Package 'IDmining'

November 25, 2016

Type Package

Date 2016-11-25

Title Intrinsic Dimension for Data Mining

Version 1.0.0

Author Jean Golay and Mohamed Laib

Maintainer Jean Golay <Jean.Golay@unil.ch>

Description Contains techniques for mining large high-dimensional data sets by using the concept of Intrinsic Dimension (ID). Here the ID is not necessarily 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 dplyr, stats

License CC BY-NC-SA 4.0

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

Encoding UTF-8

LazyData true

RoxygenNote 5.0.1

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

NeedsCompilation no

Repository CRAN

Date/Publication 2016-11-25 19:47:55

R topics documented:

IDmining-package 2
Butterfly 2
logMINDEX 3
MINDEX_SP 5

MINDID 7
RenDim 8
SwissRoll 10

Index **11**

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

Description

Contains techniques for mining large high-dimensional data sets by using the concept of Intrinsic Dimension (ID). Here the ID is not necessarily 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 <Jean.Golay@unil.ch> and Mohamed Laib <Mohamed.Laib@unil.ch>,

Maintainer: Jean Golay <Jean.Golay@unil.ch>

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 (2016). Feature Selection for Regression Problems Based on the Morisita Estimator of Intrinsic Dimension, arXiv:1602.00216.

J. Golay and M. Kanevski (2016). Unsupervised Feature Selection Based on the Morisita Estimator of Intrinsic Dimension, arXiv:1608.05581.

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).

`Butterfly` *Butterfly Data Set Generator*

Description

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

Usage

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

Arguments

`N`	The number of points to be generated.

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.

Author(s)

Jean Golay <Jean.Golay@unil.ch>

References

J. Golay, M. Leuenberger and M. Kanevski (2016). Feature Selection for Regression Problems Based on the Morisita Estimator of Intrinsic Dimension, arXiv:1602.00216.

Examples

```
N <- 1000
bf <- Butterfly(N)

## 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 or data frame where N is the number of data points and E is the number of variables (or features). The values of X are 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 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. $\ell^{(-1)}$ is the number of grid cells (or quadrats) along each axis of the Euclidean space in which the data points are embedded.
3. $\ell^{(-1)}$ is equal to $Q^{(1/E)}$ where Q is the number of grid cells and E is the number of variables (or features).
4. $\ell^{(-1)}$ 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 <Jean.Golay@unil.ch>

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

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

m <- 2
scaleQ <- seq(1,15,1) # 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)
```

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 or data frame 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 <Jean.Golay@unil.ch>

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

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

m <- 2
scaleQ <- seq(1,15,1) # 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 or data frame where N is the number of data points and E is the number of variables (or features). The values of X are 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 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. $\ell^{(-1)}$ is the number of grid cells (or quadrats) along each axis of the Euclidean space in which the data points are embedded.
3. $\ell^{(-1)}$ is equal to $Q^{(1/E)}$ where Q is the number of grid cells and E is the number of variables (or features).
4. $\ell^{(-1)}$ 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 <Jean.Golay@unil.ch>

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 (2016). Feature Selection for Regression Problems Based on the Morisita Estimator of Intrinsic Dimension, arXiv:1602.00216.

J. Golay and M. Kanevski (2016). Unsupervised Feature Selection Based on the Morisita Estimator of Intrinsic Dimension, arXiv:1608.05581.

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

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

m <- 2
scaleQ <- seq(1,15,1) # 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)))
```

`RenDim` *Renyi's Generalized Dimensions*

Description

Estimates Rényi's generalized dimensions (or Rényi's dimensions of q^{th} 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 or data frame where N is the number of data points and E is the number of variables (or features). The values of `X` are 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` = $1:5$).
`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 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. $\ell^{(-1)}$ is the number of grid cells (or quadrats) along each axis of the Euclidean space in which the data points are embedded.
3. $\ell^{(-1)}$ is equal to $Q^{(1/E)}$ where Q is the number of grid cells and E is the number of variables (or features).
4. $\ell^{(-1)}$ 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 q^{th} 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 <Jean.Golay@unil.ch>

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

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

m <- 2
scaleQ <- seq(1,15,1) # 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.

Value

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

References

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

Examples

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

Index

Butterfly, 2

IDmining *(*IDmining-package*)*, 2
IDmining-package, 2

logMINDEX, 3

MINDEX_SP, 5
MINDID, 7

RenDim, 8

SwissRoll, 10