Package 'DEMOVA'

March 15, 2016

Type Package

Title DEvelopment (of Multi-Linear QSPR/QSAR) MOdels VAlidated using Test Set

Version 1.0

Date 2016-03-15

Author Vinca Prana

Maintainer Vinca Prana <vinca.prana@free.fr>

Description Tool for the development of multi-linear QSPR/QSAR models (Quantitative structure-property/activity relationship). Theses models are used in chemistry, biology and pharmacy to find a relationship between the structure of a molecule and its property (such as activity, toxicology but also physical properties). The various functions of this package allows: selection of descriptors based of variances, intercorrelation and user expertise; selection of the best multi-linear regression in terms of correlation and robustness; methods of internal validation (Leave-One-Out, Leave-Many-Out, Y-scrambling) and external using test sets.

License GPL (>= 2)

Depends leaps

Suggests testthat

NeedsCompilation no

Repository CRAN

Date/Publication 2016-03-15 19:54:06

R topics documented:

`DEMOVA-package` *DEvelopment of (multi-linear QSPR/QSAR) MOdels VAlidated using test set.*

Description

Tool for the development of multi-linear QSPR/QSAR models (Quantitative structure-property/activity relationship). Theses models are used in chemistry, biology and pharmacy to find a relationship between the structure of a molecule and its property (such as activity, toxicology but also physical properties). The various functions of this package allows: selection of descriptors based of variances, intercorrelation and user expertise; selection of the best multi-linear regression in terms of correlation and robustness; methods of internal validation (Leave-One-Out, Leave-Many-Out, Y-scrambling) and external using test sets.

Details

Package:	DEMOVA
Type:	Package
Version:	1.0
Date:	2016-03-15
License:	GPL (>= 2)

Example of input files are avaible into the floder "tests".
data<-read.csv("NameOfInputFile.csv",header = TRUE , sep=" ")
mydesc<-data[,3:dim[2]]
Functions should be use in this order:
- preselection
- select_variables
- select_MLR
- fit
- LOO / LMO / Scramb (No specific order between these ones. Optional to do the rest)
- prediction
- graphe_3Sets

Author(s)

Vinca Prana
Maintainer: Vinca Prana <vinca.prana@free.fr>

References

1. Selassie, C. D. History of Quantitative Structure-Activity Relationship; Burger's Medicinal Chemistry and Drug Discovery Sixth Edition; John Wiley & Sons Inc., 2002; Vol. 1. (2)

2. Willett, P. Chemoinformatics: a History. Wiley Interdisciplinary Reviews: Computational Molecular Science 2011, 1, 46-56.

`fitting` *Performance of selected model*

Description

Perform a multi linear regression between property and previously selected descriptors (using select_MLR function).
Calculate R2 coefficient and the predicted values from the MLR. Trace the graph experimental values vs predicted values.

Usage

```
fitting(mydata, n, property)
```

Arguments

`mydata`	Dataframe containing names and values of response and descriptors
`n`	Number of selected descriptors of the regression (determined using select_MLR function)
`property`	Name of the studied proterty

Value

`prediction_TrainSet_Y.csv`
: File containing prediction obtained using the fitting

`Y_TrainingSet.tiff`
: Image representing experimental values vs predicted values for the training set

`fit`
: lm object return by the function

Examples

```
# First run select_MLR to define n
# y<-data[,2]
# mydata<-cbind(y,MLR)
# fit<-fitting(data,dim(MLR)[2],"Name of property")
```

graphe_3Sets *Predictions for the external validation set and graph*

Description

Calulate the predicted values for the external validation set and trace the graph experimental values vs predicted values for training, test and external validation sets.

Usage

```
graphe_3Sets(fit, mydata, mynewdata, mynewdata2, n)
```

Arguments

fit	Multi linear regression between property and selected descriptors (lm object)
mydata	Dataframe containing names and values of response and descriptors
mynewdata	Dataframe containing property and selected descriptors values for the test set
mynewdata2	Dataframe containing property and selected descriptors values for the external validation set
n	Numbers of selected descriptors of the regression (determined using select_MLR)

Value

Rext,Rext2	return a list containing the value of the determination coefficient of the test set and of the external validation set
Graphe_3sets.tiff	Image representing experimental values vs predicted values for the all three sets

Examples

```
# This function have to be run last!

## "Test_set.csv" should be with the following form
## ID property SelectedDesc1 SelectedDesc2 ...

# new_nom<-'Test_set.csv'
# newdata<-read.csv(new_nom,header=TRUE , sep=" ")
# mynewdata=newdata[,2:dim[2]]


## "External_set.csv" should be with the following form
## ID property SelectedDesc1 SelectedDesc2 ...

# new_nom2<-'External_set.csv'
# newdata2<-read.csv(new_nom2,header=TRUE , sep=" ")
# mynewdata2=newdata2[,2:dim[2]]

#graphe_3Sets(fit,mynewdata,mynewdata2,dim(MLR)[2])
```

LMO *Leave Many Out*

Description

Calculate the robustness of the equation using the leave many out method.

Usage

```
LMO(mydata, cv, n)
```

Arguments

mydata	Dataframe containing names and values of response and descriptors
cv	Numbers of fold
n	Numbers of selected descriptors of the regression (determined using Select_MLR)

Value

return Q2, the coefficient that measure the robstness

References

1. Gramatica, P. Principles of QSAR Models Validation: Internal and External. Qsar & Combinatorial Science 2007, 26, 694-701.
2. Golbraikh, A.; Tropsha, A. Beware of Q(2)! Journal of Molecular Graphics & Modelling 2002, 20, 269-276.

Examples

```
# First run Select_MLR to define n

#LMO(mydata,5,dim(MLR)[2])
#LMO(mydata,10,dim(MLR)[2])
```

LOO *Leave One Out*

Description

Calculate the robustness of the equation using the leave one out method.

Usage

```
LOO(mydata, n)
```

Arguments

mydata	Dataframe containing names and values of response and descriptors
n	Numbers of selected descriptors of the regression (determined using Select_MLR)

Value

return Q2, the coefficient that measure the robstness

References

1. Gramatica, P. Principles of QSAR Models Validation: Internal and External. Qsar & Combinatorial Science 2007, 26, 694-701.
2. Golbraikh, A.; Tropsha, A. Beware of Q(2)! Journal of Molecular Graphics & Modelling 2002, 20, 269-276.

Examples

```
# First run Select_MLR to define n

# LOO(mydata,dim(MLR)[2])
```

---

prediction *Predictions for the test set and graph*

Description

Calulate the predicted values for the test set and trace the graph experimental values vs predicted values for both training and test sets. This function also give the R2 test coefficent.

Usage

```
prediction(fit, mydata, mynewdata, n)
```

Arguments

fit	Multi linear regression between property and selected descriptors
mydata	Dataframe containing names and values of response and descriptors
mynewdata	Dataframe containing property and selected descriptors values for the test set
n	Numbers of selected descriptors of the regression (determined using Select_MLR)

Value

Exp.vs.Pred.tiff	Image representing experimental values vs predicted values for the both sets
Rext	return the value of the determination coefficient of the test set

Examples

```
# This function have to be run after choise of the model.

## "Test_set.csv" should be with the following form
## ID property SelectedDesc1 SelectedDesc2 ...

#new_nom<-'Test_set.csv'
#newdata<-read.csv(new_nom,header=TRUE , sep=" ")
#mynewdata=newdata[,2:dim[2]]

#prediction(fit,mynewdata,dim(MLR)[2])
```

`preselection` *Suppression of missing or constant descriptors*

Description

Remove descriptors with missing values and a variance lower than 0.001.

Usage

```
preselection(desc)
```

Arguments

`desc`	Dataframe containing the names of desciptors and their values

Value

return a dataframe without the removed variables

Examples

```
## The input file should be with the following form
## id_molecule propriete x1 x2 x3 ... # Header line
## molecule1 1 0.02 500 ...
## molecule2 5 0.06 600 ...

# nom<-"NameOfInputFile.csv"
# data<-read.csv(nom,header = TRUE , sep=" ")
# dim<-dim(data)
# mydesc<-data[,3:dim[2]]
# id<-data[,1]
# y<-data[,2]

# d<-preselection(mydesc)
```

scramb *scrambling*

Description

Perform the y-scrambling method that consit to permute y values and try to develop new models. They have to be unperformants in order to validate the original one. The graph R2 vs r(y,yrandom) is created.

Usage

```
scramb(mydata, k, n, cercle = FALSE)
```

Arguments

mydata	Dataframe containing names and values of response and descriptors
k	Number of random run
n	Number of selected descriptors of the regression (determined using Select_MLR)
cercle	Value is TRUE or FALSE (by default) . If it TRUE it's draw a circle around the point representinf the original model

Value

Return a list of

mean	Mean of R^2 new model
sd	RStandard deviation of R^2 new model

And also

Scramb.tiff	Description of 'comp1'
Scramb.csv	Description of 'comp2'

References

Tropsha, A.; Gramatica, P.; Gombar, V. K. The Importance of Being Earnest: Validation Is the Absolute Essential for Successful Application and Interpretation of QSPR Models. Qsar \& Combinatorial Science 2003, 22, 69-77.
Rucker, C.; Rucker, G.; Meringer, M. y-Randomization and Its Variants in QSPR/QSAR. J. Chem. Inf. Model. 2007, 47, 2345-2357.
Lindgren, F.; Hansen, B.; Karcher, W.; Sjostrom, M.; Eriksson, L. Model Validation by Permutation Tests: Applications to Variable Selection. Journal of Chemometrics 1996, 10, 521-532.

Examples

```
# First run Select_MLR to define n

# scramb(mydata,1000,nom,dim(MLR)[2])
```

select_MLR *Development of the model (multi linear regression)*

Description

From a list of descriptors and responses values, this function choose the best compromise between correlation and robustness to select the best model.

Usage

```
select_MLR(y, desc, n, method = "forward")
```

Arguments

y	Vector with values of the property/response
desc	Dataframe containing the names of desciptors and their values
n	Maximal number (integer) of desciptors for the final equation
method	Determine the method used to build the regression. Can be: "backward", "forward" (by default) or "seqrep". For more info see leaps package.

Value

Return the list of selected variables for the choosen MLR.

Examples

```
# First run Select_variables to remove descriptors with missing or constant values.
# MLR<-select_MLR(y,desc,5)
```

select_variables *Selection of descriptors*

Description

This function allow the user to select wanted descriptors between both that are intercorrelated with a correlation coefficent higher that ThresholdInterCor. The selection can also be automatic based on the correlation with the property of each variables.

Usage

```
select_variables(id, y, d, ThresholdInterCor, auto = FALSE)
```

Arguments

id	List of the names of observations
y	List of the values of the property/response
d	Dataframe containing the names of desciptors and their values (without missing or constant values)
ThresholdInterCor	Threshold value (double) of the accepted intercorrelation between descriptors (should be between 0 and 1)
auto	Two possible values: TRUE or FALSE (by default). The selection of descriptors is done automatically based on the correlation between descriptor and property (auto=TRUE) or is done manually by user (auto=FALSE)

Value

return a dataframe containing only of non intercorrelated variables

Examples

```
# Run after Preselection : d<-Preselection(desc)

# desc<-select_variables(id,y,d,0.95)
```

Index