Package ‘breakpoint’

October 20, 2014

Type Package

Title An R Package for Multiple Break-Point Detection via the Cross-Entropy Method

Version 1.0

Date 2014-01-14

Author Priyadarshana W.J.R.M. and Georgy Sofronov

Maintainer Priyadarshana W.J.R.M. <madawa.weerasinghe@mq.edu.au>

Description

Implements the cross-entropy (CE) method, which is a model based stochastic optimization technique to estimate both the number as well as the corresponding locations of break-points in biological sequences of continuous measurements (Priyadarshana and Sofronov (2014)).

License GPL(>=2)

Depends R (>= 2.5.0), ggplot2, foreach

Suggests snow, doSNOW, doMC, parallel

R topics documented:

breakpoint-package 1
CE 2
data 4

Index 5

breakpoint-package *Multiple Break-Point Detection in Continuous Measurements*

Description

The breakpoint package implements a variant of the cross-entropy (CE) method proposed in Priyadarshana and Sofronov (2014) to estimate both the number as well as the corresponding locations of break-points in biological sequences of continuous measurements. The methodology primarily built to detect multiple break-points in genomic sequences of continuous measurements. However, it can be easily extended and applied to other problems.

Details

Package:	breakpoint
Type:	Package
Version:	1.0
Date:	2013-09-16
License:	GPL 2.0

"breakpoint"" package provides estimates on both the number as well as the corresponding locations of break-points in continuous scale measurements. The algorithm utilizes the cross-entropy (CE) method, which is a model based stochastic optimization procedure to obtain the estimates. Current implementation of the methodology works as an exact search method in estimating the number of break-points. It selects the best solution from the solution space based on the modified BIC introduced in Zhang & Seigmund (2007). A parallel implementation of the algorithm can be carried-out in Unix/Linux/MAC OS X and Windows operating systems with the use of "doMC", "parallel", "snow" and "doSNOW" packages.

Author(s)

Priyadarshana, W.J.R.M. and Sofronov, G.

Maintainer: Priyadarshana, W.J.R.M. <madawa.weerasinghe@mq.edu.au>

References

Priyadarshana, W.J.R.M. and Sofronov, G. (2014) Multiple Break-Points Detection in array CGH Data via the Cross-Entropy Method. (Submitted)

Priyadarshana,W.J.R.M., and Sofronov, G. (2012) A Modified Cross Entropy Method for Detecting Multiple Change Points in DNA Count Data. In Proc.IEEE World Congress on Computational Intelligence (CEC2012), 1020-1027.

Rubinstein, R., and Kroese, D. (2004) The Cross-Entropy Method: A Unified Approach to Combinatorial Optimization, Monte-Carlo Simulation and Machine Learning. Springer-Verlag, New York.

Zhang,N.R., and Seigmund, D.O. (2007) A modified Bayes information criterion with applications to the analysis of comparative genomic hybridization data. Biometrics, 63, 22-32.

CE *Break-point Detection via the CE Method*

Description

This function carries out calculations to estimate both the number of break-points as well as their corresponding locations based on the Cross-Entropy (CE) method for sequences of continuous measurements, particulary for genomic sequences (array CGH data).

Usage

```
CE(data, N_max = 10, eps = 0.01, rho = 0.05, M = 200, h = 5,a=0.8,
x.label="Data Sequence", y.label="Value",parallel = FALSE)
```

Arguments

`data` Data to be analysed. A single column array or a dataframe.

`N_max` Maximum number of break-points. Default vlaue is 10.

`eps` The cut-off value for the Median Absolute Deviation value. Default value is 0.01.

`rho` The fraction which is used to obtain the best performing set of sample solutions (elite sample). Default value is 0.05

`M` Sample size to be used in simulating the locations of break-points from four-parameter beta distribution. Default value is 200.

`h` Minimum abberation width. Deafult is 5.

`a` Smoothing parameter value. Deafult is 0.8.

`x.label` x-axis label in the profile plot

`y.label` y-axis label in the profile plot

`parallel` A logical argument specifying if parallel computation should be carried-out (TRUE) or not (FALSE). By default it is set as 'FALSE'. In Windows OS systems "snow" functionalities are used, whereas in Unix/Linux/MAC OSX "multicore" functionalities are used to carryout parallel computations with the maximum number of cores available.

Details

The CE algorithm is a model based stochastic optimization method. In the breakpoint package it is used as an exact search method. A performance function score (modified BIC, Zhang & Seigmund (2007)) is calculated for each of the solutions generated by the four-parameter beta distribution from no change-point to the user provided maximum number of break-points. The solution that maximizes the modified BIC with respect to the number of break-points is considered as the optimal solution. Finally a vector of break-point locations are given along with the mean profile plot.

A list that contains the break-points and their corresponding locations are given in the console. The mean profile plot is also produced as an output. Furthermore, it stores information on computational time and the mean profile plot under the "CE" folder which is created in the current working directory.

Value

A list is returned containing the following items

`No.BPs` The number of break-points in the data that is estimated by the CE method

`BP.Loc` A vector of break-point locations

References

Priyadarshana, W.J.R.M. and Sofronov, G. (2014) Multiple Break-Points Detection in array CGH Data via the Cross-Entropy Method. (Submitted)

Priyadarshana,W.J.R.M., and Sofronov, G. (2012) A Modified Cross Entropy Method for Detecting Multiple Change Points in DNA Count Data. In Proc.IEEE World Congress on Computational Intelligence (CEC2012), 1020-1027.

Rubinstein, R., and Kroese, D. (2004) The Cross-Entropy Method: A Unified Approach to Combinatorial Optimization, Monte-Carlo Simulation and Machine Learning. Springer-Verlag, New York.

Zhang,N.R., and Seigmund, D.O. (2007) A modified Bayes information criterion with applications to the analysis of comparative genomic hybridization data. Biometrics, 63, 22-32.

Examples

```
data(data)
## Not run:
CE(data)

## End(Not run)
```

data *Fibroblast cell line (GM03563) data*

Description

Chromosome 1 of cell line GM03563

Usage

```
data(data)
```

Format

A single column data frame with 135 observations corresponds to chromosome 1 of cell line GM03563.

`log2ratio` Normalized average of the log base 2 test over reference ratio data

Details

This data set is extracted from a single experiments on 15 fibroblast cell lines with each array containing over 2000 (mapped) BACs spotted in triplicate discussed in Snijders et al.(2001). Data corresponds to the chromosome 1 of cell line GM03563.

References

Snijders,A.M. et al. (2001) Assembly of microarrays for genome-wide measurement of DNA copy number. Nature Genetics, 29, 263-264.

Examples

```
data(data)
## Not run:
CE(data)

## End(Not run)
```

Index

∗Topic **data**
    data, 4

breakpoint *(*breakpoint-package*)*, 1
breakpoint-package, 1

CE, 2

data, 4