Package 'BioM2'

May 16, 2024

Title Biologically Explainable Machine Learning Framework

Version 1.0.6

Author Shunjie Zhang and Junfang Chen

Maintainer Shunjie Zhang <zhang.shunjie@qq.com>

Description Biologically Explainable Machine Learning Framework for Phenotype Prediction using omics data described in Chen and Schwarz (2017) <doi:10.48550/arXiv.1712.00336>.Identifying reproducible and interpretable biological patterns from high-dimensional omics data is a critical factor in understanding the risk mechanism of complex disease. As such, explainable machine learning can offer biological insight in addition to personalized risk scoring.In this process, a feature space of biological pathways will be generated, and the feature space can also be subsequently analyzed using WGCNA (Described in Horvath and Zhang (2005) <doi:10.2202/1544-6115.1128> and Langfelder and Horvath (2008) <doi:10.1186/1471-2105-9-559>) methods.

License MIT + file LICENSE

Encoding UTF-8

RoxygenNote 7.2.3

Imports WGCNA, mlr3, CMplot, ggsci, ROCR, caret, ggplot2, ggpubr, viridis, ggthemes, ggstatsplot, htmlwidgets, jiebaR, mlr3verse, parallel, uwot, webshot, wordcloud2,ggforce, igraph, ggnetwork

Depends R (>= 4.1.0)

LazyData true

NeedsCompilation no

Repository CRAN

Date/Publication 2024-05-16 10:00:02 UTC

R topics documented:

`AddUnmapped` *Add unmapped probe*

Description

Add unmapped probe

Usage

```
AddUnmapped(
  train = NULL,
  test = NULL,
  Unmapped_num = NULL,
  Add_FeartureSelection_Method = "wilcox.test",
  anno = NULL,
  len = NULL,
  verbose = TRUE,
  cores = 1
)
```

Arguments

`train`	The input training dataset. The first column is the label or the output. For binary classes, 0 and 1 are used to indicate the class member.
`test`	The input test dataset. The first column is the label or the output. For binary classes, 0 and 1 are used to indicate the class member.
`Unmapped_num`	The number of unmapped probes.

Add_FeartureSelection_Method	Feature selection methods. Available options are c('cor', 'wilcox.test').
anno	The annotation data stored in a data.frame for probe mapping. It must have at least two columns named 'ID' and 'entrezID'. (For details, please refer to data(data("MethylAnno"))
len	The number of unmapped probes
verbose	Whether to print running process information to the console
cores	The number of cores used for computation.

Value

Matrix of unmapped probes

---

baseModel *Prediction by Machine Learning*

Description

Prediction by Machine Learning with different learners (From 'mlr3')

Usage

```
baseModel(
  trainData,
  testData,
  predMode = "probability",
  classifier,
  paramlist = NULL,
  inner_folds = 10
)
```

Arguments

trainData	The input training dataset. The first column is the label or the output. For binary classes, 0 and 1 are used to indicate the class member.
testData	The input test dataset. The first column is the label or the output. For binary classes, 0 and 1 are used to indicate the class member.
predMode	The prediction mode. Available options are c('probability', 'classification').
classifier	Learners in mlr3
paramlist	Learner parameters
inner_folds	k-fold cross validation (Only supported when testData = NULL)

Value

The predicted output for the test data.

Author(s)

Shunjie Zhang

Examples

```
library(mlr3verse)
library(caret)
library(BioM2)
data=MethylData_Test
set.seed(1)
part=unlist(createDataPartition(data$label,p=0.8))#Split data
predict=baseModel(trainData=data[part,1:10],
                  testData=data[-part,1:10],
                 classifier = 'svm')#Use 10 features to make predictions,Learner uses svm
```

BioM2 *Biologically Explainable Machine Learning Framework*

Description

Biologically Explainable Machine Learning Framework

Usage

```
BioM2(
  TrainData = NULL,
  TestData = NULL,
  pathlistDB = NULL,
  FeatureAnno = NULL,
  resampling = NULL,
  nfolds = 5,
  classifier = "liblinear",
  predMode = "probability",
  PathwaySizeUp = 200,
  PathwaySizeDown = 20,
  MinfeatureNum_pathways = 10,
  Add_UnMapped = TRUE,
  Unmapped_num = 300,
  Add_FeartureSelection_Method = "wilcox.test",
  Inner_CV = TRUE,
  inner_folds = 10,
  Stage1_FeartureSelection_Method = "cor",
  cutoff = 0.3,
  Stage2_FeartureSelection_Method = "RemoveHighcor",
  cutoff2 = 0.85,
  classifier2 = NULL,
```

```
  target = "predict",
  p.adjust.method = "fdr",
  save_pathways_matrix = FALSE,
  cores = 1,
  verbose = TRUE
)
```

Arguments

TrainData	The input training dataset. The first column is the label or the output. For binary classes, 0 and 1 are used to indicate the class member.
TestData	The input test dataset. The first column is the label or the output. For binary classes, 0 and 1 are used to indicate the class member.
pathlistDB	A list of pathways with pathway IDs and their corresponding genes ('entrezID' is used). For details, please refer to (data("GO2ALLEGS_BP"))
FeatureAnno	The annotation data stored in a data.frame for probe mapping. It must have at least two columns named 'ID' and 'entrezID'. (For details, please refer to data(data("MethylAnno"))
resampling	Resampling in mlr3verse.
nfolds	k-fold cross validation (Only supported when TestData = NULL)
classifier	Learners in mlr3
predMode	The prediction mode. Available options are c('probability', 'classification').
PathwaySizeUp	The upper-bound of the number of genes in each biological pathways.
PathwaySizeDown	The lower-bound of the number of genes in each biological pathways.
MinfeatureNum_pathways	The minimal defined pathway size after mapping your own data to pathlistDB(KEGG database/GO database).
Add_UnMapped	Whether to add unmapped probes for prediction
Unmapped_num	The number of unmapped probes
Add_FeartureSelection_Method	Feature selection methods.
Inner_CV	Whether to perform a k-fold verification on the training set.
inner_folds	k-fold verification on the training set.
Stage1_FeartureSelection_Method	Feature selection methods.
cutoff	The cutoff used for feature selection threshold. It can be any value between 0 and 1.
Stage2_FeartureSelection_Method	Feature selection methods.
cutoff2	The cutoff used for feature selection threshold. It can be any value between 0 and 1.
classifier2	Learner for stage 2 prediction(if classifier2==NULL,then it is the same as the learner in stage 1.)

target	Is it used to predict or explore potential biological mechanisms? Available options are c('predict', 'pathways').
p.adjust.method	p-value adjustment method.(holm", "hochberg", "hommel", "bonferroni", "BH", "BY",
save_pathways_matrix	Whether to output the path matrix file
cores	The number of cores used for computation.
verbose	Whether to print running process information to the console

Value

A list containing prediction results and prediction result evaluation

Examples

```
library(mlr3verse)
library(caret)
library(parallel)
library(BioM2)
data=MethylData_Test
set.seed(1)
part=unlist(createDataPartition(data$label,p=0.8))
Train=data[part,]
Test=data[-part,]
pathlistDB=GO2ALLEGS_BP
FeatureAnno=MethylAnno


pred=BioM2(TrainData = Train,TestData = Test,
           pathlistDB=pathlistDB,FeatureAnno=FeatureAnno,
           classifier='svm',nfolds=5,
           PathwaySizeUp=25,PathwaySizeDown=20,MinfeatureNum_pathways=10,
           Add_UnMapped='Yes',Unmapped_num=300,
           Inner_CV='None',inner_folds=5,
           Stage1_FeartureSelection_Method='cor',cutoff=0.3,
           Stage2_FeartureSelection_Method='None',
           target='predict',cores=1
)#(To explore biological mechanisms, set target='pathways')
```

FindParaModule *Find suitable parameters for partitioning pathways modules*

Description

Find suitable parameters for partitioning pathways modules

Usage

```
FindParaModule(
  pathways_matrix = NULL,
  control_label = NULL,
  minModuleSize = seq(10, 20, 5),
  mergeCutHeight = seq(0, 0.3, 0.1),
  minModuleNum = 20,
  power = NULL,
  exact = TRUE
)
```

Arguments

pathways_matrix	A pathway matrix generated by the BioM2(target='pathways') function.
control_label	The label of the control group (A single number, factor, or character)
minModuleSize	minimum module size for module detection. Detail for WGCNA::blockwiseModules()
mergeCutHeight	dendrogram cut height for module merging. Detail for WGCNA::blockwiseModules()
minModuleNum	Minimum total number of modules detected
power	soft-thresholding power for network construction. Detail for WGCNA::blockwiseModules()
exact	Whether to divide GO pathways more accurately

Value

A list containing recommended parameters

GO2ALLEGS_BP *An example about pathlistDB*

Description

An example about pathlistDB

Format

A list :

...

Details

A list of pathways with pathway IDs and their corresponding genes ('entrezID' is used).

GO_Ancestor *Pathways in the GO database and their Ancestor*

Description

Inclusion relationships between pathways

Format

A data frame :

...

Details

In the GO database, each pathway will have its own ancestor pathway. Map pathways in GO database to about 20 common ancestor pathways.

Source

From GO.db

GO_Ancestor_exact *Pathways in the GO database and their Ancestor*

Description

Inclusion relationships between pathways

Format

A data frame :

...

Details

In the GO database, each pathway will have its own ancestor pathway. Map pathways in GO database to about 400 common ancestor pathways.

Source

From GO.db

MethylAnno *An example about FeatureAnno for methylation data*

Description

An example about FeatureAnno for methylation data

Format

A data frame :

...

Details

The annotation data stored in a data.frame for probe mapping. It must have at least two columns named 'ID' and 'entrezID'.

MethylData_Test *An example about TrainData/TestData for methylation data*

Description

An example about TrainData/TestData for methylation data MethylData_Test.

Format

A data frame :

...

Details

The first column is the label or the output. For binary classes, 0 and 1 are used to indicate the class member.

PathwaysModule *Delineate differential pathway modules with high biological interpretability*

Description

Delineate differential pathway modules with high biological interpretability

Usage

```
PathwaysModule(
  pathways_matrix = NULL,
  control_label = NULL,
  power = NULL,
  minModuleSize = NULL,
  mergeCutHeight = NULL,
  cutoff = 70,
  MinNumPathways = 5,
  p.adjust.method = "fdr",
  exact = TRUE
)
```

Arguments

pathways_matrix	A pathway matrix generated by the BioM2(target='pathways') function.
control_label	The label of the control group (A single number, factor, or character)
power	soft-thresholding power for network construction. Detail for WGCNA::blockwiseModules()
minModuleSize	minimum module size for module detection. Detail for WGCNA::blockwiseModules()
mergeCutHeight	dendrogram cut height for module merging. Detail for WGCNA::blockwiseModules()
cutoff	Thresholds for Biological Interpretability Difference Modules
MinNumPathways	Minimum number of pathways included in the biologically interpretable difference module
p.adjust.method	p-value adjustment method.(holm", "hochberg", "hommel", "bonferroni", "BH", "BY",
exact	Whether to divide GO pathways more accurately

Value

A list containing differential module results that are highly biologically interpretable

PlotCorModule *Correlalogram for Biological Differences Modules*

Description

Correlalogram for Biological Differences Modules

Usage

```
PlotCorModule(
  PathwaysModule_obj = NULL,
  alpha = 0.7,
  begin = 0.2,
  end = 0.9,
  option = "C",
  family = "serif"
)
```

Arguments

PathwaysModule_obj	Results produced by PathwaysModule()
alpha	The alpha transparency, a number in (0,1). Detail for scale_fill_viridis()
begin	The (corrected) hue in (0,1) at which the color map begins. Detail for scale_fill_viridis().
end	The (corrected) hue in (0,1) at which the color map ends. Detail for scale_fill_viridis()
option	A character string indicating the color map option to use. Detail for scale_fill_viridis()
family	calligraphic style

Value

a ggplot object

PlotPathFearture *Visualisation of significant pathway-level features*

Description

Visualisation of significant pathway-level features

Usage

```
PlotPathFearture(
  BioM2_pathways_obj = NULL,
  pathlistDB = NULL,
  top = 10,
  p.adjust.method = "none",
  begin = 0.1,
  end = 0.9,
  alpha = 0.9,
  option = "C",
  seq = 1
)
```

Arguments

BioM2_pathways_obj	Results produced by BioM2(,target='pathways')
pathlistDB	A list of pathways with pathway IDs and their corresponding genes ('entrezID' is used). For details, please refer to (data("GO2ALLEGS_BP"))
top	Number of significant pathway-level features visualised
p.adjust.method	p-value adjustment method.(holm", "hochberg", "hommel", "bonferroni", "BH", "BY","fdr","none")
begin	The (corrected) hue in (0,1) at which the color map begins. Detail for scale_fill_viridis().
end	The (corrected) hue in (0,1) at which the color map ends. Detail for scale_fill_viridis()
alpha	The alpha transparency, a number in (0,1). Detail for scale_fill_viridis()
option	A character string indicating the color map option to use. Detail for scale_fill_viridis()
seq	Interval of x-coordinate

Value

a ggplot2 object

PlotPathInner *Visualisation Original features that make up the pathway*

Description

Visualisation Original features that make up the pathway

Usage

```
PlotPathInner(
  data = NULL,
  pathlistDB = NULL,
  FeatureAnno = NULL,
  PathNames = NULL,
  p.adjust.method = "none",
  save_pdf = FALSE,
  alpha = 1,
  cols = NULL
)
```

Arguments

data	The input omics data
pathlistDB	A list of pathways with pathway IDs and their corresponding genes ('entrezID' is used). For details, please refer to (data("GO2ALLEGS_BP"))
FeatureAnno	The annotation data stored in a data.frame for probe mapping. It must have at least two columns named 'ID' and 'entrezID'. (For details, please refer to data(data("MethylAnno"))
PathNames	A vector.A vector containing the names of pathways
p.adjust.method	p-value adjustment method.(holm", "hochberg", "hommel", "bonferroni", "BH", "BY","fdr","none")
save_pdf	Whether to save images in PDF format
alpha	The alpha transparency, a number in (0,1).
cols	palette (vector of colour names)

Value

a plot object

PlotPathNet *Network diagram of pathways-level features*

Description

Network diagram of pathways-level features

Usage

```
PlotPathNet(
  data = NULL,
  BioM2_pathways_obj = NULL,
  FeatureAnno = NULL,
  pathlistDB = NULL,
  PathNames = NULL,
  cutoff = 0.2,
  num = 20
)
```

Arguments

data	The input omics data
BioM2_pathways_obj	Results produced by BioM2()
FeatureAnno	The annotation data stored in a data.frame for probe mapping. It must have at least two columns named 'ID' and 'entrezID'. (For details, please refer to data(data("MethylAnno"))
pathlistDB	A list of pathways with pathway IDs and their corresponding genes ('entrezID' is used). For details, please refer to (data("GO2ALLEGS_BP"))
PathNames	A vector.A vector containing the names of pathways
cutoff	Threshold for correlation between features within a pathway
num	The first few internal features of each pathway that are most relevant to the phenotype

Value

a ggplot object

ShowModule *Display biological information within each pathway module*

Description

Display biological information within each pathway module

Usage

```
ShowModule(obj = NULL, ID_Module = NULL, exact = TRUE)
```

Arguments

obj	Results produced by PathwaysModule()
ID_Module	ID of the diff module
exact	Whether to divide GO pathways more accurately

Value

List containing biologically specific information within the module

---

`Stage1_FeartureSelection` *Stage 1 Fearture Selection*

Description

Stage 1 Fearture Selection

Usage

```
Stage1_FeartureSelection(
  Stage1_FeartureSelection_Method = "cor",
  data = NULL,
  cutoff = NULL,
  featureAnno = NULL,
  pathlistDB_sub = NULL,
  cores = 1,
  verbose = TRUE
)
```

Arguments

`Stage1_FeartureSelection_Method`	Feature selection methods. Available options are c(NULL, 'cor', 'wilcox.test', 'cor_rank', 'wilcox.test_rank').
`data`	The input training dataset. The first column is the label.
`cutoff`	The cutoff used for feature selection threshold. It can be any value between 0 and 1. Commonly used cutoffs are c(0.5, 0.1, 0.05, 0.01, etc.).
`featureAnno`	The annotation data stored in a data.frame for probe mapping. It must have at least two columns named 'ID' and 'entrezID'. (For details, please refer to data(data("MethylAnno"))
`pathlistDB_sub`	A list of pathways with pathway IDs and their corresponding genes ('entrezID' is used). For details, please refer to (data("GO2ALLEGS_BP"))
`cores`	The number of cores used for computation.
`verbose`	Whether to print running process information to the console

Value

A list of matrices with pathway IDs as the associated list member names.

Author(s)

Shunjie Zhang

Examples

```
library(parallel)
data=MethylData_Test
feature_pathways=Stage1_FeartureSelection(Stage1_FeartureSelection_Method='cor',
                data=data,cutoff=0,
                featureAnno=MethylAnno,pathlistDB_sub=GO2ALLEGS_BP,cores=1)
```

Stage2_FeartureSelection

Stage 2 Fearture Selection

Description

Stage 2 Fearture Selection

Usage

```
Stage2_FeartureSelection(
  Stage2_FeartureSelection_Method = "RemoveHighcor",
  data = NULL,
  label = NULL,
  cutoff = NULL,
  preMode = NULL,
  classifier = NULL,
  verbose = TRUE,
  cores = 1
)
```

Arguments

Stage2_FeartureSelection_Method	Feature selection methods. Available options are c(NULL, 'cor', 'wilcox.test', 'RemoveHighcor', 'RemoveLinear').
data	The input training dataset. The first column is the label.
label	The label of dataset
cutoff	The cutoff used for feature selection threshold. It can be any value between 0 and 1.
preMode	The prediction mode. Available options are c('probability', 'classification').
classifier	Learners in mlr3
verbose	Whether to print running process information to the console
cores	The number of cores used for computation.

Value

Column index of feature

Author(s)

Shunjie Zhang

`TransAnno` *An example about FeatureAnno for gene expression*

Description

An example about FeatureAnno for gene expression

Format

A data frame :

...

Details

The annotation data stored in a data.frame for probe mapping. It must have at least two columns named 'ID' and 'entrezID'.

`TransData_Test` *An example about TrainData/TestData for gene expression*

Description

An example about TrainData/TestData for gene expression MethylData_Test.

Format

A data frame :

...

Details

The first column is the label or the output. For binary classes, 0 and 1 are used to indicate the class member.

VisMultiModule *Visualisation of the results of the analysis of the pathway modules*

Description

Visualisation of the results of the analysis of the pathway modules

Usage

```
VisMultiModule(
  BioM2_pathways_obj = NULL,
  FindParaModule_obj = NULL,
  ShowModule_obj = NULL,
  PathwaysModule_obj = NULL,
  exact = TRUE,
  type_text_table = FALSE,
  text_table_theme = ttheme("mOrange"),
  volin = FALSE,
  control_label = 0,
  module = NULL,
  cols = NULL,
  n_neighbors = 8,
  spread = 1,
  min_dist = 2,
  target_weight = 0.5,
  size = 1.5,
  alpha = 1,
  ellipse = TRUE,
  ellipse.alpha = 0.2,
  theme = ggthemes::theme_base(base_family = "serif"),
  save_pdf = FALSE,
  width = 7,
  height = 7
)
```

Arguments

BioM2_pathways_obj	Results produced by BioM2(,target='pathways')
FindParaModule_obj	Results produced by FindParaModule()
ShowModule_obj	Results produced by ShowModule()
PathwaysModule_obj	Results produced by PathwaysModule()
exact	Whether to divide GO pathways more accurately

type_text_table
: Whether to display it in a table

text_table_theme
: The topic of this table.Detail for ggtexttable()

volin
: Can only be used when PathwaysModule_obj exists. (Violin diagram)

control_label
: Can only be used when PathwaysModule_obj exists. (Control group label)

module
: Can only be used when PathwaysModule_obj exists.(PathwaysModule ID)

cols
: palette (vector of colour names)

n_neighbors
: The size of local neighborhood (in terms of number of neighboring sample points) used for manifold approximation. Larger values result in more global views of the manifold, while smaller values result in more local data being preserved. In general values should be in the range 2 to 100.

spread
: The effective scale of embedded points. In combination with min_dist, this determines how clustered/clumped the embedded points are.

min_dist
: The effective minimum distance between embedded points. Smaller values will result in a more clustered/clumped embedding where nearby points on the manifold are drawn closer together, while larger values will result on a more even dispersal of points. The value should be set relative to the spread value, which determines the scale at which embedded points will be spread out.

target_weight
: Weighting factor between data topology and target topology. A value of 0.0 weights entirely on data, a value of 1.0 weights entirely on target. The default of 0.5 balances the weighting equally between data and target. Only applies if y is non-NULL.

size
: Scatter plot point size

alpha
: Alpha for ellipse specifying the transparency level of fill color. Use alpha = 0 for no fill color.

ellipse
: logical value. If TRUE, draws ellipses around points.

ellipse.alpha
: Alpha for ellipse specifying the transparency level of fill color. Use alpha = 0 for no fill color.

theme
: Default:theme_base(base_family = "serif")

save_pdf
: Whether to save images in PDF format

width
: image width

height
: image height

Value

a ggplot2 object

Index