Package ‘TSDFGS’

March 7, 2019

Type    Package
Title   Training Set Determination for Genomic Selection
Version 1.0
Date    2019-03-06
Author  Jen-Hsiang Ou and Chen-Tuo Liao
Maintainer Jen-Hsiang Ou<oumark.me@outlook.com>

Description Determining training set for genomic selection using a genetic algorithm (Holland J.H. (1975) <DOI:10.1145/1216504.1216510>) or simple exchange algorithm (change an individual every iteration). Three different criteria are used in both algorithms, which are r-score (Ou J.H., Liao C.T. (2018) <DOI:10.6342/NTU201802290> ), PEV-score (Akdemir D. et al. (2015) <DOI:10.1186/s12711-015-0116-6>) and CD-score (Laloe D. (1993) <DOI:10.1186/1297-9686-25-6-557>). Phenotypic data for candidate set is not necessary for all these methods. By using it, one may readily determine a training set that can be expected to provide a better training set comparing to random sampling.

URL  https://tsdfgs.oumark.me

BugReports  https://gitlab.com/oumark/TSDFGS/issues
License  GPL (>= 3)
Imports  Rcpp (>= 1.0.0)
LinkingTo Rcpp, RcppEigen
NeedsCompilation yes
Repository CRAN
Date/Publication  2019-03-07 17:42:53 UTC

R topics documented:

TSDFGS-package .......................................................... 2
cd_score .............................................................. 3
geno ................................................................. 3
optTrain ............................................................. 4
pev_score ............................................................ 5
Training Set Determination for Genomic Selection

Description

Determining training set for genomic selection using a genetic algorithm (Holland J.H. (1975) <DOI:10.11145/1216504.1216510>) or simple exchange algorithm (change an individual every iteration). Three different criteria are used in both algorithms, which are r-score (Ou J.H., Liao C.T. (2018) <DOI:10.6342/NTU201802290>), PEV-score (Akdemir D. et al. (2015) <DOI:10.1186/s12711-015-0116-6>) and CD-score (Laloe D. (1993) <DOI:10.1186/1297-9686-25-6-557>). Phenotypic data for candidate set is not necessary for all these methods. By using it, one may readily determine a training set that can be expected to provide a better training set comparing to random sampling.

Details


Author(s)

Jen-Hsiang Ou and Chen-Tuo Liao
Maintainer: Jen-Hsiang Ou<oumark.me@outlook.com>

References


See Also

STPGA
**cd_score**

*Generalized Coefficient of Determination*

**Description**


**Usage**

`cd_score(x, x0)`

**Arguments**

- `x`  Genomic metrix of training set.
- `x0`  Genomic metrix of testing set.

**Value**

A numeric score.

**Author(s)**

Jen-Hsiang Ou <oumark.me@outlook.com>

**References**


**Examples**

```r
data("rice44kPCA")
cd_score(geno[1:100,], geno[101:200,])
```

---

**Rice 44k Genomoe Data**

**Description**

A PC Matric of Rice 44k Genomoe Data

**Usage**

```r
data("rice44kPCA")
```
Format

A numeric matrix with 404 rows and 404 columns.

Source

http://www.ricediversity.org/data/index.cfm

Examples

```R
data("rice44kPCA")
dim(geno)
```

---

**optTrain** Algorithm for optimal training set determination

**Description**


**Usage**

```R
optTrain(geno, cand, n.train, subpop = NULL,
    test = NULL, method = "rScore", min.iter = NULL)
```

**Arguments**

<table>
<thead>
<tr>
<th>Argument</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>geno</td>
<td>A numeric matrix of principal components (rows: individuals; columns: PCs). To reduce computing time, one may use first k PCs by geno[,1:k].</td>
</tr>
<tr>
<td>cand</td>
<td>An integer vector of which rows of individuals are candidates of the training set in the geno matrix.</td>
</tr>
<tr>
<td>n.train</td>
<td>The size of the target training set.</td>
</tr>
<tr>
<td>subpop</td>
<td>A character vector of subpopulation’s group name. The algorithm will ignore the population structure if it remains NULL.</td>
</tr>
<tr>
<td>test</td>
<td>An integer vector of which rows of individuals are in the test set in the geno matrix. The algorithm will use an un-target method if it remains NULL.</td>
</tr>
<tr>
<td>method</td>
<td>Choices are rScore, PEV and CD. rScore will be used by default.</td>
</tr>
<tr>
<td>min.iter</td>
<td>Minimum iteration of all methods can be appointed. One should always check if the algorithm is converged or not. A minimum iteration will set by considering the candidate and test set size if it remains NULL.</td>
</tr>
</tbody>
</table>
Value

OPTtrain An integer vector of the chosen optimal training set.
TOPscore Score of each iteration. (Given by one of three criterions)
ITERscore Score of the best solution in by far. (Given by one of three criterions.)

Note

Both genetic algorithm and simple exchange algorithms do not assure convergence to global optimal, and it is highly recommended to draw the convergence plot to check it converges to the local optimal.

Author(s)

Jen-Hsiang Ou and Chen-Tuo Liao
Maintainer: Jen-Hsiang Ou<oumark.me@outlook.com>

References


Examples

```r
## LOAD EXAMPLE DATA ##
data("rice44KPCA")
out.RNN = optTrain(geno, cand = 1:404, n.train = 100)
```

<table>
<thead>
<tr>
<th>pev_score</th>
<th>PEV Score</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Description

PEV-score (Akdemir D. et al. (2015) <DOI:10.1186/s12711-015-0116-6>) is a criterion for finding a training set which derived from the covariance of the prediction of the test set.

Usage

pev_score(X, X0)
Arguments

- X : A genetic matrix of training set.
- X0 : A genetic matrix of test set.

Value

A numeric score.

Author(s)

Jen-Hsiang Ou <oumark.me@outlook.com>

References


Examples

```r
data("rice44kPCA")
pev_score(geno[1:50,],geno[51:100,])
```

---

**Description**

This data set was provided by Zhao et al. (2011) <DOI:10.1038/ncomms1467> which genotyping 44,100 SNP variants across 413 diverse accessions of O. sativa from 82 countries. We converted the genomic information into a PC matrix.

**Usage**

```r
data("rice44kPCA")
```

**Format**

```
  geno  A numeric matrix of principal components
  subpop  A character vector of subpopulation’s group name.
```

**Source**

http://www.ricediversity.org/data/index.cfm
References

Examples
data("rice44kPCA")

<table>
<thead>
<tr>
<th>r_score</th>
<th>r Score</th>
</tr>
</thead>
</table>

Description
A criterion for finding training set which derived from Pearson’s correlation between GEBVs (genomic estimated breeding value) and phenotype value of a test set.

Usage
r_score(x, x0)

Arguments
x A genetic matrix of training set.
x0 A genetic matrix of test set.

Value
A numeric score.

Author(s)
Jen-Hsiang Ou <oumark.me@outlook.com>

References

Examples
data("rice44kPCA")
r_score(geno[1:50], geno[51:100],)
**Description**

Subpopulation of Each Individuals in Rice 44k Genome Data

**Usage**

```r
data("rice44kPCA")
```

**Format**

A character vector.

**Source**


**Examples**

```r
data("rice44kPCA")
print(subpop)
```
Index

+Topic package
  TSDFGS-package, 2

cd_score, 3

geno, 3

optTrain, 4

pev_score, 5

r_score, 7

rice44KPCA, 6

STPGA, 2

subpop, 8

TSDFGS (TSDFGS-package), 2

TSDFGS-package, 2