# Package 'STGS'

October 24, 2019

**Type** Package

**Title** Genomic Selection using Single Trait

**Version** 0.1.0

**Author** Neeraj Budhlakoti, D C Mishra, Anil Rai, K K Chaturvedi

**Maintainer** Neeraj Budhlakoti <neeraj35669@gmail.com>

**Description** Genomic Selection (GS) is a latest development in animal and plant breeding where whole genome markers information is used to predict genetic merit of an individuals in a practical breeding programme. GS is one of the promising tool for improving genetic gain in animal and plants in today's scenario. This package is basically developed for genomic predictions by estimating marker effects. These marker effects further used for calculation of genotypic merit of individual i.e. genome estimated breeding values (GEBVs). Genomic selection may be based on single trait or multi traits information. This package performs genomic selection only for single traits hence named as STGS i.e. single trait genomic selection. STGS is a comprehensive package which gives single step solution for genomic selection based on most commonly used statistical methods.

**License** GPL-3

**Depends** R (>= 3.6)

**Imports** glmnet, brnn, kernlab, randomForest, rrBLUP

**Encoding** UTF-8

**LazyData** true

**RoxygenNote** 6.1.1

**NeedsCompilation** no

**Repository** CRAN

**Date/Publication** 2019-10-24 16:00:05 UTC

## R topics documented:

---

`STGS.ann` *Genomic Selection using Artificial Neural Networks(ANN)*

---

## Description

Calculates the Genomic Estimated Breeding Value based on ANN method.

## Usage

```
STGS.ann(X, Y, r)
```

## Arguments

| | |
|---|---|
| `X` | X is a design matrix of marker genotype of size n×p where n are no of Individuals under study (i.e. genotype, lines) and p are no of markers. |
| `Y` | Y is a vector of individuals of size n×1. |
| `r` | fraction of testing data (ranges from (0-1)) used during model fitting (suppose if one want to use 75% of data for model training and remaining 25% for model testing so one has to define r=0.25). |

## Details

This function fits model by dividing data into two part i.e. training sets and testing sets. Former one is used to build the models and later one for performance evaluation. The performance of model is evaluated by calculating model accuracy i.e. pearson correlation coefficient between actual phenotypic value and predicted phenotypic value. Whole procedures is repeated 25 times and accuracy is averaged.

## Value

$fit meta-data of ANN model fitting

$Pred GEBV's for genotype under study

$Accuracy model accuracy i.e. pearson correlation coefficient between actual phenotypic value and predicted phenotypic value

## References

Foresee, F. D., and M. T. Hagan. 1997. "Gauss-Newton approximation to Bayesian regularization", Proceedings of the 1997 International Joint Conference on Neural Networks.

MacKay, D. J. C. 1992. "Bayesian interpolation", Neural Computation, vol. 4, no. 3, pp. 415-447.

Nguyen, D. and Widrow, B. 1990. "Improving the learning speed of 2-layer neural networks by choosing initial values of the adaptive weights", Proceedings of the IJCNN, vol. 3, pp. 21-26.

Paulino Perez Rodriguez and Daniel Gianola (2018). brnn: Bayesian Regularization for Feed-Forward Neural Networks. R package version 0.7. https://CRAN.R-project.org/package=brnn.

## Examples

```
library(STGS)

data(wheat_data)

X<-wheat_data[,1:100]

Y<-as.data.frame(wheat_data[,101])

r<-0.25

STGS.ann(X,Y,r)
```

---

`STGS.blup` *Genomic Selection using Best Linear Unbiased Prediction (BLUP).*

---

## Description

Calculates the Genomic Estimated Breeding Value based on BLUP.

## Usage

```
STGS.blup(X, Y, r)
```

## Arguments

| | |
|---|---|
| `X` | X is a design matrix of marker genotype of size n×p where n are no of Individuals under study (i.e. genotype, lines) and p are no of markers. |
| `Y` | Y is a vector of individuals of size n×1. |
| `r` | fraction of testing data (ranges from (0-1)) used during model fitting (suppose if one want to use 75% of data for model training and remaining 25% for model testing so one has to define r=0.25). |

## Details

This function fits model by dividing data into two part i.e. training sets and testing sets. Former one is used to build the models and later one for performance evaluation. The performance of model is evaluated by calculating model accuracy i.e. pearson correlation coefficient between actual phenotypic value and predicted phenotypic value. Whole procedures is repeated 25 times and accuracy is averaged.

## Value

$Vu variance of random effect i.e. u

$Ve error variance

$beta estimate of fixed effects i.e. BLUE

$u estimate of random effects i.e. BLUP(u)

$LL maximized log-likelihood

$Pred GEBVs for genotype under study

$Accuracy model accuracy i.e. pearson correlation coefficient between actual phenotypic value and predicted phenotypic value

## References

Endelman, J.B. 2011. Ridge regression and other kernels for genomic selection with R package rrBLUP. Plant Genome 4:250-255.

Searle, S.R., G. Casella and C.E. McCulloch. 1992. Variance Components. John Wiley, Hoboken.

## Examples

```
library(STGS)

data(wheat_data)

X<-wheat_data[,1:100]

Y<-as.data.frame(wheat_data[,101])

r<-0.25

STGS.blup(X,Y,r)
```

STGS.lasso *Genomic Selection using Least Absolute Shrinkage and Selection Operator (LASSO)*

## Description

Calculates the Genomic Estimated Breeding Value using LASSO method.

## Usage

```
STGS.lasso(X, Y, r)
```

## Arguments

| | |
|---|---|
| X | X is a design matrix of marker genotype of size n×p where n are no of Individuals under study (i.e. genotype, lines) and p are no of markers. |
| Y | Y is a vector of individuals of size n×1. |
| r | fraction of testing data (ranges from (0-1)) used during model fitting (suppose if one want to use 75% of data for model training and remaining 25% for model testing so one has to define r=0.25). |

## Details

This function fits model by dividing data into two part i.e. training sets and testing sets. Former one is used to build the models and later one for performance evaluation. The performance of model is evaluated by calculating model accuracy i.e. pearson correlation coefficient between actual phenotypic value and predicted phenotypic value. Whole procedures is repeated 25 times and accuracy is averaged.

## Value

$fit Lists various coeffecient assocaited to LASSO model fitting

$Pred GEBV's for genotype under study

$Accuracy model accuracy i.e. pearson correlation coefficient between actual phenotypic value and predicted phenotypic value

## References

Tibshirani, R. 1996. Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Series B (Methodological). 267–288.

Searle, S.R., G. Casella and C.E. McCulloch. 1992. Variance Components. John Wiley, Hoboken.

Friedman, J., Hastie, T. and Tibshirani, R. (2008) Regularization Paths for Generalized Linear Models via Coordinate Descent, https://web.stanford.edu/~hastie/Papers/glmnet.pdf Journal of Statistical Software, Vol. 33(1), 1-22 Feb 2010 http://www.jstatsoft.org/v33/i01/.

Jerome Friedman, Trevor Hastie, Robert Tibshirani (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software, 33(1), 1-22. URL http://www.jstatsoft.org/v33/i01/ .

## Examples

```
library(STGS)

data(wheat_data)

X<-wheat_data[,1:100]

Y<-as.data.frame(wheat_data[,101])

r<-0.25

STGS.lasso(X,Y,r)
```

---

STGS.rf *Genomic Selection using Random Forest*

---

## Description

Calculates the Genomic Estimated Breeding Value by using Random Forest method.

## Usage

```
STGS.rf(X, Y, r)
```

## Arguments

| | |
|---|---|
| X | X is a design matrix of marker genotype of size n×p where n are no of Individuals under study (i.e. genotype, lines) and p are no of markers. |
| Y | Y is a vector of individuals of size n×1. |
| r | fraction of testing data (ranges from (0-1)) used during model fitting (suppose if one want to use 75% of data for model training and remaining 25% for model testing so one has to define r=0.25). |

## Details

This function fits model by dividing data into two part i.e. training sets and testing sets. Former one is used to build the models and later one for performance evaluation. The performance of model is evaluated by calculating model accuracy i.e. pearson correlation coefficient between actual phenotypic value and predicted phenotypic value. Whole procedures is repeated 25 times and accuracy is averaged.

## Value

$Pred GEBV's for genotype under study

$Accuracy model accuracy i.e. pearson correlation coefficient between actual phenotypic value and predicted phenotypic value

## References

Breiman, L. (2001), Random Forests, Machine Learning 45(1), 5-32.

Breiman, L (2002), "Manual On Setting Up, Using, And Understanding Random Forests V3.1",https://www.stat.berkeley.edu

A. Liaw and M. Wiener (2002). Classification and Regression by randomForest. R News 2(3), 18-22

## Examples

```
library(STGS)

data(wheat_data)

X<-wheat_data[,1:100]

Y<-as.data.frame(wheat_data[,101])

r<-0.25

STGS.rf(X,Y,r)
```

---

# STGS.rr *Genomic Selection using Ridge Regression (RR)*

---

## Description

Calculates the Genomic Estimated Breeding Value using RR.

## Usage

```
STGS.rr(X, Y, r)
```

## Arguments

| | |
|---|---|
| X | X is a design matrix of marker genotype of size n×p where n are no of Individuals under study (i.e. genotype, lines) and p are no of markers. |
| Y | Y is a vector of individuals of size n×1. |
| r | fraction of testing data (ranges from (0-1)) used during model fitting (suppose if one want to use 75% of data for model training and remaining 25 for model testing so one has to define r=0.25). |

## Details

This function fits model by dividing data into two part i.e. training sets and testing sets. Former one is used to build the models and later one for performance evaluation. The performance of model is evaluated by calculating model accuracy i.e. pearson correlation coefficient between actual phenotypic value and predicted phenotypic value. Whole procedures is repeated 25 times and accuracy is averaged.

## Value

$bhat estimate of marker effects

$Pred GEBV's for genotype under study

$Accuracy model accuracy i.e. pearson correlation coefficient between actual phenotypic value and predicted phenotypic value

## References

de los Campos, G., and P. P. Rodriguez, 2010 BLR: Bayesian Linear Regression. R package version 1.2. http://CRAN.R-project.org/package=BLR.

de los Campos, G., H. Naya, D. Gianola, J. Crossa, A. Legarra et al., 2009 Predicting quantitative traits with regression models for dense molecular markers and pedigrees. Genetics 182(1): 375–385.

## Examples

```
library(STGS)

data(wheat_data)

X<-wheat_data[,1:100]

Y<-as.data.frame(wheat_data[,101])

r<-0.25

STGS.rr(X,Y,r)
```

---

# STGS.svm *Genomic Selection using Support Vector Machine (SVM)*

---

## Description

Calculates the Genomic Estimated Breeding Value based on SVM method.

## Usage

```
STGS.svm(X, Y, r)
```

## Arguments

| | |
|---|---|
| X | X is a design matrix of marker genotype of size n×p where n are no of Individuals under study (i.e. genotype, lines) and p are no of markers. |
| Y | Y is a vector of individuals of size n×1. |
| r | fraction of testing data (ranges from (0-1)) used during model fitting (suppose if one want to use 75% of data for model training and remaining 25% for model testing so one has to define r=0.25). |

## Details

This function fits model by dividing data into two part i.e. training sets and testing sets. Former one is used to build the models and later one for performance evaluation. The performance of model is evaluated by calculating model accuracy i.e. pearson correlation coefficient between actual phenotypic value and predicted phenotypic value. Whole procedures is repeated 25 times and accuracy is averaged.

## Value

$fit List various coeffecient associated with SVM model fitting

$Pred GEBV's for genotype under study

$Accuracy model accuracy i.e. pearson correlation coefficient between actual phenotypic value and predicted phenotypic value

## References

Vapnik, V., 1995. The Nature of Statistical Learning Theory, Ed. 2. Springer, New York.

Vapnik, V., and A. Vashist, 2009. A new learning paradigm: Learning using privileged information. Neural Networks 22: 544–557.

Alexandros Karatzoglou, Alex Smola, Kurt Hornik, Achim Zeileis (2004). kernlab - An S4 Package for Kernel Methods in R. Journal of Statistical Software 11(9), 1-20. URL http://www.jstatsoft.org/v11/i09/ .

## Examples

```
library(STGS)

data(wheat_data)

X<-wheat_data[,1:100]

Y<-as.data.frame(wheat_data[,101])

r<-0.25

STGS.svm(X,Y,r)
```

---

# wheat_data *Genotyping and phenotypic dataset for wheat*

---

## Description

Dataset used in this study is subset from CIMMYT 599 wheat lines. In our sample dataset it has one response for 50 lines genotyped for 100 markers.

## Usage

```
data("wheat_data")
```

## Format

A data frame with 50 rows as genotypes with 101 columns (i.e. First 100 columns contains information of genotyped markers and last column represent data of phenotypic trait under study).

## Details

This dataset is created by taking 50 wheat lines genotyped for 100 Markers as a subset from CIMMYT 599 wheat lines. Wheat lines were genotyped using 1447 Diversity Array Technology markers generated by Triticarte Pty. Ltd. (Canberra, Australia; http://www.triticarte.com.au). These markers may take two different values i.e. their presence (1) or absence (0). Phenotypic trait used under this study is grain yield.

## Source

International Maize and Wheat Improvement Center (CIMMYT), Mexico.

## Examples

```
library(STGS)

data(wheat_data)

X<-wheat_data[,1:100] #### Extracting Genotype

Y<-as.data.frame(wheat_data[,101]) #### Extracting Phenotype
```

# Index