Package 'MBMethPred'

December 13, 2022

Type Package

Title Medulloblastoma Subgroups Prediction

Version 0.1.0

Date 2022-11-23

Description Utilizing a combination of machine learning models (Random Forest, Naive Bayes, K-Nearest Neighbor, Support Vector Machines, Extreme Gradient Boosting, and Linear Discriminant Analysis) and a deep Artificial Neural Network model, 'MBMethPred' can predict medulloblastoma subgroups, including wingless (WNT), sonic hedgehog (SHH), Group 3, and Group 4 from methylation data.

Maintainer Edris Sharif Rahmani <rahmani.biotech@gmail.com>

BugReports https://github.com/sharifrahmanie/MBMethPred/issues

License GPL

Encoding UTF-8

LazyData true

LazyDataCompression xz

RoxygenNote 7.2.2

Imports stringr, ggplot2, parallel, caTools, caret, keras, MASS, Rtsne, SNFtool, class, dplyr, e1071, pROC, randomForest, readr, reshape2, reticulate, rgl, tensorflow, xgboost

Depends R (>= 3.5.0)

Suggests knitr, rmarkdown, testthat, utils, stats, scales

Config/testthat/edition 3

VignetteBuilder knitr

NeedsCompilation no

Author Edris Sharif Rahmani [aut, ctb, cre]
(<https://orcid.org/0000-0002-7899-1663>),
Ankita Sunil Lawarde [aut, ctb]
(<https://orcid.org/0000-0001-7572-4431>),
Vijayachitra Modhukur [aut, ctb]
(<https://orcid.org/0000-0002-7123-9903>)

Repository CRAN

Date/Publication 2022-12-13 12:20:02 UTC

R topics documented:

BoxPlot *Box plot*

Description

A function to draw a box plot for the DNA methylation dataset.

Usage

```
BoxPlot(File, Projname = NULL)
```

Arguments

`File`	The output of ReadMethylFile function.
`Projname`	A name used to name the plot. The default is null.

Value

A ggplot2 object

Examples

```
data <- Data2[1:10,]
data <- cbind(rownames(data), data)
colnames(data)[1] <- "ID"
BoxPlot(File = data)
```

ConfusionMatrix *Confusion matrix*

Description

A function to calculate the confusion matrix of the machine and deep learning models. It outputs Accuracy, Precision, Sensitivity, F1-Score, Specificity, and AUC_average.

Usage

```
ConfusionMatrix(y_true, y_pred)
```

Arguments

y_true	True labels
y_pred	Predicted labels

Value

A data frame

Examples

```
set.seed(1234)
data <- Data1[1:10,]
data$subgroup <- factor(data$subgroup)
fac <- ncol(data)
split <- caTools::sample.split(data[, fac], SplitRatio = 0.8)
training_set <- subset(data, split == TRUE)
test_set <- subset(data, split == FALSE)
rf <- randomForest::randomForest(x = training_set[-fac],
                                 y = training_set[, fac],
                                 ntree = 10)
y_pred <- predict(rf, newdata = test_set[-fac])
ConfusionMatrix(y_true = test_set[, fac],
                y_pred = y_pred)
```

`Data1` *Training data*

Description

Data1 is a medulloblastoma DNA methylation dataset from a GEO series (GSE85212) and focuses on 399 as the most important probes. This dataset is used to train and test the machine and deep learning models.

Value

A data frame

Source

https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE85212

Examples

```
data(Data1)
```

`Data2` *Data2*

Description

Data2 is a medulloblastoma DNA methylation dataset (GSE85212, 50 samples) including 10000 most variable probes used for similarity network fusion.

Value

A data frame

Source

https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE85212

References

Cavalli FMG, Remke M, Rampasek L, Peacock J et al. Intertumoral Heterogeneity within Medulloblastoma Subgroups. Cancer Cell 2017 Jun 12;31(6):737-754.e6. PMID: 28609654

Examples

```
data(Data2)
```

Data3 *Data3*

Description

Data3 is an expression dataset from primary medulloblastoma samples (GSE85217, 50 samples) used for similarity network fusion.

Value

A data frame

Source

https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE85217

References

Cavalli FMG, Remke M, Rampasek L, Peacock J et al. Intertumoral Heterogeneity within Medulloblastoma Subgroups. Cancer Cell 2017 Jun 12;31(6):737-754.e6. PMID: 28609654

Examples

```
data(Data3)
```

KNearestNeighborModel *K nearest neighbor model*

Description

A function to train a K nearest neighbor model to classify medulloblastoma subgroups using the DNA methylation dataset (Illumina Infinium HumanMethylation450). Prediction is followed by training if new data is provided.

Arguments

`SplitRatio`	Train and test split ratio. A value greater or equal to zero and less than one.
`CV`	The number of folds for cross-validation. It should be greater than one.
`K`	The number of nearest neighbors.
`NCores`	The number of cores for parallel computing.
`NewData`	A methylation data from the ReadMethylFile function.

Value

A list

Examples

```
set.seed(111)
knn <- KNearestNeighborModel(SplitRatio = 0.8,
                             CV = 3,
                             K = 3,
                             NCores = 1,
                             NewData = NULL)
```

LinearDiscriminantAnalysisModel *Linear discriminant analysis model*

Description

A function to train a linear discriminant analysis model to classify medulloblastoma subgroups using the DNA methylation dataset (Illumina Infinium HumanMethylation450). Prediction is followed by training if new data is provided.

Arguments

`SplitRatio` Train and test split ratio. A value greater or equal than zero and less than one.

`CV` The number of folds for cross validation. It should be greater than one.

`NCores` The number of cores for parallel computing.

`NewData` A methylation data from ReadMethylFile function.

Value

A list

Examples

```
set.seed(123)
lda <- LinearDiscriminantAnalysisModel(SplitRatio = 0.8,
                                       CV = 2,
                                       NCores = 1,
                                       NewData = NULL)
```

`ModelMetrics` *Model metrics*

Description

A function to extract the confusion matrix information.

Usage

```
ModelMetrics(Model)
```

Arguments

`Model`	A trained model.

Value

A data frame

Examples

```
xgboost <- XGBoostModel(SplitRatio = 0.6,
                        CV = 2,
                        NCores = 1,
                        NewData = NULL)
ModelMetrics(Model = xgboost)
```

`NaiveBayesModel` *Naive bayes model*

Description

A function to train a Naive Bayes model to classify medulloblastoma subgroups using the DNA methylation dataset (Illumina Infinium HumanMethylation450). Prediction is followed by training if new data is provided.

Arguments

`SplitRatio`	Train and test split ratio. A value greater or equal to zero and less than one.
`CV`	The number of folds for cross-validation. It should be greater than one.
`Threshold`	The threshold for deciding class probability. A value greater or equal to zero and less than one.
`NCores`	The number of cores for parallel computing.
`NewData`	A methylation data from ReadMethylFile function.

Value

A list

Examples

```
set.seed(123)
nb <- NaiveBayesModel(SplitRatio = 0.8,
                      CV = 2,
                      Threshold = 0.8,
                      NCores = 1,
                      NewData = NULL)
```

NeuralNetworkModel *Artificial neural network model*

Description

A function to train an artificial neural network model to classify medulloblastoma subgroups using the DNA methylation dataset (Illumina Infinium HumanMethylation450). Prediction is followed by training if new data is provided.

Arguments

`Epochs` The number of epochs.

`NewData` A methylation data from ReadMethylFile function.

`InstallTensorFlow`
Logical. Running this function for the first time, you need to install TensorFlow library (V 2.10-cpu). Default is TRUE.

Value

A list

Examples

```
## Not run:
set.seed(1234)
ann <- NeuralNetworkModel(Epochs = 100,
                          NewData = NULL,
                          InstallTensorFlow = TRUE)

## End(Not run)
```

NewDataPredictionResult *New data prediction result*

Description

A function to output the predicted medulloblastoma subgroups by trained models.

Usage

```
NewDataPredictionResult(Model)
```

Arguments

Model	A trained model

Value

A data frame

Examples

```
set.seed(10)
fac <- ncol(Data1)
NewData <- sample(data.frame(t(Data1[,-fac])),10)
NewData <- cbind(rownames(NewData), NewData)
colnames(NewData)[1] <- "ID"
xgboost <- XGBoostModel(SplitRatio = 0.6,
                        CV = 2,
                        NCores = 1,
                        NewData = NewData)
NewDataPredictionResult(Model = xgboost)
```

RandomForestModel *Random forest model*

Description

A function to train a random forest model to classify medulloblastoma subgroups using the DNA methylation dataset (Illumina Infinium HumanMethylation450). Prediction is followed by training if new data is provided.

Arguments

`SplitRatio` Train and test split ratio. A value greater or equal to zero and less than one.

`CV` The number of folds for cross-validation. It should be greater than one.

`NTree` The number of trees to be grown.

`NCores` The number of cores for parallel computing.

`NewData` A methylation data from ReadMethylFile function.

Value

A list

Examples

```
set.seed(21)
rf <- RandomForestModel(SplitRatio = 0.8,
                        CV = 3,
                        NTree = 10,
                        NCores = 1,
                        NewData = NULL)
```

`ReadMethylFile` *Input file for prediction*

Description

A function to read DNA methylation files can be used as the new data for prediction by every model.

Usage

```
ReadMethylFile(File)
```

Arguments

`File` A data frame with tsv or csv file extension. While the first column is CpG methylation probs, starting with cg and followed by a number, other columns are samples with methylation values. All columns should be named.

Value

A data frame

Examples

```
## Not run:
methyl <- ReadMethylFile(File = ”file.csv”)

## End(Not run)
```

ReadSNFData *Input file for similarity network fusion (SNF)*

Description

A function to read user-provided file feeding into the SNF function (from the SNFtools package).

Usage

```
ReadSNFData(File)
```

Arguments

`File`	A table with tsv or csv file extension. While the first column is a character vector (e.g., gene names or IDs), columns are samples. All columns should be named.

Value

A data frame

Examples

```
## Not run:
data <- ReadSNFData(File = "file.csv")

## End(Not run)
```

RLabels *RLabels*

Description

The actual labels from the medulloblastoma DNA methylation dataset (GSE85212, 50 samples) that was used for similarity network fusion.

Value

Factor

Source

https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE85212

References

Cavalli FMG, Remke M, Rampasek L, Peacock J et al. Intertumoral Heterogeneity within Medulloblastoma Subgroups. Cancer Cell 2017 Jun 12;31(6):737-754.e6. PMID: 28609654

Examples

```
data(RLabels)
```

SimilarityNetworkFusion *Similarity network fusion (SNF)*

Description

A function to perform SNF function (from SNFtool package) and output clusters.

Usage

```
SimilarityNetworkFusion(
  Files = NULL,
  NNeighbors,
  Sigma,
  NClusters,
  CLabels = NULL,
  RLabels = NULL,
  Niterations
)
```

Arguments

Files	A list of data frames created using the ReadSNFData function or matrices.
NNeighbors	The number of nearest neighbors.
Sigma	The variance for local model.
NClusters	The number of clusters.
CLabels	A string vector to name the clusters. Optional.
RLabels	The actual label of samples to calculate the Normalized Mutual Information (NMI) score. Optional.
Niterations	The number of iterations for the diffusion process.

Value

Factor

Examples

```
data(RLabels) # Real labels
data(Data2) # Methylation
data(Data3) # Gene expression
snf <- SimilarityNetworkFusion(Files = list(Data2, Data3),
                               NNeighbors  = 13,
                               Sigma = 0.75,
                               NClusters = 4,
                               CLabels = c("Group4", "SHH", "WNT", "Group3"),
                               RLabels = RLabels,
                               Niterations = 10)
snf
```

SupportVectorMachineModel *Support vector machine model*

Description

A function to train a support vector machine model to classify medulloblastoma subgroups using the DNA methylation dataset (Illumina Infinium HumanMethylation450). Prediction is followed by training if new data is provided.

Arguments

`SplitRatio`	Train and test split ratio. A value greater or equal to zero and less than one.
`CV`	The number of folds for cross-validation. It should be greater than one.
`NCores`	The number of cores for parallel computing.
`NewData`	A methylation data from the ReadMethylFile function.

Value

A list

Examples

```
set.seed(56)
svm <- SupportVectorMachineModel(SplitRatio = 0.8,
                                 CV = 3,
                                 NCores = 1,
                                 NewData = NULL)
```

TSNEPlot *t-SNE 3D plot*

Description

A function to draw a 3D t-SNE plot for DNA methylation dataset using the K-means clustering technique.

Usage

```
TSNEPlot(File, NCluster = 4)
```

Arguments

`File`	The output of ReadMethylFile function.
`NCluster`	The number of cluster.

Value

Objects of rgl

Examples

```
set.seed(123)
data <- Data2[1:100,]
data <- data.frame(t(data))
data <- cbind(rownames(data), data)
colnames(data)[1] <- "ID"
TSNEPlot(File = data, NCluster = 4)
```

`XGBoostModel` *XGBoost model*

Description

A function to train an XGBoost model to classify medulloblastoma subgroups using the DNA methylation dataset (Illumina Infinium HumanMethylation450). Prediction is followed by training if new data is provided.

Arguments

`SplitRatio`	Train and test split ratio. A value greater or equal to zero and less than one.
`CV`	The number of folds for cross-validation. It should be greater than one.
`NCores`	The number of cores for parallel computing.
`NewData`	A methylation data from the ReadMethylFile function.

Value

A list

Examples

```
set.seed(123)
xgboost <- XGBoostModel(SplitRatio = 0.6,
                        CV = 2,
                        NCores = 1,
                        NewData = NULL)
```

Index