Package 'G2P'

August 1, 2017

Title Genomic Selection Prediction and Evalution

Version 2.0

Description Genomic Selection Prediction and Evalution.

Depends R (>= 3.2.0)

License GPL-2|GPL-3

LazyData true

RoxygenNote 5.0.1

NeedsCompilation no

Encoding UTF-8

Author Chuang Ma [aut, cre]

Maintainer Chuang Ma <chuangma2006@gmail.com>

R topics documented:

cvSampleIndex *Generate the Sample Indices of Training Sets and Testing Sets*

Description

This function be used for generating training and testing sets indices.

Usage

```
cvSampleIndex(sampleNum, cross = 5, seed = 1, randomSeed = FALSE)
```

Arguments

sampleNum	the number of samples for building genomic selection model.
cross	the fold of cross validation.
seed	Random number options,defult 1
randomSeed	logical variable,defult FALSE.

Value

A list,and each element including the $trainIdx $testIdx and cvIdx

$trainIdx The index of training samples.

$testIdx The index of testing samples.

$cvIdx The cross validation index.

Author(s)

Chuang Ma , Zhixu Qiu , Qian Cheng ,Jie Song

Examples

```
## Load example data ##
data(riceYield)
## leave-one out cross validation
a <- cvSampleIndex(sampleNum = nrow(Markers),cross = nrow(Markers),seed = 1)

## random samples cross validation
b <- cvSampleIndex(sampleNum = nrow(Markers),cross = 5,seed = 1)
## you will get a list with 5 elements and in each element,the $trainIdx amount is 80,the
## testIdx amount is 20
```

data *Example Data for G2P*

Description

The data of rice yield(SNP genotypes informations) markers A numeric matrix, each row is the each individual's SNP genotypes informations.

pheVal The real phenotype Value of each individual.

Usage

```
data(..., list = character(), package = NULL, lib.loc = NULL,
  verbose = getOption("verbose"), envir = .GlobalEnv)
```

Author(s)

Chuang Ma , Qian Cheng , Zhixu Qiu , Jie Song

Examples

```
##load rice yield datasets
data(riceYield)
```

dataCheck *Check the Markers Data and Missing Value Handling*

Description

This function is applied for data checking and missing value handling through read the genotypes informations and the phenotype values.

Usage

```
dataCheck(markers, pheVal)
```

Arguments

markers	(numeric)a matrix, each row is the each individual's SNP genotypes informations.Genotypes should be coded as 0,1,2;0 represent AA(homozygote),2 represent BB(homozygote) and 1 represent AB(heterozygote).
pheVal	(numeric)the phenotype Value of each individual.

Value

$genMat A numeric matrix including genotypes informations

$num The count of individuals

$pheVal Each individual corresponding real phenotype Value

Author(s)

Chuang Ma , Qian Cheng , Zhixu Qiu , Jie Song

Examples

```
##apply the function ##
GSData <- dataCheck(marker = Markers,pheVal = phenotype)
dim(GSData )
```

`evaluateGS` *evaluateGS*

Description

this function is used to evaluete the accuracy of predicted by genomic selection model.

Usage

```
evaluateGS(realScores, predScores, Probability = TRUE, evaMethod = "RE",
  Beta = 1, BestIndividuals = "top", topAlpha = 1:90)
```

Arguments

`realScores`	A numeric vector is the real breeding values of the validation individual for a trait.
`predScores`	A numeric vector or matrix is the prediction breeding value predicted by genomic selection model of the individuals.
`Probability`	For RE and kappa method , whether the predScores is probability? Default True.
`evaMethod`	A character vetctor is the methods selected to evaluete, which include "pearson", "kendall", "spearman", "MSE","R2" "RE", "Kappa", "auc","AUCpr","accuracy","F1","meanNDCGEvalu "NDCGEvaluation".
`Beta`	the parameter of "F1"
`BestIndividuals`	It is a stratrgy that you want to select the best individuals in the candidate groups, according to the prediction breeding value of a trait,when using RE and kappa method. if the trait was yield,flowering or disease resistance,and male flowering time to female flowering time,it is "top"(default), "buttom",and "middle", respectively. when the parameter is "top", the parameter Probability make no difference.
`topAlpha`	A numeric vector is the proportion of excellent individuals,defaulting 1:90.

Value

a list inculding evaluation results with methods which user select.

Author(s)

Chuang Ma , Zhixu Qiu , Qian Cheng ,Jie Song

Examples

```
data(riceYield)
########## predicting breeding value
predlist <-  G2P(cross = 10,seed = 1 ,cpus = 3,markers  = Markers,pheVal  = phenotype,modelMethods = c("rrBLUP","R
predMartix <- NULL
for(ii in 1:10){predMartix <- rbind(predMartix,predlist[[ii]])}
######## evaluate the accuracy of the prediction result
evaluareTest <- evaluateGS(realScores = predMartix[,1],predScores = predMartix[,2:3],evaMethod = c("pearson", "k
########## exhibit the evaluation value
REMat <- evaluareTest$RE
result_diplay(plotMartix = REMat,plotType = "graph")
```

---

feature_assess *Feature Selection*

Description

This function score each SNP set, you can screen of high grade of SNP for subsequent modeling, in order to simplify the operation and improve the precision of feature selection.(methods including Gini, Accuracy, rrBLUP)

Usage

```
feature_assess(markers, pheVal, method = c("rrBLUP", "Gini", "Accuracy"),
  ntree = 500, importance = TRUE, posPercentage = 0.4,
  BestIndividuals = c("top", "middle", "buttom"))
```

Arguments

markers	a numeric matrix, each row is the each individual's SNP genotypes informations.Genotypes should be coded as 0,1,2;0 represent AA(homozygote),2 represent BB(homozygote) and 1 represent AB(heterozygote);missing (NA) alleles are not allowed.
pheVal	the phenotype Value of each individual(numeric)
method	the method of feature selction including "Gini" "Accuracy" "rrBLUP"
ntree	the number of random forest decision tree,defult 500
posPercentage	phenotypic good proportion in Classification,defult 0.4
importantce	whether the results of variable importance,defult TRUE
betterPhenotypePosition	the better phenotype position including "top","buttom",defult "top"

Value

A numeric mode score of each position of SNPs

Author(s)

Chuang Ma , Zhixu Qiu , Qian Cheng ,Jie Song

Examples

```
## feature selection with Gini ##
Gini_selection <- feature_assess(markers = Markers,pheVal = phenotype,method = "Gini",
ntree = 500, importance = TRUE,posPercentage = 0.40, BestIndividuals = "top")

## feature selection with Acc ##
Acc_selection <- feature_assess(markers = Markers,pheVal = phenotype,method = "Accuracy",
ntree = 500, importance = TRUE,posPercentage = 0.40, BestIndividuals = "top")

## feature selection with rrBLUP ##
rrBLUP_selection <- feature_assess(markers = Markers,pheVal = phenotype,method = "rrBLUP",
posPercentage = 0.40, BestIndividuals = "top")
```

G2P *G2P*

Description

this function is apply cross validation to test Genomic Selection model trained by different methods and datas.

Usage

```
G2P(cross = 5, seed = 1, cpus = 1, markers, pheVal,
  modelMethods = "SVC", nIter = 7000, burnIn = 500, thin = 5,
  saveAt = "", S0 = NULL, df0 = 5, R2 = 0.5, weights = NULL,
  verbose = FALSE, rmExistingFiles = TRUE, groups = NULL,
  importance = FALSE, posPercentage = 0.4, BestIndividuals = c("top"),
  ntree = 500, nodesize = 1, kernel = c("radial"), gamma = 1,
  cost = 2^(-9), outputModel = TRUE, ...)
```

Arguments

markers	(numeric) a matrix, each row is the each individual's SNP genotypes informations.Genotypes should be coded as 0,1,2;0 represents AA(homozygote),2 represents BB(homozygote) and 1 represents AB(heterozygote);missing (NA) alleles are not allowed
pheVal	(numeric)the phenotype Value of each individual.

`nIter, burnIn, thin`
: (integer) the number of iterations, burn-in and thinning,default nIter 7000,burnIn 500,thin 5.

`saveAt`
: (string) this may include a path and a pre-fix that will be added to the name of the files that are saved as the program runs,default ""

`S0,`
: df0 (numeric) The scale parameter for the scaled inverse-chi squared prior assigned to the residual variance, only used with Gaussian outcomes. In the parameterization of the scaled-inverse chi square in BGLR the expected values is S0/(df0-2). The default value for the df parameter is 5. If the scale is not specified a value is calculated so that the prior mode of the residual variance equals var(y)*R2 (see below). For further details see the vignettes in the package or http://genomics.cimmyt.org/BGLR-extdoc.pdf.Default S0 NULL,df0 5.

`R2`
: (numeric, 0<R2<1) The proportion of variance that one expects, a priori, to be explained by the regression. Only used if the hyper-parameters are not specified; if that is the case, internaly, hyper-paramters are set so that the prior modes are consistent with the variance partition specified by R2 and the prior distribution is relatively flat at the mode. For further details see the vignettes in the package or http://genomics.cimmyt.org/BGLR-extdoc.pdf.Defult 0.5

`weights`
: (numeric, n) a vector of weights, may be NULL. If weights is not NULL, the residual variance of each data-point is set to be proportional to the square of the weight. Only used with Gaussian outcomes.

`verbose`
: (logical) if TRUE the iteration history is printed, default FALSE

`rmExistingFiles`
: (logical) if TRUE removes existing output files from previous runs, default TRUE.

`groups`
: (factor) a vector of the same length of y that associates observations with groups, each group will have an associated variance component for the error term.

`importance`
: RandomForest parameter:Should importance of predictors be assessed?Defualt FALSE

`posPercentage`
: (numeric)the percentage positive samples in all samples.1 > posPercentage > 0.

`BestIndividuals`
: BestIndividuals It is a position that the best individuals (positive samples) in a training group, according to the breeding values of a training group's trait. if the trait was yield,flowering or disease resistance,and male flowering time to female flowering time,it is "top"(default), "buttom",and "middle" of the breeding values, respectively.

`ntree`
: RandomForest parameter:Number of trees to grow. This should not be set to too small a number, to ensure that every input row gets predicted at least a few times.Defualt 500

`kernel`
: svm parameter the kernel used in training and predicting. You might consider changing some of the following parameters, depending on the kernel type.(linear,polynomial,sigmoid,radi
radial

`gamma`
: svm parameter parameter needed for all kernels except linear (default: 1/(data dimension))

`cost`
: svm cost of c.

outputModel	if true return the list of training model.
model	the model to fit."BayesA", "BayesB", "BayesC", "BL", "BRR", "RKHS","rrBLUP","SVR"

Value

a matrix or a list: if evaluation = FALSE a matrix with two column,and the first column is the true phenotype value,the second column is the prediction score.

if evaluation = TRUE a list including predition result and all evaluation method result

Author(s)

Chuang Ma , Zhixu Qiu , Qian Cheng ,Jie Song

Examples

```
data(riceYield)
########## predicting breeding value
predlist <-  G2P(cross = 10,seed = 1 ,cpus = 3,markers  = Markers,pheVal  = phenotype,modelMethods = c("rrBLUP","R
```

GSmachine *Fit machine learning model*

Description

This function can fit several machine learning models of genomic selection such as svm (support vector machine),ramdomforest

Usage

```
GSmachine(markers, pheVal, modelMethods = "SVC", posPercentage = 0.4,
  BestIndividuals = c("top"), ntree = 500, nodesize = 1,
  kernel = c("radial"), gamma = 1, cost = 2^(-9))
```

Arguments

markers	(numeric)a matrix, each row is the each individual's SNP genotypes informations.Genotypes should be coded as 0,1,2;0 represent AA(homozygote),2 represent BB(homozygote) and 1 represent AB(heterozygote);missing (NA) alleles are not allowed.
pheVal	(numeric)the phenotype value of each individual.
modelMethods	the methods is built genomic selection model. "SVR" or "SVC represent a regression or classification model build by using svm, and also "RFR" or "RFC" is a ramdomforest methods to build a regression or classification model
posPercentage	(numeric,1 > posPercentage > 0)the percentage of positive samples for a trait in training groups.

`BestIndividuals`
: It is a position that the best individuals (positive samples) in a training group, according to the breeding values of a training group's trait. if the trait was yield,flowering or disease resistance,and male flowering time to female flowering time,it is "top"(default), "buttom",and "middle" of the breeding values, respectively.

`ntree`
: ramdomforest parameter (integer)Number of trees to grow. This should not be set to too small a number, to ensure that every input row gets predicted at least a few times,default 500.

`nodesize`
: ramdomforest parameter Minimum size of terminal nodes. Setting this number larger causes smaller trees to be grown (and thus take less time). Note that the default values are different for classification (1) and regression (5).

`kernel`
: svm parameter the kernel used in training and predicting. You might consider changing some of the following parameters, depending on the kernel type.(linear,polynomial,sigmoid,radi
radial

`gamma`
: svm parameter parameter needed for all kernels except linear (default: 1/(data dimension))

`cost`
: svm parameter cost of constraints violation (default: 2^(-9))-it is the 'C'-constant of the regularization term in the Lagrange formulation.

`posNegSampleList`
: (integer) a list of row number of positive and negative samples $posSampleIndex row number of positive samples $negSampleIndex row number of negative samples

`mtry`
: ramdomforest parameter Number of variables randomly sampled as candidates at each split. Note that the default values are different for classification (sqrt(p) where p is number of variables in x) and regression (p/3)

Value

a machine model which is enable to predict

Author(s)

Chuang Ma , Qian Cheng , Zhixu Qiu , Jie Song

Examples

```
## Load example data ##
data(riceYield)

## Fit RFR model ##
machine_model <- GSmachine(markers = Markers,pheVal = phenotype,modelMethods = "RFR")

## Fit classification model(RFC) ##
machine_model <- GSmachine(markers = Markers,pheVal = phenotype,modelMethods = "RFC",posPercentage = 0.4,ntree =
```

GSReModel *Fit Regression Model*

Description

This function can fit several regression models of genomic selection such as BayesA, BayesB,BayesC,BRR(BayesBayesian Ridge Regression),BL(Bayesian LASSO),RHKS(Bayesian Reproducing Kernel Hilbert Space),etc.

Usage

```
GSReModel(markers, pheVal, modelMethods, nIter = 7000, burnIn = 500,
  thin = 5, saveAt = "", S0 = NULL, df0 = 5, R2 = 0.5,
  weights = NULL, verbose = FALSE, rmExistingFiles = TRUE,
  groups = NULL, ntree = 500, importance = FALSE, ...)
```

Arguments

markers	(numeric)a matrix, each row is the each individual's SNP genotypes informations.Genotypes should be coded as 0,1,2or-1,0,1;0(-1) represent AA(homozygote),2(1) represent BB(homozygote) and 1(0) represent AB(heterozygote);missing (NA) alleles are not allowed.
pheVal	(numeric)the phenotype value of each individual.
modelMethods	the model to fit."BayesA", "BayesB", "BayesC", "BL", "BRR", "RKHS","rrBLUP".
nIter, burnIn, thin	(integer) the number of iterations, burn-in and thinning,default nIter 7000,burnIn 500,thin 5.
saveAt	(string) this may include a path and a pre-fix that will be added to the name of the files that are saved as the program runs,default "".
S0,	df0 (numeric) The scale parameter for the scaled inverse-chi squared prior assigned to the residual variance, only used with Gaussian outcomes. In the parameterization of the scaled-inverse chi square in BGLR the expected values is S0/(df0-2). The default value for the df parameter is 5. If the scale is not specified a value is calculated so that the prior mode of the residual variance equals var(y)*R2 (see below). For further details see the vignettes in the package or http://genomics.cimmyt.org/BGLR-extdoc.pdf.Default S0 NULL,df0 5.
R2	(numeric, 0<R2<1) The proportion of variance that one expects, a priori, to be explained by the regression. Only used if the hyper-parameters are not specified; if that is the case, internaly, hyper-paramters are set so that the prior modes are consistent with the variance partition specified by R2 and the prior distribution is relatively flat at the mode. For further details see the vignettes in the package or http://genomics.cimmyt.org/BGLR-extdoc.pdf.Defult 0.5
weights	(numeric, n) a vector of weights, may be NULL. If weights is not NULL, the residual variance of each data-point is set to be proportional to the square of the weight. Only used with Gaussian outcomes.
verbose	(logical) if TRUE the iteration history is printed, default FALSE.

`rmExistingFiles`
: (logical) if TRUE removes existing output files from previous runs, default TRUE.

`groups`
: (factor) a vector of the same length of y that associates observations with groups, each group will have an associated variance component for the error term.

`ntree`
: RandomForest parameter:Number of trees to grow. This should not be set to too small a number, to ensure that every input row gets predicted at least a few times.Defualt 500.

`importance`
: RandomForest parameter:Should importance of predictors be assessed? Defualt FALSE.

Value

A regression model which is enable to predict

Author(s)

Chuang Ma , Qian Cheng , Zhixu Qiu , Jie Song

Examples

```
## Load example data ##
data(riceYield)

## Fit rrBLUP model ##
rrBLUP_model <- GSReModel(markers =Markers,pheVal = phenotype,modelMethods = "rrBLUP")
```

---

predictGS *Prediction with Trained Model from Geomic Selection Model*

Description

This function is give the prediction score of a new GS data by using already model.

Usage

```
predictGS(testMat, trainModel, modelMethods = "SVC")
```

Arguments

`testMat`
: (numeric)a matrix, each row is the each testing sets or new GS data individual's SNP genotypes informations.Genotypes should be coded as 0,1,2;0 represent AA(homozygote),2 represent BB(homozygote) and 1 represent AB(heterozygote);missing (NA) alleles are not allowed.

`trainModel`
: The trained model.It's type must be similar whith modelMethods.

`modelMethods`
: (character)the type name of training model including "BayesA", "BayesB", "BayesC", "BL", "BRR", "RKHS","rrBLUP","SVC","SVR","RFC","RFR".

Value

The prediction result of testing sets which predicted through already models

Author(s)

Chuang Ma , Qian Cheng , Zhixu Qiu , Jie Song

Examples

```
## Load example data ##
data(riceYield)

## Fit rrBLUP model ##
rrBLUP_model <- GSReModel(markers =Markers,pheVal = phenotype,modelMethods = "rrBLUP")

## Predict 1-20 subset of all example data with already rrBLUP model ##
res <- predictGS(testMat = Markers[1:20,],trainModel = rrBLUP_model,modelMethods = "rrBLUP")
```

---

randomSeed *Generate Random Seed*

Description

This funcation is appplied for generating random seed with current system time

Usage

```
randomSeed()
```

Value

(numeric) A random seed

Author(s)

Chuang Ma , Qian Cheng , Zhixu Qiu , Jie Song

Examples

```
## generate the random seed ##
randomSeed()
```

`result_diplay` *exhibit evaluation result and data structure*

Description

exhibit evaluation result and data structure from genomic selection.

Usage

```
result_diplay(markers, plotMartix, centers = 4,
  plotType = c("population_structure"),
  colorSet = rainbow(ncol(plotMartix)), main = NULL,
  ylab = "Relative efficiency", legend.name = colnames(plotMartix))
```

Arguments

`markers`	the marker data for genomic selection.
`plotMartix`	numeric matrix of the values to be plotted,which is given from evaluating result.
`centers`	either the number of clusters, say k, or a set of initial (distinct) cluster centres. If a number, a random set of (distinct) rows in x is chosen as the initial centres.
`plotType`	the type of plot is including "population_structure" of markers and "graph" and "heatmap " of evaluate value.
`colorSet`	vector of colors used to plot "graph".
`main`	the title of the plot.
`ylab`	a title for the y axis.
`legend.name`	a character or expression vector of length >= 1 to appear in the legend,used to plolt graph

Author(s)

Chuang Ma , Qian Cheng , Zhixu Qiu , Jie Song

Examples

```
data(riceYield)
result_diplay(markers = Markers,centers = 4,plotType = c("population_structure"))
```

sampleClassify *Generate Positive and Negative Samples for Training*

Description

This function can be use to generate positive and negative samples for training.The positive samples represent the excellent individuals which's breeding values we expect to obtain in your research.And the negative samples represent the lower breeding values of individuals.

Usage

```
sampleClassify(pheVal, posPercentage = 0.4, BestIndividuals = c("top",
  "middle", "buttom"))
```

Arguments

pheVal	(numeric)the breeding values of each individual.
posPercentage	(numeric,1 > posPercentage > 0)the percentage of positive samples for a trait in training groups.
BestIndividuals	It is a position that the best individuals (positive samples) in a training group, according to the breeding values of a training group's trait. if the trait was yield,flowering or disease resistance,and male flowering time to female flowering time,it is "top"(default), "buttom",and "middle" of the breeding values, respectively.

Value

A list of row number of positive and negative samples $posSampleIndex Index of positive samples $negSampleIndex Index of negative samples

Author(s)

Chuang Ma , Qian Cheng , Zhixu Qiu , Jie Song

Examples

```
##  percentage of positive samples is 0.4 ##
sampleClassify(phenotype,posPercentage = 0.4, BestIndividuals = "top")
```

Index