Package 'IGST'

January 31, 2020

Type Package

Title Informative Gene Selection Tool

Version 0.1.0

Author Nitesh Kumar Sharma, Dwijesh Chandra Mishra, Neeraj Budhlakoti and Md. Samir Farooqi

Maintainer Nitesh Kumar Sharma <sharmanitesh.iasri@gmail.com>

Description Mining informative genes with certain biological meanings are important for clinical diagnosis of disease and discovery of disease mechanisms in plants and animals. This process involves identification of relevant genes and removal of redundant genes as much as possible from a whole gene set. This package selects the informative genes related to a specific trait using gene expression dataset. These trait specific genes are considered as informative genes. This package returns the informative gene set from the high dimensional gene expression data using a combination of methods SVM and MRMR (for feature selection) with bootstrapping procedure.

Depends R (>= 3.5)

Imports e1071, BootMRMR

License GPL-3

Encoding UTF-8

LazyData true

NeedsCompilation no

Repository CRAN

Date/Publication 2020-01-31 16:00:06 UTC

R topics documented:

`IGST.bootmrmrsvm.pval.cutoff`

Gene selection using SVM and MRMR feature selection techniques with bootstrapping procedure

Description

The informative gene set which have maximum relevance with target class or trait and minimum redundancy among genes based on statistical significance values computed from the combination of SVM and MRMR feature selection techniques with bootstrapping procedure.

Usage

```
IGST.bootmrmrsvm.pval.cutoff (x, y, s, Q, v, re)
```

Arguments

`x`	x is a n by p data frame of gene expression values where rows represent genes and columns represent samples. Each cell entry represents the expression level of a gene in a sample or subject (row names of x as gene names or gene ids).
`y`	y is a p by 1 numeric vector with entries 1 or -1 representing sample labels, where, 1\/-1 represents the sample label of subjects orsamples for stress/control condition(for two class problems).
`s`	s is a scalar representing the size of the informative gene set to be obtained.
`Q`	Q is a scalar representing the quartile value of the rank scores of genes (lies within 1\/N to 1), usually the second quartile, i.e. 0.5 or third quartile i.e. 0.75 may be taken.
`v`	v is a scalar representing the weightage of a method and must be within 0 and 1.
`re`	re is a scalar representing the number of bootstrap generated, re must be sufficiently large (i.e. number of times bootstrap samples are generated.

Value

The function returns a list of the genes\/informative gene set which are highly relevant to the particular trait or condition under investigation and minimal redundant among themselves without any spurious association among the genes.

Author(s)

Nitesh Kumar Sharma, Dwijesh Chandra Mishra, Neeraj Budhlakoti and Md. Samir Farooqi

References

Das, S., Rai, A., Mishra, D. C., & Rai, S. N. (2018). Statistical approach for selection of biologically informative genes. Gene, 655, 71-83.

Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. Machine Learning, 46(1-3), 389-422.

Wang, J., Chen, L., Wang, Y., Zhang, J., Liang, Y., & Xu, D. (2013). A computational systems biology study for understanding salt tolerance mechanism in rice. PLoS One, 8(6), e64929.

Examples

```
#################################
library(IGST)
data(rice_cold)
x<-rice_cold[-1,]
y<-rice_cold[1,]
y<-as.matrix(y)
y<-as.vector(y)
s<-10
Q<-0.5
v<-0.5
re<-10
IGST.bootmrmrsvm.pval.cutoff (x, y, s, Q, v, re)
```

IGST.bootmrmrsvm.weight.cutoff

Identification of informative gene set based on weights obtained from SVM and MRMR feature selection technique with bootstrapping procedure

---

Description

The function enables to find set of informative genes based on weights which are obtained by maximizing the relevancy of genes with classes or condition or trait and minimizing the redundancy among genes from the combination of SVM and MRMR feature selection techniques with bootstrapping procedure.

Usage

```
IGST.bootmrmrsvm.weight.cutoff (x, y, s, v, re)
```

Arguments

x	x is a n by p data frame of gene expression values where rows represent genes and columns represent samples. Each cell entry represents the expression level of a gene in a sample or subject (row names of x as gene names or gene ids).
y	y is a p by 1 numeric vector with entries 1 or -1 representing sample labels, where, 1 or -1 represents the sample label of subjects or samples for stress or control condition(for two class problems).

s s is a scalar representing the size of the informative gene set to be obtained.

v v is a scalar representing the weightage of a method and must be within 0 and 1.

re re is a scalar representing the number of bootstrap generated, re must be sufficiently large (i.e. number of times bootstrap samples are generated.

Value

The function returns a set of genes, which are highly informative to the trait or condition under consideration based on weights given by the combination of SVM and MRMR feature selection techniques with bootstrapping procedure.

Author(s)

Nitesh Kumar Sharma, Dwijesh Chandra Mishra, Neeraj Budhlakoti and Md. Samir Farooqi

References

Ding, C., & Peng, H. (2005). Minimum redundancy feature selection from microarray gene expression data. Journal of Bioinformatics and Computational Biology, 3(02), 185-205.

Mishra DC, Kumar S, Lal SB, Saha A, Chaturvedi KK, Budhlakoti N, et al.(2018) TAGPT: A Web Server for Prediction of Trait Associated Genes using Gene Expression Data. Annals of Genetics and Genetic Disorder. 1(1): 1003.

Examples

```
#################################
library(IGST)
data(rice_cold)
x<-rice_cold[-1,]
y<-rice_cold[1,]
y<-as.matrix(y)
y<-as.vector(y)
s<-10
#Q<-0.5
v<-0.5
re<-10
IGST.bootmrmrsvm.weight.cutoff (x, y, s, v, re)
```

IGST.pval.bootmrmrsvm *Computation of statistical significance values for genes using SVM and MRMR feature selection technique with bootstrapping procedure*

Description

The function computes the statistical significance values for the genes from the non-parametric test "H0: i-th gene is not informative against H1: i-th gene is informative" for selection of informative genes using SVM and MRMR feature selection technique with bootstrapping procedure.

Usage

```
IGST.pval.bootmrmrsvm(x, y, re, Q, v)
```

Arguments

`x` x is a n by p data frame of gene expression values where rows represent genes and columns represent samples. Each cell entry represents the expression level of a gene in a sample or subject (row names of x as gene names or gene ids).

`y` y is a p by 1 numeric vector with entries 1 or -1 representing sample labels, where, 1 or -1 represents the sample label of subjects or samples for stress or control condition(for two class problems).

`Q` Q is a scalar representing the quartile value of the rank scores of genes (lies within 1\VN to 1), usually the second quartile, i.e. 0.5 or third quartile i.e. 0.75 may be taken.

`v` v is a scalar representing the weightage of a method and must be within 0 and 1.

`re` re is a scalar representing the number of bootstrap generated, re must be sufficiently large (i.e. number of times bootstrap samples are generated.

Value

The function returns a vector of p-values for all the genes from the given statistical test in the dataset using SVM and MRMR feature selection technique with bootstrapping procedure.

Author(s)

Nitesh Kumar Sharma, Dwijesh Chandra Mishra, Neeraj Budhlakoti and Md. Samir Farooqi

References

Das, S., Rai, A., Mishra, D. C., & Rai, S. N. (2018). Statistical approach for selection of biologically informative genes. Gene, 655, 71-83.

Examples

```
#################################
library(IGST)
data(rice_cold)
x<-rice_cold[-1,]
y<-rice_cold[1,]
y<-as.matrix(y)
y<-as.vector(y)
#s<-10
Q<-0.5
v<-0.5
re<-10
IGST.pval.bootmrmrsvm (x, y, re, Q, v)
```

IGST.weight.bootmrmrsvm

Computation of weights for informative genes or gene set selection using SVM and MRMR feature selection technique with bootstrapping procedure

Description

The function computes the weights associated with each genes for a given dataset using SVM and MRMR feature selection technique with bootstrapping procedure.

Usage

```
IGST.weight.bootmrmrsvm (x, y, re, v)
```

Arguments

`x` x is a n by p data frame of gene expression values where rows represent genes and columns represent samples. Each cell entry represents the expression level of a gene in a sample or subject (row names of x as gene names or gene ids).

`y` y is a p by 1 numeric vector with entries 1 or -1 representing sample labels, where, 1\/-1 represents the sample label of subjects or samples for stress or control condition(for two class problems).

`v` v is a scalar representing the weightage of a method and must be within 0 and 1.

`re` re is a scalar representing the number of bootstrap generated, re must be sufficiently large (i.e. number of times bootstrap samples are generated.

Value

The function returns a vector of weights associated with each genes computed from SVM and MRMR feature selection technique with bootstrapping procedure for a given dataset.

Author(s)

Nitesh Kumar Sharma, Dwijesh Chandra Mishra, Neeraj Budhlakoti and Md. Samir Farooqi

References

Wang, J., Chen, L., Wang, Y., Zhang, J., Liang, Y., & Xu, D. (2013). A computational systems biology study for understanding salt tolerance mechanism in rice. PLoS One, 8(6), e64929.

Ding, C., & Peng, H. (2005). Minimum redundancy feature selection from microarray gene expression data. Journal of Bioinformatics and Computational Biology, 3(02), 185-205.

Mishra DC, Kumar S, Lal SB, Saha A, Chaturvedi KK, Budhlakoti N, et al.(2018) TAGPT: A Web Server for Prediction of Trait Associated Genes using Gene Expression Data. Annals of Genetics and Genetic Disorder. 1(1): 1003.

Examples

```
##################################
library(IGST)
data(rice_cold)
x<-rice_cold[-1,]
y<-rice_cold[1,]
y<-as.matrix(y)
y<-as.vector(y)
#s<-10
#Q<-0.5
v<-0.5
re<-10
IGST.weight.bootmrmrsvm (x, y, re, v)
```

`rice_cold` *A gene expression dataset of rice under cold stress*

Description

This data has gene expression values of 250 genes over 36 samples or subjects for a cold vs. control study in rice. These 36 samples belong to either of cold stress or control condition (two class problem). This gene expression data is balanced type as the first 18 samples are under cold stress and the later 18 samples are under control condition. The first row of the data contains the samples or subjects labels with entries are 1 and -1, where the label '1' and '-1' represent samples generated under cold stress and control condition respectively.

Usage

```
data("rice_cold")
```

Format

A data frame with 250 rows as genes with 36 columns as samples or subjects. Each column (sample) represent the gene expression values of genes. Each column as microarray samples with labels -1 or 1 represents control or cold stress respectively.

Details

The data is created by taking 250 genes from the large number of genes from NCBI GEO database. The rows are the genes and columns are the samples or subjects. The first half of the samples or subjects are generated under cold stress condition and other half under control condition. The first row of the data contains the samples/subjects labels with entries are 1 and -1, where the label '1' and '-1' represent samples generated under cold stress and control condition respectively.

Source

Gene Expression Omnibus: NCBI gene expression and hybridization array data repository.ncbi.nlm.nih.gov/geo/.

Examples

```
####################################
library(IGST)
data(rice_cold)
```

Index