Package 'EBPRS'

November 3, 2018

Type Package

Title Derive Polygenic Risk Score Based on Emprical Bayes Theory

Version 1.0.1

Author Shuang Song [aut, cre], Wei Jiang [aut], Lin Hou [aut] and Hongyu Zhao [aut]

Maintainer Shuang Song <songs15@mails.tsinghua.edu.cn>

Description EB-PRS is a novel method that leverages information for effect sizes across all the markers to improve the prediction accuracy. No parameter tuning is needed in the method, and no external information is needed. This R-package provides the calculation of polygenic risk scores from the given training summary statistics and testing data. We can use EB-PRS to extract main information, estimate Empirical Bayes parameters, derive polygenic risk scores for each individual in testing data, and evaluate the PRS according to AUC and predictive r2.

License GPL-3

Depends ROCR, data.table, methods

Encoding UTF-8

LazyData true

RoxygenNote 6.1.0

NeedsCompilation no

Repository CRAN

Date/Publication 2018-11-03 15:30:11 UTC

R topics documented:

EBPRSpackage *Extract information from raw data*

Description

The first step of the algorithm, to clean the dataset and extract information from raw data. (Please notice that there are some requirements for the training and testing datasets.)

Usage

```
EBPRSpackage()
```

Details

EB-PRS is a novel method that leverages information for effect sizes across all the markers to improve the prediction accuracy. No parameter tuning is needed in the method, and no external information is needed. This R-package provides the calculation of polygenic risk scores from the given training summary statistics and testing data. We can use EB-PRS to extract main information, estimate Empirical Bayes parameters, derive polygenic risk scores for each individual in testing data, and evaluate the PRS according to AUC and predictive r2.

Package:	EBPRS
Type:	Package
Date:	2018-10

The package contains four main functions for users.

1. `extractInfo`. We use this to extract important information from training dataset and test dataset. There is a strict requirement for the format of imput, which is detailedly illustrated in details in `extractInfo`. Here we mention that the we recommend users first use package `plink2R` from github to read plink files into R, and the data transfered by `read_plink` from `plink2R` can be directly used as our input. A merge of training set and testing set will also be made.

`plink2R` can be installed using this command:

```
options(unzip = "internal")
devtools::install_github("gabraham/plink2R/plink2R")
```

2. `findPara`. We use this function to estimate parameters from the processed training set.

3. `generateScore`. This is the core function to generate polygenic risk score using our algorithm based on Empirical Bayes Theory.

4. `validate`. We use this to validate the performance of the PRS.

Author(s)

Shuang Song, Wei Jiang, Lin Hou and Hongyu Zhao

References

Song, S., Jiang, W., Hou, L. and Zhao, H. Leveraging effect size distributions to improve polygenic risk scores derived from genome-wide association studies. *Submitted*.

See Also

`extractInfo`, `findPara`, `generateScore`, `validate`,

https://github.com/gabraham/plink2R

`extractInfo` *Extract information from raw data*

Description

The first step of the algorithm, to clean the dataset and extract information from raw data. (Please notice that there are some requirements for the training and testing datasets.)

Usage

```
extractInfo(trainpath, test)
```

Arguments

`trainpath`	train dataset path
`test`	test dataset(list) including fam, bed, bim(generated from plink files, plink2R::read_plink is recommended)

Details

The raw training data should be a file with 8 columns including CHROM, POS, A1, A2, BETA, P, SNP, N in order. The CHROM column should only be a number from 1 to 22. The SNP column is the rsid number.

"test" file can be generated from read_plink("test_plink_file") The raw testing data could be the files transformed from plink2R (using plink bfiles).

test is a list including fam (6 columns with information on samples), bim (6 columns with information on SNPs), bed (genotypes 0, 1, 2)

Value

A list including processed training data (train) and testing data (bed, bim, fam)

Author(s)

Shuang Song, Wei Jiang, Lin Hou and Hongyu Zhao

References

Song, S., Jiang, W., Hou, L. and Zhao, H. Leveraging effect size distributions to improve polygenic risk scores derived from genome-wide association studies. *Submitted*.

See Also

https://github.com/gabraham/plink2R

`findPara` *Derive the parameters*

Description

All the input files can be generated from 'extractInfo'

Usage

```
findPara(train, bed, bim, fam, N1, N0)
```

Arguments

`train`	train set after processed by 'extractInfo'
`bed`	bed file after processed by 'extractInfo'
`bim`	bim file after processed by 'extractInfo'
`fam`	fam file after processed by 'extractInfo'
`N1`	case number
`N0`	control number

Value

A list including estimated mu (muHat) estimated sigma2 (sigmaHat2) estimated proportion of non-associated SNPs (pi0) estimated variance of effect sizes of associated SNPs (sigma02)

Author(s)

Shuang Song, Wei Jiang, Lin Hou and Hongyu Zhao

References

Song, S., Jiang, W., Hou, L. and Zhao, H. Leveraging effect size distributions to improve polygenic risk scores derived from genome-wide association studies. *Submitted*.

See Also

`extractInfo`

generateScore *Calculate the polygenic risk scores*

Description

Function that generates PRS for each individual.

Usage

```
generateScore(sigmaHat2, muHat, X)
```

Arguments

sigmaHat2	parameters generated by 'findPara', estimated variations
muHat	parameters generated by 'findPara', estimated effect sizes
X	bed file

Value

Polygenic risk scores for each individual calculated by the EBPRS model (S).

Author(s)

Shuang Song, Wei Jiang, Lin Hou and Hongyu Zhao

References

Song, S., Jiang, W., Hou, L. and Zhao, H. Leveraging effect size distributions to improve polygenic risk scores derived from genome-wide association studies. *Submitted*.

See Also

extractInfo

findPara

Examples

```
S <- generateScore(sigmaHat2=rep(1,100),muHat=rep(0.1,100),
X=matrix(sample(0:1,size=1000,replace=TRUE),10,100) )
```

validate *Validate the performance of EBPRS*

Description

Provide the performance evaluated by predictive r2 and AUC.

Usage

```
validate(fam, score)
```

Arguments

fam	fam file after processed by 'extractInfo'
score	polygenic score generated by 'generateScore'

Author(s)

Shuang Song, Wei Jiang, Lin Hou and Hongyu Zhao

References

Song, S., Jiang, W., Hou, L. and Zhao, H. Leveraging effect size distributions to improve polygenic risk scores derived from genome-wide association studies. *Submitted*.

See Also

extractInfo

findPara

generateScore

Examples

```
validate(fam=matrix(sample(0:1,20*6,replace=TRUE),ncol=6),score=rnorm(20,0,1))
```

Index