

Package ‘CCMtools’

February 14, 2012

Version 1.0

Date 2010-23-02

Title Clustering through “Correlation Clustering Model” (CCM) and cluster analysis tools.

Author Mathieu Vrac <mathieu.vrac@lsce.ipsl.fr>

Maintainer Mathieu Vrac <mathieu.vrac@lsce.ipsl.fr>

Depends R (>= 1.8.0), mclust, class, tree,mvtnorm

Description This package proposes a clustering method called “Correlation Clustering Model” (CCM) based on mixture of canonical correlation analysis (CCA). It also provides some tools for cluster analysis.

License GPL (>= 2)

Repository CRAN

Date/Publication 2010-02-26 14:28:39

R topics documented:

CCM	2
CWGLI	3
DI	4
Info.Criterion	5
learn.and.project.clusters	6
Percent.bad.and.false.classif.per.cluster	7
WGP	8

Index	10
--------------	-----------

CCM

*Clustering through "Correlation Clustering Model" (CCM) method***Description**

This function performs a simultaneous clustering of two matched datasets (e.g., daily local- and large-scale atmospheric data), such that each cluster maximises the correlation between the two datasets. This clustering is based on a mixture of canonical correlation analyses (CCAs).

Usage

```
CCM(Nc, NS, DataA.tbc, DataS.tbc, NN, DataStation, init="block", ITmax=15, rq=0)
```

Arguments

Nc	Number of clusters required.
NS	Number of locations (i.e., weather stations) for the local-scale time series.
DataA.tbc	Dataset corresponding to the large-scale data to be clustered. This is a matrix $M \times NN$, where M corresponds to the number of large-scale locations (e.g., GCM or RCM grid cells), and NN to the length of the time series (e.g., number of days). Note that this matrix is used in CCM without any transformation. For example, if a principal component analysis (PCA) has to be performed, this must be done BEFORE entering CCM.
DataS.tbc	Dataset corresponding to the local-scale (station) data to be clustered. This is a matrix $NS \times NN$. Similarly as for DataA.tbc, note that this matrix is used in CCM without any transformation, and that if a PCA has to be performed, this must be done BEFORE entering CCM.
NN	Length of the time series (e.g, number of days for daily time series).
DataStation	Local-scale (stations) dataset on which the information criterion will be calculated. It usually is the same as DataS.tbc but can be different according to the application (e.g., DataS.tbc is the result of a PCA) or to the goal to reach.
init	Initializing method for the clusters. Six methods are available: <ul style="list-style-type: none"> - "block": Blocks initialization (the default) - "12345": Each day is alternatively allocated to a cluster (For example, if 3 clusters required, day1 goes to C1, day2 to C2, day3 to C3, day4 to C1, day5 to C2, etc.) - "Kmeans": Initialization by the k-means algorithm - "Mixtn": Same a "12345" but for length 12 (instead of length 1) - "EMw": Initialization by the EM clustering algorithm applied onto the w (i.e., large-scale) canonical variates resulting from a CCA performed between DataA.tbc and DataS.tbc - "EMvw": Same as "EMw" but EM is applied onto both v (local) and w (large-scale) canonical variates.
ITmax	Maximum number of iterations (default is 15) is the algorithm di not converge.

`rq` Value (of the local-scale variable of interest) on which the information criterion (IC) is calculated (default is `rq=0`). `rq` can be the 90th percentile of the data. In that case, CCM will try to find clusters for which the extremes are well discriminated. A high IC means a good discrimination (in terms of local-scale variable) between the clusters.

Details

For details about the CCM method, see the reference below.

M. Vrac, P. Yiou. "Weather regimes designed for local precipitation modelling: Application to the Mediterranean basin". JGR-Atmospheres, doi:10.1029/2009JD012871, 2010

Author(s)

M. Vrac (mathieu.vrac@lscce.ipsl.fr)

See Also

[Info.Criterion](#)

Examples

```
## Example
```

CWGLI	<i>Conditional Weighted Global Log-Intensity (CWGLI) cost function for precipitation</i>
-------	--

Description

This function calculates the Conditional Weighted Global Log-Intensity (CWGLI) cost function for precipitation. This error is supposed to be due to a misclassification (i.e., the elements should be classified according to sequence `cl.check`, while they are in practice allocated to the clusters according to the sequence `cl.proj`).

Usage

```
CWGLI(cl.check, cl.proj, DataS.check)
```

Arguments

<code>cl.check</code>	Reference sequence (i.e., numerical vector) of clusters (i.e., this is the sequence we should have).
<code>cl.proj</code>	Sequence of clusters in practice.
<code>DataS.check</code>	Precipitation data time series

Details

For details about this cost function, see the reference below.

M. Vrac, P. Yiou. "Weather regimes designed for local precipitation modelling: Application to the Mediterranean bassin". Submitted, JGR-A, 2009

Value

Returns a numerical vector corresponding to the CWGLI cost function

Author(s)

M. Vrac (mathieu.vrac@lsce.ipsl.fr)

See Also

[DI](#), [WGP](#)

Examples

```
## Example
```

DI

Daily Intensity (DI) cost function for precipitation

Description

This function calculates the Daily Intensity (DI) cost function for precipitation. This error is supposed to be due to a misclassification (i.e., the elements should be classified according to sequence `cl.check`, while they are in practice allocated to the clusters according to the sequence `cl.proj`).

Usage

```
DI(cl.calibration, cl.check, cl.proj, DataS.Calibration, DataS.check)
```

Arguments

<code>cl.calibration</code>	Sequence of clusters for time period 1 (TP1)
<code>cl.check</code>	Reference sequence (i.e., numerical vector) of clusters (i.e., this is the sequence we should have) for time period 2 (TP2).
<code>cl.proj</code>	Sequence of clusters in practice for TP2.
<code>DataS.Calibration</code>	Precipitation data time series for TP1.
<code>DataS.check</code>	Precipitation data time series for TP2.

Details

For details about these cost functions, see the reference below.

M. Vrac, P. Yiou. "Weather regimes designed for local precipitation modelling: Application to the Mediterranean bassin". Submitted, JGR-A, 2009

Value

Returns a list with arguments:

DI	Daily Intensity (DI) error costs, depending on station s and day d .
MDI	Means DI costs, depending on station s .
RMDI	Regional MDI providing a global (spatial and temporal) view of the DI error cost.

Author(s)

M. Vrac (mathieu.vrac@lsce.ipsl.fr)

See Also

[DI](#), [WGP](#)

Examples

```
## Example
```

Info.Criterion	<i>Computes the Information Criterion (IC) of a clustering result, for values higher than r.</i>
----------------	---

Description

This function computes the Information Criterion (IC) of a clustering result.

Usage

```
Info.Criterion(NS, DataS, r, totCL, Nc, cl)
```

Arguments

NS	Number of locations (i.e., weather stations) for the local-scale time series on which IC is calculated.
DataS	Dataset corresponding to local-scale (station) data on which IC is calculated. This is a matrix $NS \times NN$, where NN is the number of days (i.e., length of the time series).
r	Value for which the IC is calculated (see details).

totCL	Vector of numbers of elements (e.g., days) in each cluster.
Nc	Number of clusters.
cl	Vector containing the sequence of clusters (length(cl) is NN).

Details

The IC is computed as $IC = \sum_{i=1}^K \ln_{i,r} - (p_r * n_i)$, where
 $n_{i,r}$ = \# of days in cluster i that receive a rainfall amount $> r$
 p_r = proba of such rainy days in the whole population
 n_i = \# of days in cluster i

Author(s)

M. Vrac (mathieu.vrac@lscce.ipsl.fr)

See Also

[CCM](#)

Examples

```
## Example
```

```
learn.and.project.clusters
```

Learning of attribution of clusters, and projection of new days to clusters

Description

This function (1) learns how to attribute days to clusters based on the sequence of predictors and associated sequence of clusters.

Usage

```
learn.and.project.clusters(DataCalibration, DataToBeProjected, cl.calibration, allocmet, DataS.Calibr
```

Arguments

DataCalibration

Values of the predictor variable for the calibration set (can be a matrix).

DataToBeProjected

Values of the predictor variable for the projection set.

cl.calibration Numerical vector corresponding to the sequence of clusters (i.e., calibration set).

`allocmet` Name of the attribution method.(The 12 possibilities are: "Euclid.dist.A", "Euclid.dist.w1", "Euclid.dist.w2", "CART.A", "CART.w", "CART.A.and.w", "knnA", "knnA10", "Gaussian.A", "Gaussian.w", "MM", "MMw")

`DataS.Calibration` Values of other predictor variables for the calibration set. This is sometimes needed, according to the attribution method (`allocmet`) to be used (needed for "Euclid.dist.w1", "Euclid.dist.w2", "CART.w", "CART.A.and.w", "Gaussian.w", "MMw").

Value

Returns a list with two objects:

`cl` The sequence of clusters defined from the predictors for the projection set.

`tot` Number of elements per cluster for projection set.

Author(s)

M. Vrac (mathieu.vrac@lsce.ipsl.fr)

Examples

```
## Example
```

`Percent.bad.and.false.classif.per.cluster`
Percentage of bad and false classification

Description

This function computes the percentage of bad and false classification of a sequence of clusters (`new.cl`) according to a reference sequence (`cl`).

Usage

```
Percent.bad.and.false.classif.per.cluster(cl, new.cl)
```

Arguments

`cl` Reference sequence of clusters.

`new.cl` Sequence of clusters to be compared to the reference sequence.

Value

Returns a list containing the following elements:

tot	Global percentage of bad classification
BadPerCluster	Percentage of bad classification per cluster
FalsePerCluster	Percentage of false classification per cluster
mat.att	Global matrix of attribution (row = cl, colomn = new.cl)

Author(s)

M. Vrac (mathieu.vrac@lsce.ipsl.fr)

See Also

[learn.and.project.clusters](#)

Examples

```
## Example
```

WGP	<i>Weighted Global Probability (WGP) cost function for precipitation occurrence</i>
-----	---

Description

This function calculates the Weighted Global Probability (WGP) cost function for precipitation occurrence. This error is supposed to be due to a misclassification (i.e., the elements should be classified according to sequence `cl.check`, while they are in practice allocated to the clusters according to the sequence `cl.proj`).

Usage

```
WGP(cl.check, cl.proj, DataS.check)
```

Arguments

<code>cl.check</code>	Reference sequence (i.e., numerical vector) of clusters.
<code>cl.proj</code>	Sequence of clusters in practice.
<code>DataS.check</code>	Precipitation data time series

Details

For details about this cost function, see the reference below.

M. Vrac, P. Yiou. "Weather regimes designed for local precipitation modelling: Application to the Mediterranean bassin". Submitted, JGR-A, 2009

Value

Returns a numerical vector corresponding to the WGP cost function

Author(s)

M. Vrac (mathieu.vrac@lscce.ipsl.fr)

See Also

[DI](#), [CWGLI](#)

Examples

```
## Example
```

Index

*Topic **cluster**

CCM, [2](#)

CWGLI, [3](#)

DI, [4](#)

Info.Criterion, [5](#)

learn.and.project.clusters, [6](#)

Percent.bad.and.false.classif.per.cluster,
[7](#)

WGP, [8](#)

*Topic **models**

CCM, [2](#)

CWGLI, [3](#)

DI, [4](#)

WGP, [8](#)

CCM, [2](#), [6](#)

CWGLI, [3](#), [9](#)

DI, [4](#), [4](#), [5](#), [9](#)

Info.Criterion, [3](#), [5](#)

learn.and.project.clusters, [6](#), [8](#)

Percent.bad.and.false.classif.per.cluster,
[7](#)

WGP, [4](#), [5](#), [8](#)