Package ‘aricode’

June 26, 2020

Type Package

Title Efficient Computations of Standard Clustering Comparison Measures

Version 1.0.0

Maintainer Julien Chiquet <julien.chiquet@inrae.fr>

Description Implements an efficient O(n) algorithm based on bucket-sorting for fast computation of standard clustering comparison measures. Available measures include adjusted Rand index (ARI), normalized information distance (NID), normalized mutual information (NMI), adjusted mutual information (AMI), normalized variation information (NVI) and entropy, as described in Vinh et al (2009) <doi:10.1145/1553374.1553511>. Include AMI (Adjusted Mutual Information) since version 0.1.2, a modified version of ARI (MARI) and simple Chi-square distance since version 1.0.0.

License GPL (>= 3)

URL https://github.com/jchiquet/aricode (dev version)

BugReports https://github.com/jchiquet/aricode/issues

LazyData TRUE

Encoding UTF-8

Imports Matrix, Rcpp

Suggests testthat, spelling

LinkingTo Rcpp

RooxygenNote 7.1.0

Language en-US

NeedsCompilation yes

Author Julien Chiquet [aut, cre] (<https://orcid.org/0000-0002-3629-3429>),
Guillem Rigail [aut],
Martina Sundqvist [aut],
Valentin Dervieux [ctb]

Repository CRAN

Date/Publication 2020-06-26 14:30:03 UTC
R topics documented:

AMI ................................................................. 2
ARI ................................................................. 3
aricode ............................................................ 3
Chi2 ................................................................. 4
clustComp ........................................................ 5
entropy ............................................................ 6
MARI ............................................................... 6
MARIraw .......................................................... 7
NID ................................................................. 8
NMI ................................................................. 8
NVI ................................................................. 9
RI ................................................................. 10
sortPairs .......................................................... 11

Index 12

AMI

Adjusted Mutual Information

Description

A function to compute the adjusted mutual information between two classifications

Usage

AMI(c1, c2)

Arguments

c1 a vector containing the labels of the first classification. Must be a vector of
characters, integers, numerics, or a factor, but not a list.

c2 a vector containing the labels of the second classification.

Value

a scalar with the adjusted rand index.

See Also

ARI, RI, NID, NVI, NMI, clustComp

Examples

data(iris)
c1 <- cutree(hclust(dist(iris[, -5])), 4)
AMI(c1, iris$$Species)
Description
A function to compute the adjusted rand index between two classifications

Usage
ARI(c1, c2)

Arguments
- c1: a vector containing the labels of the first classification. Must be a vector of characters, integers, numerics, or a factor, but not a list.
- c2: a vector containing the labels of the second classification.

Value
a scalar with the adjusted rand index.

See Also
RI, NID, NVI, NMI, clustComp

Examples
```
data(iris)
c1 <- cutree(hclust(dist(iris[, -5])), 4)
ARI(c1, iris$Species)
```

Description
A package for efficient computations of standard clustering comparison measures. Most of the available measures are described in the paper of Vinh et al, JMLR, 2009 (see reference below).

Details
Traditional implementations (e.g., function adjustedRandIndex of package mclust) are in Omega(n + u v) where n is the size of the vectors the classifications of which are to be compared, u and v are the respective number of classes in each vectors. Here, the implementation is in Theta(n), plus the gain of speed due to the C++ code.
Functions in aricode

The functions included in aricode are:

- ARI: computes the adjusted rand index
- Chi2: computes the Chi-square statistic
- MARI: computes the modified adjusted rand index (Sundqvist et al, in preparation)
- MARIraw: computes the raw version of the modified adjusted rand index
- RI: computes the rand index
- NVI: computes the normalized variation information
- NID: computes the normalized information distance
- NMI: computes the normalized mutual information
- AMI: computes the adjusted mutual information
- entropy: computes the conditional and joint entropies
- clustComp: computes all clustering comparison measures at once

Author(s)

Julien Chiquet <julien.chiquet@inrae.com>
Guillem Rigaill <guillem.rigaill@inrae.fr>
Martina Sundqvist <martina.sundqvist@agroparistech.fr>

References


See Also

ARI, RI, NID, NVI, AMI, NMI, entropy, clustComp

Chi2

Chi-square statistics

Description

A function to compute the Chi-2 statistics

Usage

Chi2(c1, c2)
**clustComp**

**Arguments**
- **c1** a vector containing the labels of the first classification. Must be a vector of characters, integers, numerics, or a factor, but not a list.
- **c2** a vector containing the labels of the second classification.

**Value**
- a scalar with the chi-square statistics.

**See Also**
- ARI, NID, NVI, NMI, clustComp

**Examples**
```r
data(iris)
c1 <- cutree(hclust(dist(iris[, -5])), 4)
Chi2(c1, iris$Species)
```

---

**Description**
A function to measure similarity between two classifications.

**Usage**
```
clustComp(c1, c2)
```

**Arguments**
- **c1** a vector containing the labels of the first classification. Must be a vector of characters, integers, numerics, or a factor, but not a list.
- **c2** a vector containing the labels of the second classification.

**Value**
- a list with the RI, ARI, NMI, NVI and NID.

**See Also**
- RI, NID, NVI, NMI, ARI

**Examples**
```r
data(iris)
c1 <- cutree(hclust(dist(iris[, -5])), 4)
clustComp(c1, iris$Species)
```
entropy

Description
A function to compute the empirical entropy for two vectors of classification and the joint entropy

Usage
entropy(c1, c2)

Arguments
- c1: a vector containing the labels of the first classification. Must be a vector of characters, integers, numerics, or a factor, but not a list.
- c2: a vector containing the labels of the second classification.

Value
a list with the two conditional entropies, the joint entropy and output of sortPairs.

Examples
```r
data(iris)
c1 <- cutree(hclust(dist(iris[,-5])), 4)
entropy(c1, iris$Species)
```

MARI

Description
A function to compute a modified adjusted rand index between two classifications as proposed by Sundqvist et al. in prep, based on a multinomial model.

Usage
MARI(c1, c2)

Arguments
- c1: a vector containing the labels of the first classification. Must be a vector of characters, integers, numerics, or a factor, but not a list.
- c2: a vector containing the labels of the second classification.
Value

a scalar with the modified ARI.

See Also

ARI, NID, NVI, NMI, clustComp

Examples

data(iris)
c1 <- cutree(hclust(dist(iris[-5])), 4)
MARIraw(c1, iris$Species)
**NID**

*Normalized information distance (NID)*

**Description**

A function to compute the NID between two classifications.

**Usage**

```r
NID(c1, c2)
```

**Arguments**

- `c1`: a vector containing the labels of the first classification. Must be a vector of characters, integers, numerics, or a factor, but not a list.
- `c2`: a vector containing the labels of the second classification.

**Value**

A scalar with the normalized information distance.

**See Also**

`RI`, `NMI`, `NVI`, `ARI`, `clustComp`.

**Examples**

```r
data(iris)
c1 <- cutree(hclust(dist(iris[, -5])), 4)
NID(c1, iris$Species)
```

**NMI**

*Normalized mutual information (NMI)*

**Description**

A function to compute the NMI between two classifications.

**Usage**

```r
NMI(c1, c2, variant = c("max", "min", "sqrt", "sum", "joint"))
```

**Examples**

```r
data(iris)
c1 <- cutree(hclust(dist(iris[, -5])), 4)
NMI(c1, iris$Species)
```
**NVI**

**Arguments**
- `c1` a vector containing the labels of the first classification. Must be a vector of characters, integers, numerics, or a factor, but not a list.
- `c2` a vector containing the labels of the second classification.
- `variant` a string in ("max", "min", "sqrt", "sum", "joint"): different variants of NMI. Default use "max".

**Value**
a scalar with the normalized variation of information.

**See Also**
- RI, NID, NVI, ARI, clustComp

**Examples**
```r
data(iris)
c1 <- cutree(hclust(dist(iris[, -5])), 4)
NVI(c1, iris$Species)
```

---

**Description**
A function to compute the NVI between two classifications

**Usage**

```
NVI(c1, c2)
```

**Arguments**
- `c1` a vector containing the labels of the first classification. Must be a vector of characters, integers, numerics, or a factor, but not a list.
- `c2` a vector containing the labels of the second classification.

**Value**
a scalar with the normalized variation of information.

**See Also**
- RI, NID, NMI, ARI, clustComp
**Examples**

```r
data(iris)
c1 <- cutree(hclust(dist(iris[, -5])), 4)
NVI(c1, iris$Species)
RI(cl, iris$Species)
```

<table>
<thead>
<tr>
<th>RI</th>
<th>Rand Index</th>
</tr>
</thead>
</table>

**Description**

A function to compute the rand index between two classifications

**Usage**

```r
RI(c1, c2)
```

**Arguments**

- `c1`: a vector containing the labels of the first classification. Must be a vector of characters, integers, numerics, or a factor, but not a list.
- `c2`: a vector containing the labels of the second classification.

**Value**

a scalar with the rand index.

**See Also**

ARI, NID, NVI, NMI, clustComp

**Examples**

```r
data(iris)
c1 <- cutree(hclust(dist(iris[, -5])), 4)
RI(cl, iris$Species)
```
sortPairs

Description

A function to sort pairs of integers or factors and identify the pairs

Usage

sortPairs(c1, c2, spMat = FALSE)

Arguments

c1 a vector of length n with value between 0 and N1 < n
c2 a vector of length n with value between 0 and N2 < n
spMat logical: send back the contingency table as sparsely encoded (cost more than the algorithm itself). Default is FALSE
Index

AMI, 2, 4
ARI, 2, 3, 4, 5, 7–10
aricode, 3

Chi2, 4
clustComp, 2–5, 5, 7–10

entropy, 4, 6

MARI, 6
MARIRaw, 7

NID, 2–5, 7, 8, 9, 10
NMI, 2–5, 7, 8, 9, 10
NVI, 2–5, 7–9, 9, 10

RI, 2–5, 8, 9, 10

sortPairs, 11