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 *Training Set Determination for Genomic Selection*

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.

Details

The package is used to determine the optimal training set in a highly structured, mild structured and diverse population. The function "optTrain" use a genetic algorithm or simple exchange algorithm to evaluate an optimal solution using one of the criteria (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>), CD-score(Laloe D. (1993) <DOI:10.1186/1297-9686-25-6-557>)).

Author(s)

Jen-Hsiang Ou and Chen-Tuo Liao

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

References

Akdemir D., Sanchez JI., Jannink JL. (2015), Optimization of genomic selection training populations with a genetic algorithm. Genetic Selection Evolution 47:38.\

Laloe D. (1993), Precision and information in linear models of genetic evolution. Genetics Selection Evolution 25:557.\

Ou J.H., Liao C.T. (2018), Training set determination for genomic selection. National Taiwan University Master Thesis.

See Also

STPGA

cd_score *Generalized Coefficient of Determination*

Description

A criterion for finding optimal training set using generalized coefficient of determination (Laloe D. (1993) <DOI:10.1186/1297-9686-25-6-557>).

Usage

```
cd_score(x, x0)
```

Arguments

x	Genomic metrix of training set.
x0	Genomic metric of testing set.

Value

A numeric score.

Author(s)

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

References

Laloe D. (1993), Precision and information in linear models of genetic evaluation. Genetics Selection Evolution 25:557.

Examples

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

geno *Rice 44k Genomoe Data*

Description

A PC Matric of Rice 44k Genomoe Data

Usage

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

Format

A numeric matrix with 404 rows and 404 columns.

Source

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

Examples

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

---

optTrain *Algorithm for optimal training set determination*

Description

It uses a genetic algorithm or simple exchange algorithm with three different criteria (r-score (J.H. Ou et al., (2019) <DOI:10.6342/NTU201802290>), PEV-score (Akdemir D. et al., (2015) <DOI:10.1186/s12711-015-0116-6>), CD-score (Laloe D. (1993) <DOI:10.1186/1297-9686-25-6-557>)) to determine an optimal training set.

Usage

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

Arguments

geno	A numeric matrix of principal components (rows: individuals; columns: PCs). To reduce computing time, one may use first k PCs by geno[,1:k].
cand	An integer vector of which rows of individuals are candidates of the training set in the geno matrix.
n.train	The size of the target training set.
subpop	A character vector of subpopulation's group name. The algorithm will ignore the population structure if it remains NULL.
test	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.
method	Choices are rScore, PEV and CD. rScore will be used by default.
min.iter	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.

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

Akdemir D., Sanchez JI., Jannink JL. (2015), Optimization of genomic selection training populations with a genetic algorithm. GenSetic Selection Evolution 47:38.\

Laloe D. (1993), Precision and information in linear models of genetic evolution. Genetics Selection Evolution 25:557.\

Holland J. H. (1975) Adaptation in Natural and Artificial Systems. University of Michigan Press.

Examples

```
## LOAD EXAMPLE DATA ##
data("rice44kPCA")
out.RNN = optTrain(geno, cand = 1:404, n.train = 100)
```

---

pev_score *PEV Score*

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

Akdemir D. et al., (2015), Optimization of genomic selection training populations with a genetic algoritthm. Genetics Selection Evolution 47:38.\

Kennedy B.W., Trus D., (1993), Considerations on genetic connectedness between management units under an animal model. J Anim Sci. 1993 Sep;71(9):2341-52

Examples

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

---

rice44kPCA *44k genome rice data*

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

```
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

Keyan Zhao, Chih-Wei Tung, Georgia C. Eizenga, Mark H. Wright, M. Liakat Ali, Adam H. Price, Gareth J. Norton, M. Rafiqul Islam, Andy Reynolds, Jason Mezey, Anna M. McClung, Carlos D. Bustamante & Susan R. McCouch (2011). Genome-wide association mapping reveals a rich genetic architecture of complex traits in Oryza sativa. Nat Comm 2:467 | DOI: 10.1038/ncomms1467, Published Online 13 Sep 2011.

Examples

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

`r_score` *r Score*

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

Ou J.H., Liao C.T. (2018), Training set determination for genomic selection. National Taiwan University Master Thesis.

Examples

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

subpop *Rice 44k Genome Data*

Description

Subpopulation of Each Individuals in Rice 44k Genome Data

Usage

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

Format

A character vector.

Source

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

Examples

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

Index