Package: IDmining (via r-universe)

November 5, 2024

Type Package

Title Intrinsic Dimension for Data Mining

Version 1.0.7

Author Jean Golay [aut, cre], Mohamed Laib [aut]

Maintainer Jean Golay <jeangolay@gmail.com>

Description Contains techniques for mining large and high-dimensional data sets by using the concept of Intrinsic Dimension (ID). Here the ID is not necessarily an integer. It is extended to fractal dimensions. And the Morisita estimator is used for the ID estimation, but other tools are included as well.

Imports data.table, doParallel, parallel, foreach, stats, utils

License CC BY-NC-SA 4.0

URL https://www.sites.google.com/site/jeangolayresearch/

Encoding UTF-8

RoxygenNote 7.1.1

Note The authors are grateful to Mikhail Kanevski, Michael Leuenberger, Carmen D. Vega Orozco and Fabian Guignard for many fruitful discussions about the use of intrinsic dimension in data mining.

Repository https://jeangolay.r-universe.dev

RemoteUrl https://github.com/jeangolay/idmining

RemoteRef HEAD

RemoteSha 5d12d610e2111be78013365ca7b9bc3a769f8a46

Contents

`IDmining-package` *IDmining: Intrinsic Dimension for Data Mining*

Description

Contains techniques for mining large and high-dimensional data sets by using the concept of Intrinsic Dimension (ID). Here the ID is not necessarily an integer. It is extended to fractal dimensions. And the Morisita estimator is used for the ID estimation, but other tools are included as well.

Author(s)

Jean Golay <jeangolay@gmail.com> and Mohamed Laib <laib.med@gmail.com>,

Maintainer: Jean Golay <jeangolay@gmail.com>

References

J. Golay and M. Kanevski (2015). A new estimator of intrinsic dimension based on the multipoint Morisita index, Pattern Recognition 48 (12):4070–4081.

J. Golay, M. Leuenberger and M. Kanevski (2017). Feature selection for regression problems based on the Morisita estimator of intrinsic dimension, Pattern Recognition 70:126–138.

J. Golay and M. Kanevski (2017). Unsupervised feature selection based on the Morisita estimator of intrinsic dimension, Knowledge-Based Systems 135:125-134.

J. Golay, M. Leuenberger and M. Kanevski (2015). Morisita-based feature selection for regression problems. Proceedings of the 23rd European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN), Bruges (Belgium).

See Also

Useful links:

- https://www.sites.google.com/site/jeangolayresearch/

`Butterfly` *Butterfly Data Set Generator*

Description

Generates a random simulation of the butterfly data set with a given number of points.

Usage

```
Butterfly(N=10000)
```

Arguments

`N`	The number of points to be generated (by default: `N = 10000`).

Value

A $N \times 9$ `data.frame`. The first eight columns are the input variables, and the last one is the output (or target) variable Y.

Author(s)

Jean Golay <jeangolay@gmail.com>

References

J. Golay, M. Leuenberger and M. Kanevski (2016). Feature selection for regression problems based on the Morisita estimator of intrinsic dimension, Pattern Recognition 70:126–138.

Examples

```
bf <- Butterfly(1000)

## Not run:
require(colorRamps)
require(rgl)

c <- cut(bf$Y,breaks=64)
cols <- matlab.like(64)[as.numeric(c)]

plot3d(bf$X1,bf$X2,bf$Y,col=cols,radius=0.10,type="s",
       xlab="",ylab="",zlab="",box=F)
axes3d(lwd=3,cex.axis=3)
grid3d(c("x+","y-","z"),col="black",lwd=1)

## End(Not run)
```

`logMINDEX` *The Multipoint Morisita Index in 1, 2 or Higher Dimensions*

Description

Computes the ln values of the multipoint Morisita index in 1, 2 or higher dimensional spaces.

Usage

```
logMINDEX(X, scaleQ=1:5, mMin=2, mMax=2)
```

Arguments

`X`	A $N \times E$ `matrix`, `data.frame` or `data.table` where N is the number of data points and E is the number of variables (or features). Each variable is rescaled to the $[0, 1]$ interval by the function.
`scaleQ`	Either a single value or a vector. It contains the value(s) of ℓ^{-1} chosen by the user (by default: `scaleQ = 1:5`).
`mMin`	The minimum value of m (by default: `mMin = 2`).
`mMax`	The maximum value of m (by default: `mMax = 2`).

Details

1. ℓ is the edge length of the grid cells (or quadrats). Since the variables (and consenquently the grid) are rescaled to the $[0, 1]$ interval, ℓ is equal to 1 for a grid consisting of only one cell.
2. ℓ^{-1} is the number of grid cells (or quadrats) along each axis of the Euclidean space in which the data points are embedded.
3. ℓ^{-1} is equal to $Q^{(1/E)}$ where Q is the number of grid cells and E is the number of variables (or features).
4. ℓ^{-1} is directly related to δ (see References).
5. δ is the diagonal length of the grid cells.

Value

A `data.frame` containing the $\ln$ value of the m-Morisita index for each value of $\ln(\delta)$ and m. Notice also that the values of $\ln(\delta)$ are provided with regard to the $[0, 1]$ interval.

Author(s)

Jean Golay <jeangolay@gmail.com>

References

J. Golay and M. Kanevski (2015). A new estimator of intrinsic dimension based on the multipoint Morisita index, Pattern Recognition 48 (12):4070–4081.

Examples

```
sim_dat <- SwissRoll(1000)

m <- 2
scaleQ <- 1:15 # It starts with a grid of 1^E cell (or quadrat).
               # It ends with a grid of 15^E cells (or quadrats).
lnmMI <- logMINDEX(sim_dat, scaleQ, m, m)

dev.new(width=5, height=4)
plot(exp(lnmMI[,1]),exp(lnmMI[,2]),pch=19,col="black",xlab="",ylab="")
title(xlab = expression(delta), cex.lab = 1.5,line = 2.5)
title(ylab = expression(I['2,'*delta]), cex.lab = 1.5,line = 2.5)

dev.new(width=5, height=4)
plot(lnmMI[,1],lnmMI[,2],pch=19,col="black",xlab="",ylab="")
title(xlab = expression(paste("log(",delta,")")), cex.lab = 1.5,line = 2.5)
title(ylab = expression(paste("log(",I['2,'*delta],")")), cex.lab = 1.5,line = 2.5)
```

MBFR *Morisita-Based Filter for Regression Problems*

Description

Executes the MBFR algorithm for supervised feature selection.

Usage

```
MBFR(XY, scaleQ, m=2, C=NULL)
```

Arguments

XY	A $N \times E$ matrix, data.frame or data.table where N is the number of data points, E is the number of variables (i.e. the input variables also called "features" + the output variable). The last column contains the values of the output variable. And each variable (input + output) is rescaled to the $[0, 1]$ interval by the function.
scaleQ	A vector containing the values of ℓ^{-1} chosen by the user (see Details).
m	The value of the parameter m (by default: m=2).
C	The number of steps of the SFS procedure (by default: C = E-1).

Details

1. ℓ is the edge length of the grid cells (or quadrats). Since the data (and consenquently the grid) are rescaled to the $[0, 1]$ interval, ℓ is equal to 1 for a grid consisting of only one cell.
2. ℓ^{-1} is the number of grid cells (or quadrats) along each axis of the Euclidean space in which the data points are embedded.

3. ℓ^{-1} is equal to $Q^{(1/E)}$ where Q is the number of grid cells and E is the number of variables (or features).
4. ℓ^{-1} is directly related to δ (see References).
5. δ is the diagonal length of the grid cells.
6. The values of ℓ^{-1} in `scaleQ` must be chosen according to the linear part of the log-log plot relating the log values of the multipoint Morisita index to the log values of δ (or, equivalently, to the log values of ℓ^{-1}) (see `logMINDEX`).

Value

A list of five elements:

1. a vector containing the identifier numbers of the original features in the order they are selected through the Sequential Forward Selection (SFS) search procedure.
2. the names of the corresponding features.
3. the corresponding values of $Diss$.
4. the ID estimate of the output variable.
5. a $C \times 3$ matrix containing: (column 1) the ID estimates of the subsets retained by the SFS procedure with the target variable; (column 2) the ID estimates of the subsets retained by the SFS procedure without the output variable; (column 3) the values of $Diss$ of the subsets retained by the SFS procedure.

Author(s)

Jean Golay <jeangolay@gmail.com>

References

J. Golay, M. Leuenberger and M. Kanevski (2017). Feature selection for regression problems based on the Morisita estimator of intrinsic dimension, Pattern Recognition 70:126–138.

J. Golay, M. Leuenberger and M. Kanevski (2015). Morisita-based feature selection for regression problems.Proceedings of the 23rd European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN), Bruges (Belgium).

Examples

```
## Not run:
bf <- Butterfly(10000)

fly_select <- MBFR(bf, 5:25)
var_order  <- fly_select[[2]]
var_perf   <- fly_select[[3]]

dev.new(width=5, height=4)
plot(var_perf,type="b",pch=16,lwd=2,xaxt="n",xlab="",ylab="",
     ylim=c(0,1),col="red",panel.first={grid(lwd=1.5)})
axis(1,1:length(var_order),labels=var_order)
mtext(1,text = "Added Features (from left to right)",line = 2.5,cex=1)
```

```
mtext(2,text = "Estimated Dissimilarity",line = 2.5,cex=1)

## End(Not run)
```

MBFR_parallel *Morisita-Based Filter for Regression Problems (Parallel)*

Description

Executes the MBFR algorithm on a chosen number of workers (CPU parallel computing).

Usage

```
MBFR_parallel(XY, scaleQ, m=2, C=NULL, ncores=4)
```

Arguments

XY	A $N \times E$ matrix, data.frame or data.table where N is the number of data points, E is the number of variables (i.e. the input variables also called "features" + the output variable). The last column contains the values of the output variable. And each variable (input + output) is rescaled to the $[0, 1]$ interval by the function.
scaleQ	A vector containing the values of ℓ^{-1} chosen by the user (see Details).
m	The value of the parameter m (by default: m=2).
C	The number of steps of the SFS procedure (by default: C = E-1).
ncores	Number of workers (by default: ncores = 4)

Details

1. ℓ is the edge length of the grid cells (or quadrats). Since the data (and consenquently the grid) are rescaled to the $[0, 1]$ interval, ℓ is equal to 1 for a grid consisting of only one cell.
2. ℓ^{-1} is the number of grid cells (or quadrats) along each axis of the Euclidean space in which the data points are embedded.
3. ℓ^{-1} is equal to $Q^{(1/E)}$ where Q is the number of grid cells and E is the number of variables (or features).
4. ℓ^{-1} is directly related to δ (see References).
5. δ is the diagonal length of the grid cells.
6. The values of ℓ^{-1} in scaleQ must be chosen according to the linear part of the $\log$-$\log$ plot relating the $\log$ values of the multipoint Morisita index to the $\log$ values of δ (or, equivalently, to the $\log$ values of ℓ^{-1}) (see logMINDEX).

Value

A list of five elements:

1. a vector containing the identifier numbers of the original features in the order they are selected through the Sequential Forward Selection (SFS) search procedure.
2. the names of the corresponding features.
3. the corresponding values of $Diss$.
4. the ID estimate of the output variable.
5. a $C \times 3$ matrix containing: (column 1) the ID estimates of the subsets retained by the SFS procedure with the target variable; (column 2) the ID estimates of the subsets retained by the SFS procedure without the output variable; (column 3) the values of $Diss$ of the subsets retained by the SFS procedure.

Author(s)

Jean Golay <jeangolay@gmail.com>

References

J. Golay, M. Leuenberger and M. Kanevski (2017). Feature selection for regression problems based on the Morisita estimator of intrinsic dimension, Pattern Recognition 70:126–138.

J. Golay, M. Leuenberger and M. Kanevski (2015). Morisita-based feature selection for regression problems.Proceedings of the 23rd European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN), Bruges (Belgium).

Examples

```
## Not run:
bf <- Butterfly(10000)

fly_select <- MBFR_parallel(bf, 5:25, ncores=2)
var_order  <- fly_select[[2]]
var_perf   <- fly_select[[3]]

dev.new(width=5, height=4)
plot(var_perf,type="b",pch=16,lwd=2,xaxt="n",xlab="",ylab="",
     ylim=c(0,1),col="red",panel.first={grid(lwd=1.5)})
axis(1,1:length(var_order),labels=var_order)
mtext(1,text = "Added Features (from left to right)",line = 2.5,cex=1)
mtext(2,text = "Estimated Dissimilarity",line = 2.5,cex=1)

bf_large <- Butterfly(10^5)
system.time(MBFR(bf_large, 5:25))
system.time(MBFR_parallel(bf_large, 5:25))

## End(Not run)
```

MBRM *Morisita-Based Filter for Redundancy Minimization*

Description

Executes the MBRM algorithm for unsupervised feature selection.

Usage

```
MBRM(X, scaleQ, m=2, C=NULL, ID_tot=NULL)
```

Arguments

X	A $N \times E$ matrix, data.frame or data.table where N is the number of data points and E is the number of variables (or features). Each variable is rescaled to the $[0, 1]$ interval by the function.
scaleQ	A vector containing the values of ℓ^{-1} chosen by the user (see Details).
m	The value of the parameter m (by default: m=2).
C	The number of steps of the SFS procedure (by default: C = E).
ID_tot	The value of the full data ID if it is known a priori (by default: the value of ID_tot is estimated using the Morisita estimator of ID witin the function).

Details

1. ℓ is the edge length of the grid cells (or quadrats). Since the the variables (and consenquently the grid) are rescaled to the $[0, 1]$ interval, ℓ is equal to 1 for a grid consisting of only one cell.
2. ℓ^{-1} is the number of grid cells (or quadrats) along each axis of the Euclidean space in which the data points are embedded.
3. ℓ^{-1} is equal to $Q^{(1/E)}$ where Q is the number of grid cells and E is the number of variables (or features).
4. ℓ^{-1} is directly related to δ (see References).
5. δ is the diagonal length of the grid cells.
6. The values of ℓ^{-1} in scaleQ must be chosen according to the linear part of the $\log$-$\log$ plot relating the $\log$ values of the multipoint Morisita index to the $\log$ values of δ (or, equivalently, to the $\log$ values of ℓ^{-1}) (see logMINDEX).

Value

A list of four elements:

1. a vector containing the identifier numbers of the original features in the order they are selected through the Sequential Forward Selection (SFS) search procedure.
2. the names of the corresponding features.
3. the corresponding ID estimates.
4. the ID estimate of the full data set.

Author(s)

Jean Golay <jeangolay@gmail.com>

References

J. Golay and M. Kanevski (2017). Unsupervised feature selection based on the Morisita estimator of intrinsic dimension, Knowledge-Based Systems 135:125-134.

Examples

```
## Not run:
bf <- Butterfly(10000)

bf_select <- MBRM(bf[,-9], 5:25)
var_order <- bf_select[[2]]
var_perf  <- bf_select[[3]]

dev.new(width=5, height=4)
plot(var_perf,type="b",pch=16,lwd=2,xaxt="n",xlab="", ylab="",
     col="red",ylim=c(0,max(var_perf)),panel.first={grid(lwd=1.5)})
axis(1,1:length(var_order),labels=var_order)
mtext(1,text="Added Features (from left to right)",line=2.5,cex=1)
mtext(2,text="Estimated ID",line=2.5,cex=1)

## End(Not run)
```

`MBRM_parallel` *Morisita-Based Filter for Redundancy Minimization (Parallel)*

Description

Executes the MBRM algorithm for unsupervised feature selection (CPU parallel computing).

Usage

```
MBRM_parallel(X, scaleQ, m=2, C=NULL, ID_tot=NULL, ncores=4)
```

Arguments

`X`	A $N \times E$ `matrix`, `data.frame` or `data.table` where N is the number of data points and E is the number of variables (or features). Each variable is rescaled to the $[0, 1]$ interval by the function.
`scaleQ`	A vector containing the values of ℓ^{-1} chosen by the user (see Details).
`m`	The value of the parameter m (by default: `m=2`).
`C`	The number of steps of the SFS procedure (by default: `C = E`).
`ID_tot`	The value of the full data ID if it is known a priori (by default: the value of ID_tot is estimated using the Morisita estimator of ID witin the function).
`ncores`	Number of workers (by default: `ncores = 4`).

Details

1. ℓ is the edge length of the grid cells (or quadrats). Since the the variables (and consenquently the grid) are rescaled to the $[0, 1]$ interval, ℓ is equal to 1 for a grid consisting of only one cell.
2. ℓ^{-1} is the number of grid cells (or quadrats) along each axis of the Euclidean space in which the data points are embedded.
3. ℓ^{-1} is equal to $Q^{(1/E)}$ where Q is the number of grid cells and E is the number of variables (or features).
4. ℓ^{-1} is directly related to δ (see References).
5. δ is the diagonal length of the grid cells.
6. The values of ℓ^{-1} in `scaleQ` must be chosen according to the linear part of the $\log$-$\log$ plot relating the $\log$ values of the multipoint Morisita index to the $\log$ values of δ (or, equivalently, to the $\log$ values of ℓ^{-1}) (see `logMINDEX`).

Value

A list of four elements:

1. a vector containing the identifier numbers of the original features in the order they are selected through the Sequential Forward Selection (SFS) search procedure.
2. the names of the corresponding features.
3. the corresponding ID estimates.
4. the ID estimate of the full data set.

Author(s)

Jean Golay <jeangolay@gmail.com>

References

J. Golay and M. Kanevski (2017). Unsupervised feature selection based on the Morisita estimator of intrinsic dimension, Knowledge-Based Systems 135:125-134.

Examples

```
bf <- Butterfly(10000)

bf_select <- MBRM_parallel(bf[,-9], 5:25, ncores=2)
var_order <- bf_select[[2]]
var_perf  <- bf_select[[3]]

## Not run:
dev.new(width=5, height=4)
plot(var_perf,type="b",pch=16,lwd=2,xaxt="n",xlab="", ylab="",
     col="red",ylim=c(0,max(var_perf)),panel.first={grid(lwd=1.5)})
axis(1,1:length(var_order),labels=var_order)
mtext(1,text="Added Features (from left to right)",line=2.5,cex=1)
mtext(2,text="Estimated ID",line=2.5,cex=1)
```

```
bf_large <- Butterfly(10^5)
system.time(MBRM(bf_large[,-9], 5:25))
system.time(MBRM_parallel(bf_large[,-9], 5:25))

## End(Not run)
```

MINDEX_SP *The Multipoint Morisita Index for Spatial Patterns*

Description

Computes the multipoint Morisita index for spatial patterns (i.e. 2-dimensional patterns).

Usage

```
MINDEX_SP(X, scaleQ=1:5, mMin=2, mMax=5, Wlim_x=NULL, Wlim_y=NULL)
```

Arguments

X	A $N \times 2$ matrix, data.frame or data.table containing the X and Y coordinates of N data points. The X coordinates must be given in the first column and the Y coordinates in the second column.
scaleQ	Either a single value or a vector. It contains the value(s) of $Q^{(1/2)}$ chosen by the user where Q is the number of cells (or quadrats) of the $2D$ grid (by default: scaleQ = 1:5).
mMin	The minimum value of m (by default: mMin = 2).
mMax	The maximum value of m (by default: mMax = 5).
Wlim_x	A vector controlling the spatial extent of the $2D$ gird along the X axis. It consists of two real values, i.e. Wlim_x <- c(a,b) where $b > a$ (by default: Wlim_x <- c(min(X[,1]),max(X[,1]))).
Wlim_y	A vector controlling the spatial extent of the $2D$ gird along the Y axis. It consists of two real values, i.e. Wlim_y <- c(a,b) where $b > a$ (by default: Wlim_y <- c(min(X[,2]),max(X[,2]))).

Details

1. $Q^{(1/2)}$ is the number of grid cells (or quadrats) along each of the two axes.
2. $Q^{(1/2)}$ is directly related to δ (see References).
3. δ is the diagonal length of the grid cells.

Value

A data.frame containing the value of the m-Morisita index for each value of δ and m.

Author(s)

Jean Golay <jeangolay@gmail.com>

References

J. Golay, M. Kanevski, C. D. Vega Orozco and M. Leuenberger (2014). The multipoint Morisita index for the analysis of spatial patterns, Physica A 406:191–202.

L. Telesca, J. Golay and M. Kanevski (2015). Morisita-based space-clustering analysis of Swiss seismicity, Physica A 419:40–47.

L. Telesca, M. Lovallo, J. Golay and M. Kanevski (2016). Comparing seismicity declustering techniques by means of the joint use of Allan Factor and Morisita index, Stochastic Environmental Research and Risk Assessment 30(1):77-90.

Examples

```
sim_dat <- SwissRoll(1000)

m <- 2
scaleQ <- 1:15 # It starts with a grid of 1^2 cell (or quadrat).
               # It ends with a grid of 15^2 cells (or quadrats).
mMI <- MINDEX_SP(sim_dat[,c(1,2)], scaleQ, m, 5)

plot(mMI[,1],mMI[,2],pch=19,col="black",xlab="",ylab="")
title(xlab=expression(delta),cex.lab=1.5,line=2.5)
title(ylab=expression(I['2,'*delta]),cex.lab=1.5,line=2.5)

## Not run:
require(colorRamps)
colfunc <- colorRampPalette(c("blue","red"))
color <- colfunc(4)
dev.new(width=5,height=4)
plot(mMI[5:15,1],mMI[5:15,2],pch=19,col=color[1],xlab="",ylab="",
     ylim=c(1,max(mMI[,5])))
title(xlab=expression(delta),cex.lab=1.5,line=2.5)
title(ylab=expression(I['2,'*delta]),cex.lab=1.5,line=2.5)
for(i in 3:5){
  points(mMI[5:15,1],mMI[5:15,i],pch=19,col=color[i-1])
}
legend.text<-c("m=2","m=3","m=4","m=5")
legend.pch=c(19,19,19,19)
legend.lwd=c(NA,NA,NA,NA)
legend.col=c(color[1],color[2],color[3],color[4])
legend("topright",legend=legend.text,pch=legend.pch,lwd=legend.lwd,
       col=legend.col,ncol=1,text.col="black",cex=0.9,box.lwd=1,bg="white")

xlim_l <- c(-5,5)     # By default, the spatial extent of the grid is set so
ylim_l <- c(-6,6)     # that it is the same as the spatial extent of the data.
xlim_s <- c(-0.6,0.2) # But it can be modified to cover either a larger (l)
ylim_s <- c(-1,0.5)   # or a smaller (s) study area (or validity domain).

mMI_l <- MINDEX_SP(sim_dat[,c(1,2)], scaleQ, m, 5, xlim_l, ylim_l)
mMI_s <- MINDEX_SP(sim_dat[,c(1,2)], scaleQ, m, 5, xlim_s, ylim_s)

## End(Not run)
```

MINDID *The (Multipoint) Morisita Index for Intrinsic Dimension Estimation*

Description

Estimates the intrinsic dimension of data using the Morisita estimator of intrinsic dimension.

Usage

```
MINDID(X, scaleQ=1:5, mMin=2, mMax=2)
```

Arguments

X	A $N \times E$ matrix, data.frame or data.table where N is the number of data points and E is the number of variables (or features). Each variable is rescaled to the $[0, 1]$ interval by the function.
scaleQ	A vector (at least two values). It contains the values of ℓ^{-1} chosen by the user (by default: scaleQ = 1:5).
mMin	The minimum value of m (by default: mMin = 2).
mMax	The maximum value of m (by default: mMax = 2).

Details

1. ℓ is the edge length of the grid cells (or quadrats). Since the variables (and consenquently the grid) are rescaled to the $[0, 1]$ interval, ℓ is equal to 1 for a grid consisting of only one cell.
2. ℓ^{-1} is the number of grid cells (or quadrats) along each axis of the Euclidean space in which the data points are embedded.
3. ℓ^{-1} is equal to $Q^{(1/E)}$ where Q is the number of grid cells and E is the number of variables (or features).
4. ℓ^{-1} is directly related to δ (see References).
5. δ is the diagonal length of the grid cells.

Value

A list of two elements:

1. a data.frame containing the $\ln$ value of the m-Morisita index for each value of $\ln(\delta)$ and m. The values of $\ln(\delta)$ are provided with regard to the $[0, 1]$ interval.
2. a data.frame containing the values of S_m and M_m for each value of m.

Author(s)

Jean Golay <jeangolay@gmail.com>

References

J. Golay and M. Kanevski (2015). A new estimator of intrinsic dimension based on the multipoint Morisita index, Pattern Recognition 48 (12):4070–4081.

J. Golay, M. Leuenberger and M. Kanevski (2017). Feature selection for regression problems based on the Morisita estimator of intrinsic dimension, Pattern Recognition 70:126–138.

J. Golay and M. Kanevski (2017). Unsupervised feature selection based on the Morisita estimator of intrinsic dimension, Knowledge-Based Systems 135:125-134.

J. Golay, M. Leuenberger and M. Kanevski (2015). Morisita-based feature selection for regression problems. Proceedings of the 23rd European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN), Bruges (Belgium).

Examples

```
sim_dat <- SwissRoll(1000)

scaleQ <- 1:15 # It starts with a grid of 1^E cell (or quadrat).
               # It ends with a grid of 15^E cells (or quadrats).
mMI_ID <- MINDID(sim_dat, scaleQ[5:15])

print(paste("The ID estimate is equal to",round(mMI_ID[[1]][1,3],2)))
```

MINDID_FMC *Functional Measure of Clustering Using the Morisita Estimator of ID*

Description

Computes the functional m-Morisita index for a given set of threshold values.

Usage

```
MINDID_FMC(XY, scaleQ, m=2, thd)
```

Arguments

XY	A $N \times E$ matrix, data.frame or data.table where N is the number of data points and E is the number of variables (i.e. the input variables + the variable measured at each measurement station). The last column contains the variable measured at each measurement station. And each input variable is rescaled to the [0,1] interval by the function. Typically, the input variables are the X and Y coordinates of the measurement stations, but other or additional variables can be considered as well.
scaleQ	A vector containing the values of ℓ^{-1} chosen by the user (see Details).
m	The value of the parameter m (by default: m=2).
thd	Either a single value or a vector. It contains the value(s) of the threshold(s).

Details

1. ℓ is the edge length of the grid cells (or quadrats). Since the input variables (and consenquently the grid) are rescaled to the $[0, 1]$ interval, ℓ is equal to 1 for a grid consisting of only one cell.
2. ℓ^{-1} is the number of grid cells (or quadrats) along each axis of the Euclidean space in which the data points are embedded.
3. ℓ^{-1} is equal to $Q^{(1/E)}$ where Q is the number of grid cells and E is the number of variables (or features).
4. ℓ^{-1} is directly related to δ (see References).
5. δ is the diagonal length of the grid cells.

Value

A `vector` containing the value(s) of the m-Morisita slope, S_m, for each threshold value.

Author(s)

Jean Golay <jeangolay@gmail.com>

References

J. Golay, M. Kanevski, C. D. Vega Orozco and M. Leuenberger (2014). The multipoint Morisita index for the analysis of spatial patterns, Physica A 406:191–202.

J. Golay and M. Kanevski (2015). A new estimator of intrinsic dimension based on the multipoint Morisita index, Pattern Recognition 48 (12):4070–4081.

L. Telesca, J. Golay and M. Kanevski (2015). Morisita-based space-clustering analysis of Swiss seismicity, Physica A 419:40–47.

Examples

```
## Not run:
bf    <- Butterfly(10000)
bf_SP <- bf[,c(1,2,9)]

m      <- 2
scaleQ <- 5:25
thd    <- quantile(bf_SP$Y,probs=c(0,0.1,0.2,0.3,
                                   0.4,0.5,0.6,
                                   0.7,0.8,0.9))

nbr_shuf    <- 100
Sm_thd_shuf <- matrix(0,length(thd),nbr_shuf)
for (i in 1:nbr_shuf){
  bf_SP_shuf      <- cbind(bf_SP[,1:2],sample(bf_SP$Y,length(bf_SP$Y)))
  Sm_thd_shuf[,i] <- MINDID_FMC(bf_SP_shuf, scaleQ, m, thd)
}
mean_shuf <- apply(Sm_thd_shuf,1,mean)

dev.new(width=6, height=4)
matplot(1:10,Sm_thd_shuf,type="l",lty=1,col=rgb(1,0,0,0.25),
```

```
        ylim=c(-0.05,0.05),ylab=bquote(S[.(m)]),xaxt="n",
        xlab="",cex.lab=1.2)
axis(1,1:10,labels = FALSE)
text(1:10,par("usr")[3]-0.01,srt=45,ad=1,
     labels=c("0_100", "10_100","20_100","30_100",
              "40_100","50_100","60_100",
              "70_100","80_100","90_100"),xpd=T,font=2,cex=1)
mtext("Thresholds",side=1,line=3.5,cex=1.2)
lines(1:10,mean_shuf,type="b",col="blue",pch=19)

legend.text<-c("Shuffled","mean")
legend.pch=c(NA,19)
legend.lwd=c(2,2)
legend.col=c("red","blue")
legend("topleft",legend=legend.text,pch=legend.pch,lwd=legend.lwd,
       col=legend.col,ncol=1,text.col="black",cex=1,box.lwd=1,bg="white")

## End(Not run)
```

RenDim *Renyi's Generalized Dimensions*

Description

Estimates Rényi's generalized dimensions (or Rényi's dimensions of qth order). It is mainly for $q = 2$ that the result is used as an estimate of the intrinsic dimension of data.

Usage

```
RenDim(X, scaleQ=1:5, qMin=2, qMax=2)
```

Arguments

X	A $N \times E$ `matrix`, `data.frame` or `data.table` where N is the number of data points and E is the number of variables (or features). Each variable is rescaled to the $[0, 1]$ interval by the function.
scaleQ	A vector (at least two values). It contains the values of ℓ^{-1} chosen by the user (by default: `scaleQ = 1:5`).
qMin	The minimum value of q (by default: `qMin = 2`).
qMax	The maximum value of q (by default: `qMax = 2`).

Details

1. ℓ is the edge length of the grid cells (or quadrats). Since the variables (and consenquently the grid) are rescaled to the $[0, 1]$ interval, ℓ is equal to 1 for a grid consisting of only one cell.
2. ℓ^{-1} is the number of grid cells (or quadrats) along each axis of the Euclidean space in which the data points are embedded.

3. ℓ^{-1} is equal to $Q^{(1/E)}$ where Q is the number of grid cells and E is the number of variables (or features).
4. ℓ^{-1} is directly related to δ (see References).
5. δ is the diagonal length of the grid cells.

Value

A list of two elements:

1. a `data.frame` containing the value of Rényi's information of qth order (computed using the natural logarithm) for each value of $\ln(\delta)$ and q. The values of $\ln(\delta)$ are provided with regard to the $[0, 1]$ interval.
2. a `data.frame` containing the value of D_q for each value of q.

Author(s)

Jean Golay <jeangolay@gmail.com>

References

C. Traina Jr., A. J. M. Traina, L. Wu and C. Faloutsos (2000). Fast feature selection using fractal dimension. Proceedings of the 15th Brazilian Symposium on Databases (SBBD 2000), João Pessoa (Brazil).

E. P. M. De Sousa, C. Traina Jr., A. J. M. Traina, L. Wu and C. Faloutsos (2007). A fast and effective method to find correlations among attributes in databases, Data Mining and Knowledge Discovery 14(3):367-407.

J. Golay and M. Kanevski (2015). A new estimator of intrinsic dimension based on the multipoint Morisita index, Pattern Recognition 48 (12):4070–4081.

H. Hentschel and I. Procaccia (1983). The infinite number of generalized dimensions of fractals and strange attractors, Physica D 8(3):435-444.

Examples

```
sim_dat <- SwissRoll(1000)

scaleQ <- 1:15 # It starts with a grid of 1^E cell (or quadrat).
               # It ends with a grid of 15^E cells (or quadrats).
qRI_ID <- RenDim(sim_dat[,c(1,2)], scaleQ[5:15])

print(paste("The ID estimate is equal to",round(qRI_ID[[1]][1,2],2)))
```

`SwissRoll` *Swiss Roll Data Set Generator*

Description

Generates random points on the Swiss Roll manifold.

Usage

```
SwissRoll(N=10000)
```

Arguments

`N`	The number of points to be generated (by default: `N = 10000`).

Value

A $N \times 3$ `data.frame` containing the coordinates of the Swiss roll data points embedded in $\mathbb{R}^3$.

References

J. A. Lee and M. Verleysen (2007). Nonlinear Dimensionality Reduction, Springer, New York.

Examples

```
sim_dat <- SwissRoll(1000)
```

Index