R documentation

of 'SimuChemPC.Rd'

December 1, 2013

SimuChemPC *SimuChemPC*

Description

This function excutes a simulation to compare 4 methods for predicting potent compounds. These methods are Random selection, EI selection, 1NN selection and GP selection.

Usage

```
SimuChemPC( dataFile, seedFile, method, repeatExperiment = 25)
```

Arguments

dataFile	dataFile specifies address of dataset file to use.
seedFile	seedFile is an input random seed file which should be generated before. Simulation process uses it to randomise test and learning data selection.
method	method a string value to specify prediction and learning method. Its value can be one of random, 1NN, EI or GP.
repeatExperiment	repeatExperiment a integer value that declares number that the experiment repeats. In our published experiment it was 25.

Details

This function withholds 4 simulation methods to predict potent compounds . There exist a set of sample seed and dataset files in the package which belong the relevant paper mentioned in the reference. method can be random, 1NN , EI or GP. The explanation of the abbreviations is listed below.

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

1NN selection: The compound for 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 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.

`Format of Input files` Each Input file is a 2-dim matrix of real values in which the last column contains target values and the rest are attributes.

`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 untill 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 test set.

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

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

2. We do normalizatoin 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

It stores a sequence of predicted values of a selected method into a datafile. The file name will consist method and input data file name.

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

```
library(gpr)
library(SimuChemPC)
seedpath = "seeds_for_random_generatorMatlab.txt"
seedFile = system.file("extdata", seedpath , package="SimuChemPC")
datapath = "11407_Descriptors_Potency.txt"
dataFile  = system.file("extdata", datapath , package="SimuChemPC")
method  = "random"
repeatExperiment = 1
SimuChemPC( dataFile, seedFile, method , repeatExperiment)
```

Index