

# Package ‘C443’

June 7, 2020

**Type** Package

**Title** See a Forest for the Trees

**Version** 3.0.0

**Imports** MASS, partykit, rpart, RColorBrewer, grDevices, gridExtra, ggplot2, cluster, parallel, igraph, reshape2, qgraph, stats, graphics,plyr

**LazyData** true

**Encoding** UTF-8

**Date** 2020-06-06

**Description** Get insight into a forest of classification trees, by calculating similarities between the trees, and subsequently clustering them. Each cluster is represented by it's most central cluster member. The package implements the methodology described in Sies & Van Mechelen (2020) <doi:10.1007/s00357-019-09350-4>.

**License** GPL (>= 2)

**RoxygenNote** 7.1.0

**NeedsCompilation** no

**Author** Aniek Sies [aut, cre],  
Iven Van Mechelen [ths]

**Maintainer** Aniek Sies <aniek.sies@kuleuven.be>

**Repository** CRAN

**Date/Publication** 2020-06-06 23:40:08 UTC

## R topics documented:

clusterforest . . . . .	2
clusters.clusterforest . . . . .	4
drugs . . . . .	5
medoidtrees.clusterforest . . . . .	7
plot.clusterforest . . . . .	7
print.clusterforest . . . . .	8
summary.clusterforest . . . . .	9

treesimilarities.clusterforest . . . . .	9
treesource . . . . .	10
treesource.clusterforest . . . . .	11
treesource.default . . . . .	12

<b>Index</b>	<b>13</b>
--------------	-----------

---

clusterforest	<i>Clustering the classification trees in a forest based on similarities</i>
---------------	--

---

## Description

A function to get insight into a forest of classification trees by clustering the trees in a forest using Partitioning Around Medoids (PAM, Kaufman & Rousseeuw, 2009), based on user provided similarities, or based on similarities calculated by the package based on a similarity measure chosen by the user (see Sies & Van Mechelen, 2020).

## Usage

```
clusterforest(
  fulldata,
  treedata,
  trees,
  simmatrix = NULL,
  m = NULL,
  tol = NULL,
  weight = NULL,
  fromclus = 1,
  toclus = 1,
  treecov = NULL
)
```

## Arguments

fulldata	The full/original dataset
treedata	A list of dataframes on which the trees are based
trees	A list of trees of class party, classes inheriting from party (e.g., glmtree), or classes that can be coerced to party (i.e., rpart, Weka_tree, XMLnode).
simmatrix	A similaritymatrix with the similarities between all trees. Should be square, symmetric and have ones on the diagonal. Default=NULL
m	Similarity measure that should be used to calculate similarities, in the case that no similarity matrix was provided by the user. Default=NULL. m=1 is based on counting equal predictors or predictor-split point combinations (Equation 3 or 6 in the Supplementary Materials of Sies & Van Mechelen (2020)); m=2 is the measure of Shannon & Banks (1999), based on counting the number of equal paths from rootnode to leafs (See also Sies & Van Mechelen (2020), Equation 2); m=3 is based on the agreement in predicted classification labels (Chipman,

1998, see also Sies & Van Mechelen (2020), Equation 7);  $m=4$  is based on the agreement of partitions (Chipman, 1998; see also Sies & Van Mechelen (2020), Equation 5);  $m=5$  is based on counting equal elementary conjunctions of trees transformed to disjunctive normal form ((limited to binary predictors), see Sies & Van Mechelen (2020), Supplementary materials Equation 15);  $m=6$  is based on comparing sets of predictors or predictor split point combinations (taking into account directions of the splits) associated with a leaf, while also taking into account the classification label of that leaf (see Sies & Van Mechelen (2020), Supplementary Materials Equation 13 and 14).

tol	A vector with for each predictor a number that defines the tolerance zone within which two split points of the predictor in question are assumed equal. For example, if the tolerance for predictor X is 1, then a split on that predictor in tree A will be assumed equal to a split in tree B as long as the splitpoint in tree B is within the splitpoint in tree A + or - 1. Only applicable for $m=1$ and $m=6$ . Default=NULL
weight	If 1, the number of dissimilar paths in the Shannon and Banks measure ( $m=2$ ), should be weighted by 1/their length (Otherwise they are weighted equally). Only applicable for $m=2$ . Default=NULL
fromclus	The lowest number of clusters for which the PAM algorithm should be run. Default=1.
toclus	The highest number of clusters for which the PAM algorithm should be run. Default=1.
treecov	A vector with the covariate value for each tree in the forest.

### Details

The user should provide the number of clusters that the solution should contain, or a range of numbers. In the latter case, the resulting clusterforest object will contain results for each solution. On top of this clusterforest object, several methods, such as plot, print and summary, can be used.

### Value

The function returns an object of class clusterforest, with attributes:

medoids	the position of the medoid trees in the forest (i.e., which element of the list of partytrees)
medoidtrees	the medoid trees
clusters	The cluster to which each tree in the forest is assigned
avgsilwidth	The average silhouette width for each solution (see Kaufman and Rousseeuw, 2009)
agreement	For each solution, the agreement between the predicted class label for each observation based on the forest as a whole, and those based on the medoids only (see Sies & Van Mechelen,2020)
withinsim	Within cluster similarity for each solution (see Sies & Van Mechelen, 2020)
treesimilarities	Similarity matrix on which clustering was based
treecov	covariate value for each tree in the forest

## References

Kaufman, L., & Rousseeuw, P. J. (2009). *Finding groups in data: an introduction to cluster analysis* (Vol. 344). John Wiley & Sons.

Sies, A. & Van Mechelen I. (2020). C443: An R-package to see a forest for the trees. *Journal of Classification*.

Shannon, W. D., & Banks, D. (1999). Combining classification trees using MLE. *Statistics in medicine*, 18(6), 727-740.

Chipman, H. A., George, E. I., & McCulloh, R. E. (1998). Making sense of a forest of trees. *Computing Science and Statistics*, 84-92.

## Examples

```
require(MASS)
require(rpart)
#Function to draw a bootstrap sample from a dataset
DrawBoots <- function(dataset, i){
  set.seed(2394 + i)
  Boot <- dataset[sample(1:nrow(dataset), size = nrow(dataset), replace = TRUE),]
  return(Boot)
}

#Function to grow a tree using rpart on a dataset
GrowTree <- function(x,y,BootsSample, minsplit = 40, minbucket = 20, maxdepth =3){
  controlrpart <- rpart.control(minsplit = minsplit, minbucket = minbucket, maxdepth = maxdepth,
    maxsurrogate = 0, maxcompete = 0)
  tree <- rpart(as.formula(paste(noquote(paste(y, "~")), noquote(paste(x, collapse="+"))))),
    data = BootsSample, control = controlrpart)
  return(tree)
}

#Use functions to draw 20 bootstrapsamples and grow a tree on each sample
Boots<- lapply(1:10, function(k) DrawBoots(Pima.tr ,k))
Trees <- lapply(1:10, function (i) GrowTree(x=c("npreg", "glu", "bp", "skin",
"bmi", "ped", "age"), y="type", Boots[[i]] ))

#Turn the lists of dataframes and rpart trees to a forest object
ClusterForest<- clusterforest(fulldata=Pima.tr,treedata=Boots,trees=Trees,m=1,
fromclus=1, toclus=5)
```

---

clusters.clusterforest

*Get the cluster assignments for a solution of a clusterforest object*

---

## Description

A function to get the cluster assignments for a given solution of a clusterforest object.

**Usage**

```
clusters.clusterforest(x, ..., solution = 1)
```

**Arguments**

x	A clusterforest object
...	Additional arguments
solution	The solution for which cluster assignments should be returned. Default = 1

---

drugs	<i>Drug consumption data set</i>
-------	----------------------------------

---

**Description**

A dataset collected by Fehrman et al. (2017), freely available on the UCI Machine Learning Repository (Lichman, 2013) containing records of 1885 respondents regarding their use of 18 types of drugs, and their measurements on 12 predictors. # All predictors were originally categorical and were quantified by Fehrman et al. (2017). The meaning of the values can be found on <https://archive.ics.uci.edu/ml/datasets/Drug+consumption+%28quantified%29>. The original response categories for each drug were: never used the drug, used it over a decade ago, or in the last decade, year, month, week, or day. We transformed these into binary response categories, where 0 (non-user) consists of the categories never used the drug and used it over a decade ago and 1 (user) consists of all other categories.

**Usage**

```
drugs
```

**Format**

A data frame with 1185 rows and 32 variables:

**ID** Respondent ID

**Age** Age of respondent

**Gender** Gender of respondent, where 0.48 denotes female and -0.48 denotes male

**Edu** Level of education of participant

**Country** Country of current residence of participant

**Ethn** Ethnicity of participant

**Neuro** NEO-FFI-R Neuroticism score

**Extr** NEO-FFI-R Extraversion score

**Open** NEO-FFI-R Openness to experience score

**Agree** NEO-FFI-R Agreeableness score

**Consc** NEO-FFI-R Conscientiousness score

**Impul** Impulsiveness score measured by BIS-11  
**Sensat** Sensation seeking score measured by ImpSS  
**Alc** Alcohol user (1) or non-user (0)  
**Amphet** Amphetamine user (1) or non-user (0)  
**Amyl** Amyl nitrite user (1) or non-user (0)  
**Benzos** Benzodiazepine user (1) or non-user (0)  
**Caff** Caffeine user (1) or non-user (0)  
**Can** Cannabis user (1) or non-user (0)  
**Choco** Chocolate user (1) or non-user (0)  
**Coke** Coke user (1) or non-user (0)  
**Crack** Crack user (1) or non-user (0)  
**Ecst** Ecstasy user (1) or non-user (0)  
**Her** Heroin user (1) or non-user (0)  
**Ket** Ketamine user (1) or non-user (0)  
**Leghighs** Legal Highs user (1) or non-user (0)  
**LSD** LSD user (1) or non-user (0)  
**Meth** Methadone user (1) or non-user (0)  
**Mush** Magical Mushroom user (1) or non-user (0)  
**Nico** Nicotine user (1) or non-user (0)  
**Semeron** Semeron user (1) or non-user (0), fictitious drug to identify over-claimers  
**VSA** volatile substance abuse user(1) or non-user (0)

### Source

<https://archive.ics.uci.edu/ml/machine-learning-databases/00373/>

### References

Fehrman, E., Muhammad, A. K., Mirkes, E. M., Egan, V., & Gorban, A. N. (2017). *The Five Factor Model of personality and evaluation of drug consumption risk*. In *Data Science* (pp. 231-242). Springer, Cham.

Lichman, M. (2013). *UCI machine learning repository*.

---

 medoidtrees.clusterforest

*Get the medoid trees for a solution of a clusterforest object*


---

### Description

A function to get the medoid trees for a given solution of a clusterforest object.

### Usage

```
medoidtrees.clusterforest(x, ..., solution = 1)
```

### Arguments

x	A clusterforest object
...	Additional arguments
solution	The solution for which medoid trees should be returned. Default = 1

---

 plot.clusterforest *Plot a clusterforest object*


---

### Description

A function that can be used to plot a clusterforest object, either by returning plots with information on the cluster solutions (e.g., average silhouette width), or plots of the medoid trees of each solution.

### Usage

```
## S3 method for class 'clusterforest'
plot(x, ..., solution = NULL)
```

### Arguments

x	A clusterforest object
...	Additional arguments
solution	The solution to plot the medoid trees from. Default = NULL

### Details

This function can be used to plot a clusterforest object in two ways. If it's used with as only argument the clusterforest object, then the average silhouette width, agreement in predicted labels between medoids in solution and forest, and within cluster similarity measures are plotted for each solution. These plots may be helpful in deciding how many clusters are needed to summarize the forest (see Sies & Van Mechelen, 2020).

If the function is used with two arguments (the clusterforest object and the number of the solution), then the medoid tree(s) of that solution are plotted.

## References

Sies, A. & Van Mechelen I. (2020). C443: An R-package to see a forest for the trees. *Journal of Classification*.

## Examples

```
require(MASS)
require(rpart)
#Function to draw a bootstrap sample from a dataset
DrawBoots <- function(dataset, i){
  set.seed(2394 + i)
  Boot <- dataset[sample(1:nrow(dataset), size = nrow(dataset), replace = TRUE),]
  return(Boot)
}

#Function to grow a tree using rpart on a dataset
GrowTree <- function(x,y,BootsSample, minsplit = 40, minbucket = 20, maxdepth =3){
  controlrpart <- rpart.control(minsplit = minsplit, minbucket = minbucket,
  maxdepth = maxdepth, maxsurrogate = 0, maxcompete = 0)
  tree <- rpart(as.formula(paste(noquote(paste(y, "~")),
  noquote(paste(x, collapse="+")))), data = BootsSample,
  control = controlrpart)
  return(tree)
}

#Use functions to draw 20 bootstrapsamples and grow a tree on each sample
Boots<- lapply(1:10, function(k) DrawBoots(Pima.tr ,k))
Trees <- lapply(1:10, function (i) GrowTree(x=c("npreg", "glu", "bp",
"skin", "bmi", "ped", "age"), y="type",
Boots[[i]] ))

ClusterForest<- clusterforest(fulldata=Pima.tr,treedata=Boots,trees=Trees,m=1,
fromclus=1, toclus=5)
plot(ClusterForest)
plot(ClusterForest,2)
```

---

print.clusterforest    *Print a clusterforest object*

---

## Description

A function that can be used to print a clusterforest object.

## Usage

```
## S3 method for class 'clusterforest'
print(x, ..., solution = 1)
```



**Arguments**

x	A clusterforest object
...	Additional arguments
solution	The solution to print the medoid trees from. Default = NULL

---

summary.clusterforest *Summarize a clusterforest object*

---

**Description**

A function to summarize a clusterforest object.

**Usage**

```
## S3 method for class 'clusterforest'  
summary(object, ...)
```

**Arguments**

object	A clusterforest object
...	Additional arguments

---

treesimilarities.clusterforest  
*Get the similarity matrix that was used to create a clusterforest object*

---

**Description**

A function to get the similarity matrix used to obtain a clusterforest object.

**Usage**

```
treesimilarities.clusterforest(x, ...)
```

**Arguments**

x	A clusterforest object
...	Additional arguments

---

treesource	<i>Mapping the tree clustering solution to a known source of variation underlying the forest</i>
------------	--

---

### Description

A function that can be used to get insight into a clusterforest solution, in the case that there is a known source of variation underlying the forest. It visualizes the number of trees from each source that belong to each cluster.

### Usage

```
treesource(clusterforest, solution)
```

### Arguments

clusterforest	The clusterforest object
solution	The solution

### Value

multiplot	For each value of the source, a bar plot with the number of trees that belong to each cluster
heatmap	A heatmap with for each value of the source, the number of trees that belong to each cluster

### Examples

```
require(rpart)
data_Amphet <- drugs[,c("Amphet", "Age", "Gender", "Edu", "Neuro", "Extr", "Open", "Agree",
"Consc", "Impul", "Sensat")]
data_cocaine <- drugs[,c("Coke", "Age", "Gender", "Edu", "Neuro", "Extr", "Open", "Agree",
"Consc", "Impul", "Sensat")]
data_LSD <- drugs[,c("LSD", "Age", "Gender", "Edu", "Neuro", "Extr", "Open", "Agree",
"Consc", "Impul", "Sensat")]

#Function to draw a bootstrap sample from a dataset
DrawBoots <- function(dataset, i){
  set.seed(2394 + i)
  Boot <- dataset[sample(1:nrow(dataset), size = nrow(dataset), replace = TRUE),]
  return(Boot)
}

#Function to grow a tree using rpart on a dataset
GrowTree <- function(x,y,BootsSample, minsplit = 40, minbucket = 20, maxdepth = 3){

  controlrpart <- rpart.control(minsplit = minsplit, minbucket = minbucket, maxdepth = maxdepth,
```

```

maxsurrogate = 0, maxcompete = 0)
tree <- rpart(as.formula(paste(noquote(paste(y, "~")), noquote(paste(x, collapse="+")))),
  data = BootsSample, control = controlrpart)
return(tree)
}

#Draw bootstrap samples and grow trees
BootsA<- lapply(1:5, function(k) DrawBoots(data_Amphet,k))
BootsC<- lapply(1:5, function(k) DrawBoots(data_cocaine,k))
Boots = c(BootsA,BootsC)

TreesA <- lapply(1:5, function (i) GrowTree(x=c ("Age", "Gender", "Edu", "Neuro",
"Extr", "Open", "Agree","Consc", "Impul","Sensat"), y="Amphet", BootsA[[i]] ))
TreesC <- lapply(1:5, function (i) GrowTree(x=c ( "Age", "Gender", "Edu", "Neuro",
"Extr", "Open", "Agree", "Consc", "Impul","Sensat"), y="Coke", BootsC[[i]] ))
Trees=c(TreesA,TreesC)

#Cluster the trees
ClusterForest<- clusterforest(fulldata=drugs,treedata=Boots,trees=Trees,m=1,
fromclus=2, toclus=2, treecov=rep(c("Amphet","Coke"),each=5))

#Link cluster result to known source of variation
treesource(ClusterForest, 2)

```

---

treesource.clusterforest

*Mapping the tree clustering solution to a known source of variation  
underlying the forest*

---

## Description

A function that can be used to get insight into a clusterforest solution, in the case that there is a known source of variation underlying the forest. It visualizes the number of trees from each source that belong to each cluster.

## Usage

```
## S3 method for class 'clusterforest'
treesource(clusterforest, solution)
```

## Arguments

clusterforest	The clusterforest object
solution	The solution

---

treesource.default	<i>Mapping the tree clustering solution to a known source of variation underlying the forest</i>
--------------------	--

---

**Description**

A function that can be used to get insight into a clusterforest solution, in the case that there is a known source of variation underlying the forest. It visualizes the number of trees from each source that belong to each cluster.

**Usage**

```
## Default S3 method:  
treesource(clusterforest, solution)
```

**Arguments**

clusterforest	The clusterforest object
solution	The solution

# Index

## \*Topic **datasets**

drugs, [5](#)

clusterforest, [2](#)

clusters.clusterforest, [4](#)

drugs, [5](#)

medoidtrees.clusterforest, [7](#)

plot.clusterforest, [7](#)

print.clusterforest, [8](#)

summary.clusterforest, [9](#)

treesimilarities.clusterforest, [9](#)

treesource, [10](#)

treesource.clusterforest, [11](#)

treesource.default, [12](#)