R documentation

of ‘SimuChemPC.Rd’

August 28, 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, dataset, seedFile, simulationType , repeatExperiment =25)
```

Arguments

`dataFile` `dataFile` specifies address of dataset file to use.

`dataset` `dataset` is selected name of a dataset. RData output file will have this name.

`seedFile` `seedFile` is an input random seed file which should be generated before. Simulation process uses it to randomise test and learning data selection.

`simulationType` `simulationType` a string value to specify simulation type. Its value can be 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 predicting potent compounds . There exist a set of sample seed and dataset files in the package which belong the relevant paper mentioned in the reference. `simulationType` 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 We pick a compound for which maximum expected improvement is reached and then add to train data.

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

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

When coloring is FALSE returns a Steiner tree in form of a new igraph object. When coloring is TRUE returns a list that consists of two objects. The first is a steiner tree and the second object is a colored version of the input graph with distinguished steiner nodes and terminals.

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")
dataset=11407
datapath=paste("",dataset,"_Descriptors_Potency.txt",sep="")
dataFile  = system.file("extdata", datapath , package="SimuChemPC")
simulationType = "GP"
```

```
repeatExperiment = 1
SimuChemPC( dataFile, dataset, seedFile, simulationType , repeatExperiment)
```

Index