Package 'Harvest.Tree'

January 20, 2025

Title Harvest the Classification Tree

Version 1.1

Date 2015-07-30

Author Bingyuan Liu/Yan Yuan/Qian Shi

Maintainer Bingyuan Liu <adler1016@gmail.com>

Depends R (>= 3.0.1)

Imports rpart,stats

Description Aimed at applying the Harvest classification tree algorithm, modified algorithm of classic classification tree.The harvested tree has advantage of deleting redundant rules in trees, leading to a simplify and more efficient tree model.It was firstly used in drug discovery field, but it also performs well in other kinds of data, especially when the region of a class is disconnected. This package also improves the basic harvest classification tree algorithm by extending the field of data of algorithm to both continuous and categorical variables. To learn more about the harvest classification tree algorithm, you can go to http://www.stat.ubc.ca/Research/TechReports/techreports/220.pdf for more information.

License GPL-2

NeedsCompilation no

Repository CRAN

Date/Publication 2015-07-31 00:54:59

Contents

extrule *Bound of rules*

Description

This function takes in a ruleset and output the lower and upper bounds of each rule.

Usage

```
extrule(myrules, varname)
```

Arguments

myrules	A 3 column matrix output of function "hughs.path.rpart"
varname	the names of x variables

Value

A p*2 matrix, p is the length of varname. The first column is the lower bound, the second column is the upper bound. The default lower bound is "-Inf",the default upper bound is "Inf". row corresponse to x variables ordered in the data matrix given to rpart.

harfunc *A harvested classification tree*

Description

Basic function to apply the harvest algorithm to the training data set, computing whether we can harvest any nodes based on the classic classification tree algorithm.

Usage

```
harfunc(rpart.object, data, varname, sig = 0.95)
```

Arguments

rpart.object	classification result of training data from traditional classification tree(rpart function).
data	original training data where 'y' stores classmembership
varname	the name of each explaanatory variables
sig	significance level (default 0.95)

Value

the list of orginial result of classification, likelihood improvment and harvested classification result.

`harvest` *A harvested classification tree*

Description

The main function of the package, aiming at develop the harvest classification tree. Training data input and

Usage

```
harvest(training, num.var, numeric.info, sig = 0.95)
```

Arguments

`training`	original data where 'y' stores classmembership 0 and 1,in the first column, with explanatory variable stores in the second to the last column.
`num.var`	number of explanatory variables
`numeric.info`	the vector stores the number of which variable is continuous
`sig`	significance level (default 0.95)

Details

The function will return the harvested tree model. Missing values are allowed, and they will be treated accordingly. To use the trained tree model to predict, you can use `predict` function in this package.

Value

An object of class "harvest", which is the result of algorithm with the following elements for each nodes(nodes are ordered in sequence of harvesting):

`rule` constriants of the node

`total` total number of data points in the node

`'1'` the number of data points belonging to class 1 in the node

`'logchange'` the improvement of log likelihood of deleting the redundent rules by the algorithm for the node

Examples

```
data(training)
harvest(training,4,3)
```

`Harvest.Tree` *Harvest the classification tree*

Description

Aimed at applying the Harvest classification tree algorithm, modified algorithm of classic classification tree.The harvested tree has advantage of deleting redundant rules in trees, leading to a simplify and more efficient tree model.It was firstly used in drug discovery field, but it also performs well in other kinds of data, especially when the region of a class is disconnected. This package also improves the basic harvest classification tree algorithm by extending the field of data of algorithm to both continuous and categorical variables.

To learn more about the harvest classification tree algorithm, you can go to http://www.stat.ubc.ca/Research/TechReports/tech for more information.

Details

Package:	Harvest.Tree
Type:	Package
Version:	1.1
Date:	2015-07-30
License:	GPL-2

The main function of package called 'harvest', it can be used to analyze the data which is stored in a data frame, where first column stores the class of response data, and the second to last column stores explantory variables accordingly.The 'predict' funciton offers function to predict the unclassified data based on training model. The 'harfunc' function is the fundemental part of 'harvest', which can be used to analyze the data which has already been classified by rpart function(traditional classification tree). Please check the help file of these three functions for more information.

Author(s)

Bingyuan Liu \ Yan Yuan \ Qian Shi

Maintainer: Bingyuan Liu<adler1016@gmail.com>

`predict` *Predictions from a harvested tree*

Description

The function `predict` computes the prediction of membership from a new data set classified by harvested classification model of training data.

Usage

```
predict(harfunc.object, data, num.var)
```

Arguments

harfunc.object	the output of harfunc function.
data	test data
num.var	number of explaining varibles

Details

To run the predict function, a trained harvested classification tree formed by harvest function is required.

Value

pred.mat is a data frame stored the information of result of prediction with the following columns:

belong the node that data point belongs to

possibility the probability of point being in class 1

predict the simple perdict based on whether probability is larger than 0.5.

---

rank.nodes *Ranking of nodes*

Description

Rank harvested node by lower p value

Usage

```
rank.nodes(harfunc.object)
```

Arguments

harfunc.object	an object of class "harfunc"

Value

the ranked harvest nodes

rulesets *A logical matrix for a terminal node*

Description

Return a logical matrix of the rule sets which define a terminal node

Usage

```
rulesets(noden, newsim, varn, nodenumb)
```

Arguments

noden	a terminal node defined by a set of rules, from function "treemat"
newsim	data to be harvested
varn	x variable names
nodenumb	all the labels of terminal nodes

Value

A nxnn logical matrix, n=number of data points to be harvested, nn=number of rules defining a terminal node. Each column of the matrix corresponding to a node that is defined by one variable/rule, its name corresponds to that variable. Note the original terminal node is just the intersection of these nodes.

training *training*

Description

A simulated data set of symptoms of breast cancer patients

Usage

```
data(training)
```

Format

A data frame with 300 observations on the following 5 variables.

y a numeric vector

x1 a factor with levels 21 22 23 24 25 26 27 28 29

x2 a factor with levels 39- 40-49 50-69 70-74 75+

x3 a numeric vector

x4 a factor with levels 2004 2005 2006 2007 2008 2009 2010

Source

simulated data for breast cancer diagnosis

Index