Package 'DMBC'

July 1, 2016

Type Package

Title Dirichlet-Multinomial Bayes Classifier (DMBC) for Microbiome Classification

Version 0.1.0

Date 2016-06-22

Description
This package implements the machine-learning method described by Gao et al (2016) for microbiome classification using a Bayes classifier based on the Dirichlet-Multinomial distribution. In addition to classification, the package also identifies a subset of microbial taxa that can achieve the maximum classification accuracy.

Depends R(>= 3.2.0),
dirmult,
AUC,
ecodist,
MGLM,
caret,
e1071

Imports vegan

License MIT

LazyData TRUE

RoxygenNote 5.0.1

R topics documented:

best_cm *Confusion Matrix for the Best Model*

Description

Given an `loocv()` output and `Cal_AUC()` output, this function calculates the confusion matrix for the label produced from the model with optimized number of features.

Usage

```
best_cm(CV = cv, auc_out = auc_out)
```

Arguments

`CV`	Output object from loocv().
`auc_out`	Output object form Cal_AUC()

Value

A list of confusionMatrix class.

Examples

```
data(training)
cv = loocv(training)
auc_out = Cal_AUC(cv)
best_cm(CV=cv,auc_out=auc_out)
```

CalPrb *Calculate Likelihood based on Dirichlet-multinomial distribution estimated parameters*

Description

This function estimates parameters from Dirichlet-multinomial distribution.

Usage

```
CalPrb(FS_out = FS_out, testSet = testSet, col_start = 3, type_col = 2,
  HighestRank = nrow(FS_out$Feature))
```

Arguments

FS_out	An object from the FeatureSelection()
testSet	A test set in data frame or matrix form. The colnames should have the same bacteria (features) as in the training set.
col_start	An index indicating at which column is the beginning of bacteria (features) data. The default is the 3rd column.
type_col	An index indicating at which column is group/type variable. The default is the 2nd column.
HighestRank	The top number of features inclueded in model. The default is all the features left after filtering.

Value

A data frame with 17 columns, each row represents a model estimation output.

Examples

```
data(training)

#### Take one row as testSet ####
idx <- sample(1:nrow(training),1)
test <- training[idx,]
train <- training[-idx,]

CalPrb(FS(train),test) # This may take up to one minute
```

Cal_AUC *Calculate Area Under the ROC Curve from Leave-One-Out Cross-Validation*

Description

This function calculates the AUC for each selected subset of features from DMBC cross-validation results, and output the set of features which provide the greatest AUC.

Usage

```
Cal_AUC(CV = cv)
```

Arguments

CV	The output object from loocv()

Value

A data frame of 6 columns, comparison group1 label, comparison group2 label, number of features included in the model, AUC, AUC * prior probability, and selected features.

Examples

```
data(training)
Cal_AUC(loocv(training))
```

dmbc_predict *Predict probability from DMBC model*

Description

Based on a set of optimized features from training set, this function predicts the posterior probability for two class labels.

Usage

```
dmbc_predict(data = data, testset = testset, auc_out = auc_out,
  col_start = 3, type_col = 2, Prior1 = 0.5, Prior2 = 1 - Prior1)
```

Arguments

data	Validation dataset with rows are samples, columns are features. The first column should be the sample ID, second column group variable (Disease type, the label you want to classify on).
testset	Lable unknown testset without sample IDs.
auc_out	Output object of Cal_AUC() from a validation set.
col_start	An index indicating at which column is the beginning of bacteria (features) data in the validation set. The default is the 3rd column.
type_col	An index indicating at which column is group/type variable in the validation set. The default is the 2nd column.
Prior1	Prevalence of label1 according to literature or experience. Default is 0.5.
Prior2	Prevalence of label2 according to literature or experience. Default is 0.5.

Examples

```
#load the DMBC library
library(DMBC)

#load training dataset
data(training)

#load test dataset
data(test)

#calculate AUC based on training set
auc_out <- Cal_AUC(loocv(training))

#predict unknown test set using training set and auc results.
dmbc_predict(data=training,testset=test,auc_out=auc_out)
```

FS *Filter and Feature Selection based on Wilcoxon Rank-Sum test*

Description

Given a training set, this function performs feature selection based on several thresholds: (1). Average relative abudance in each cohort class (minimum relative abundance by default is 0.25

Usage

```
FS(training = data, type_col = 2, col_start = 3, Cutoff_mean = 5e-04,
  Cutoff_ratio = 0.1, totalReadsCutoff = 500)
```

Arguments

training	A data frame of training set.
type_col	An index indicating at which column is group/type variable. The default is the 3rd column.
col_start	An index indicating at which column is the beginning of bacteria (features) data. Default is the 2nd column.
Cutoff_mean	The minimum average relative abundance allowed in filtering step. Default is 0.0005.
Cutoff_ratio	The non-zero ratio cutoff in filtering features.Default value is 0.1.
totalReadsCutoff	The minimum allowed total reads per sample. Any sample has less than this number of total reads will be removed. Default is 500.

Value

A list of 2: Feature and CountData.

Feature A list of selected features sorted by their Wilcoxon P values.

CountData A data frame containing balanced data.

Examples

```
data(training)
FS(training)
```

loocv *Leave one out cross-validation*

Description

Implementation of leave-one-out cross-validataion. It takes in an entire dataset, with first column the sample IDs, second column the group variable (classificatoin variable/disease type), using "Type" as its colum names, and third column the beginning of count table.

Usage

```
loocv(data = data, type_col = 2, col_start = 3, Cutoff_mean = 5e-04,
  Cutoff_ratio = 0.1, totalReadsCutoff = 500)
```

Arguments

data	A data frame validation set with first column the sample IDs, second column the group variable (classificatoin variable/disease type), and third column the beginning of count table.
type_col	An index indicating at which column is group/type variable. The default is the 3rd column.
col_start	An index indicating at which column is the beginning of bacteria (features) data. Default is the 2nd column.
Cutoff_mean	The minimum average relative abundance allowed in filtering step. Default is 0.0005.
Cutoff_ratio	The non-zero ratio cutoff in filtering features.Default value is 0.1.
totalReadsCutoff	The minimum allowed total reads per sample. Any sample has less than this number of total reads will be removed. Default is 500.

Value

A list of CalPrb() results. This output will be further used for Cal_AUC() in order to estimate the final model.

Examples

```
data(training)
loocv(training)
```

---

test *Example Testing Data*

Description

This is a simulated testing dataset with no label for each sample.

Usage

```
data(test)
```

Format

An object of class `data.frame` with 6 rows and 7 columns.

Examples

```
data(test)
```

`training` *Example Training Data*

Description

This is a simulated testing dataset with label for each sample.

Usage

```
data(training)
```

Format

An object of class `data.frame` with 20 rows and 9 columns.

Examples

```
data(training)
```

Index