Package 'M3C'

April 12, 2018

Title Monte Carlo Consensus Clustering

Version 1.0.0

Description A central task in genomic data analyses for stratified medicine is class discovery which is accomplished through clustering. However, an unresolved problem with current clustering algorithms is they do not test the null hypothesis and derive p values. To solve this, we developed a novel hypothesis testing framework that uses consensus clustering called Monte Carlo Consensus Clustering (M3C). M3C use a multi-core enabled Monte Carlo simulation to generate a distribution of stability scores for each number of clusters using null datasets with the same gene-gene correlation structure as the real one. These distributions are used to derive p values and a beta distribution is fitted to the data to cheaply estimate p values beyond the limits of the simulation. M3C improves accuracy, allows rejection of the null hypothesis, removes systematic bias, and uses p values to make class number decisions. We believe M3C deals with a major pitfall in current automated class discovery tools.

Depends R (>= 3.4.0)

License AGPL-3

Encoding UTF-8

LazyData true

Imports ggplot2, Matrix, doSNOW, NMF, RColorBrewer, cluster, parallel, foreach, doParallel, matrixcalc

Suggests knitr, rmarkdown

VignetteBuilder knitr

RoxygenNote 6.0.1

biocViews Clustering, GeneExpression, Transcription, RNASeq, Sequencing

NeedsCompilation no

Author Christopher John [aut, cre]

Maintainer Christopher John <chris.r.john86@gmail.com>

R topics documented:

clustersim *clustersim: A cluster simulator for testing clustering algorithms*

Description

clustersim: A cluster simulator for testing clustering algorithms

Usage

```
clustersim(n, n2, r, K, alpha, wobble, print = FALSE, seed = NULL)
```

Arguments

n	Numerical value: The number of samples, it must be square rootable
n2	Numerical value: The number of features
r	Numerical value: The radius to define the initial circle (use approx n/100)
K	Numerical value: How many clusters to simulate
alpha	Numerical value: How far to pull apart the clusters
wobble	Numerical value: The degree of noise to add to the sample co ordinates
print	Logical flag: whether to print the PCA into current directory
seed	Numerical value: fixes the seed if you want to repeat results

Value

A list: containing 1) matrix with simulated data in it

Examples

```
res <- clustersim(225, 900, 8, 4, 0.75, 0.025, print = TRUE, seed=123)
```

desx *GBM clinical annotation data*

Description

This is the clinical annotation data from the GBM dataset, it contains the class of the tumour which is one of: classical, mesenchymal, neural, proneural. It is a data frame with 2 columns and 50 rows.

Author(s)

Chris John <chris.r.john86@gmail.com>

References

Verhaak, Roel GW, et al. "Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1." Cancer cell 17.1 (2010): 98-110.

M3C *M3C: Monte Carlo Consensus Clustering*

Description

This function runs M3C, which is a hypothesis testing framework for consensus clustering. The basic idea is to use a multi-core enabled monte carlo simulation to drive the creation of a null distribution of stability scores. The monte carlo simulations maintains the correlation structure of the input data. Then the null distribution is used to compare the reference scores with the real scores and a empirical p value is calculated for every value of K. We also use the relative cluster stability index as an alternative metric which is just based on a comparison against the reference mean, the advantage being it requires fewer iterations. Small p values are estimated cheaply using a beta distribution that is inferred using parameter estimates from the monte carlo simulation.

Usage

```
M3C(mydata, montecarlo = TRUE, cores = 1, iters = 100, maxK = 10,
  des = NULL, ref_method = c("reverse-pca", "chol"), repsref = 100,
  repsreal = 100, clusteralg = c("pam", "km"), distance = "euclidean",
  pacx1 = 0.1, pacx2 = 0.9, printres = FALSE, printheatmaps = FALSE,
  showheatmaps = FALSE, seed = NULL, removeplots = FALSE)
```

Arguments

mydata	Data frame or matrix: Contains the data, with samples as columns and rows as features
montecarlo	Logical flag: whether to run the monte carlo simulation or not (recommended: TRUE)
cores	Numerical value: how many cores to split the monte carlo simulation over
iters	Numerical value: how many monte carlo iterations to perform (default: 100, recommended: 100-1000)
maxK	Numerical value: the maximum number of clusters to test for, K (default: 10)
des	Data frame: contains annotation data for the input data for automatic reordering (optional)
ref_method	Character string: refers to which reference method to use (recommended: leaving as default)
repsref	Numerical value: how many reps to use for the monte carlo reference data (suggest 100)
repsreal	Numerical value: how many reps to use for the real data (recommended: 100)
clusteralg	String: dictates which algorithm to use for M3C (recommended: leaving as default)
distance	String: dictates which distance metric to use for M3C (recommended: leaving as default)
pacx1	Numerical value: The 1st x co-ordinate for calculating the pac score from the CDF (default: 0.1)
pacx2	Numerical value: The 2nd x co-ordinate for calculating the pac score from the CDF (default: 0.9)

printres	Logical flag: whether to print all results into current directory
printheatmaps	Logical flag: whether to print all the heatmaps into current directory
showheatmaps	Logical flag: whether to show the heatmaps on screen (can be slow)
seed	Numerical value: fixes the seed if you want to repeat results, set the seed to 123 for example here
removeplots	Logical flag: whether to remove all plots (recommended: leaving as default)

Value

A list, containing: 1) the stability results and 2) all the output data (another list) 3) reference stability scores (see vignette for more details on how to easily access)

Examples

```
res <- M3C(mydata, cores=1, iters=100, ref_method = 'reverse-pca', montecarlo = TRUE,printres = FALSE,
maxK = 10, showheatmaps = FALSE, repsreal = 100, repsref = 100,printheatmaps = FALSE, seed = 123, des = desx)
```

---

mydata *GBM expression data*

Description

This is the expression data from the GBM dataset. It is a data frame with 50 columns and 1740 rows.

Author(s)

Chris John <chris.r.john86@gmail.com>

References

Verhaak, Roel GW, et al. "Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1." Cancer cell 17.1 (2010): 98-110.

Index