Package 'SimuChemPC'

February 16, 2014

Type Package

Title Simulation process of 4 selection methods in predicting chemical potent compounds

Version 1.2

Date 2014-02-16

Author Mohsen Ahmadi

Maintainer Mohsen Ahmadi <mohsen_ahmadi989@yahoo.com>

Description This package provides simulation process of 4 selection methods in predicting potent compounds

License GPL-3

Depends stats, graphics, grDevices, utils, R (>= 2.13)

R topics documented:

`predictChemPC` *predictChemPC*

Description

This function performs a prediction method from 4 predicting potent compounds methods of this package. These methods are RA, EI, NN and GP.

Usage

```
predictChemPC(xTrain, yTrain, xTest, loghyper, method="RA")
```

Arguments

xTrain	m * n martrix of train data.
yTrain	m * 1 matrix of target values.
xTest	j * n matrix of test data.
loghyper	3 * 1 matrix of loghyper parameters which is the output of trainChemPC function.
method	One of "EI", "GP", "NN" or "RA".

Details

This function withholds 4 methods to predict potent compounds.

method is one of: EI A compound for which maximum expected potency improvement is reached. GP A compound holding maximum predicted potency in test data is selected. NN A compound that is nearest (Tonimito Coefficient as distance measure) to the most potent compound in training data is selected. RA As it's name suggests, a compound is selected randomly.

Feature selection Feature selection employed in this package is based on Spearman Rank Correlation such that before each training step those attributes in which revealed a significant Spearman rank correlation with the logarithmic potency values (q-value < 5 are computed from original p-values via the multiple testing correction method by Benjamini and Hochberg.

Value

It returns index of most potent compound in original test set w.r.t. selected method.

Author(s)

Mohsen Ahmadi

References

1.Predicting Potent Compounds via Model-Based Global Optimization, Journal of Chemical Information and Modeling, 2013, 53 (3), pp 553-559, M Ahmadi, M Vogt, P Iyer, J Bajorath, H Froehlich. 2.Software MOE is used to calculate the numerical descriptors in data sets. Ref: http://www.chemcomp.com/MOE-Molecular_Operating_Environment.htm 3.ChEMBL was the source of the compound data and potency annotations in data sets. Ref: https://www.ebi.ac.uk/chembl/

Examples

```
x = as.data.frame(array(1:100, dim=c(20,5)))
y = as.matrix(as.numeric(array(1:20, dim=c(20,1))))
xstar = as.data.frame(array(5:105, dim=c(20,5)))
loghyper = trainChemPC(x, y)
index = predictChemPC(x, y, xstar, loghyper, method="RA")
```

SimuChemPC *SimuChemPC*

Description

This function executes a simulation to compare 4 methods for predicting potent compounds. These methods are EI, GP, NN and RA.

Usage

```
SimuChemPC(dataX, dataY, method="RA", experiment=1)
```

Arguments

dataX	m * n martrix of data (features/descriptors).
dataY	m * 1 matrix of target values (potencies).
method	One of "EI", "GP", "NN" or "RA".
experiment	An integer value that indicates a number by which experiment repeats. In our published experiment it was set to 25.

Details

This function withholds 4 simulation methods to predict potent compounds . method can be RA, NN , EI or GP. The explanation of the abbreviations is listed below.

RA selection: One compound will be selected randomly and added to train data each time.

NN selection: The compound which is nearest (based on Tonimito Coefficient) to the most potent compound in training data is selected and added to train data.

EI selection A compound for which maximum expected potency improvement is reached, is selected and then it is added to train data.

GP selection A compound holding maximum potency in test data is selected.

Feature selection Feature selection employed in this package is based on Spearman Rank Correlation such that before each training step those attributes in which revealed a significant Spearman rank correlation with the logarithmic potency values (q-value < 5 are computed from original p-values via the multiple testing correction method by Benjamini and Hochberg.

The purpose of simulation step Simulation step is employed to select the compound(in the case where input files are chemical compounds) in which maximal expected potency improvement is met. Subsequently, this compound is added to train data and simulation continues until all test data are consumed. Finally, the number of simulation steps is determined which the algorithm used to select the most potent compound in the "original" test set.

In this code, given our data sets (chemical compounds), we do the followings:

1. We split our data into two distinguish parts namely Train and Test data

2. We do normalization on both parts

3. We employ a specific feature selection algorithm (i.e. Multiple Testing Correction) to overcome high dimensionality

4. Then we benefit Gaussian Process Regression in order to learn our model iteratively such that in each iteration training data are trained, the model is learnt and prediction is done for test data. One

compound holding specific property will be added to train data and the progress will repeat until no test data is left.

Result of this work is accepted in the Journal of Chemical Information and Modeling within the subject "Predicting Potent Compounds via Model-Based Global Optimization".

Value

returns a matrix (m * experiment) of original potencies in test set.

Author(s)

Mohsen Ahmadi

References

1. Predicting Potent Compounds via Model-Based Global Optimization, Journal of Chemical Information and Modeling, 2013, 53 (3), pp 553-559, M Ahmadi, M Vogt, P Iyer, J Bajorath, H Froehlich. 2. Software MOE is used to calculate the numerical descriptors in data sets. Ref: http://www.chemcomp.com/MOE-Molecular_Operating_Environment.htm 3. ChEMBL was the source of the compound data and potency annotations in data sets. Ref: https://www.ebi.ac.uk/chembl/

Examples

```
x = as.data.frame(array(1:100, dim=c(20,5)))
y = as.matrix(as.numeric(array(1:20, dim=c(20,1))))
SimuChemPC(x, y, "RA", 5)
```

---

trainChemPC *trainChemPC*

Description

Apply Gaussian Process Regression to learn the model.

Usage

```
trainChemPC(xTrain, yTrain)
```

Arguments

xTrain	m * n martrix of train data.
yTrain	m * 1 matrix of target values.

Details

This function performs training step of GP or EI by finding loghyper parameters.

Value

It returns a vector that holds calculated loghyper parameters.

Author(s)

Mohsen Ahmadi

References

1. Predicting Potent Compounds via Model-Based Global Optimization, Journal of Chemical Information and Modeling, 2013, 53 (3), pp 553-559, M Ahmadi, M Vogt, P Iyer, J Bajorath, H Froehlich. 2. Software MOE is used to calculate the numerical descriptors in data sets. Ref: http://www.chemcomp.com/MOE-Molecular_Operating_Environment.htm 3. ChEMBL was the source of the compound data and potency annotations in data sets. Ref: https://www.ebi.ac.uk/chembl/

Examples

```
x = as.data.frame(array(1:100, dim=c(20,5)))
y = as.matrix(as.numeric(array(1:20, dim=c(20,1))))
loghyper = trainChemPC(x, y)
```

Index