Package 'tame'

February 23, 2023

Title Timing, Anatomical, Therapeutic and Chemical Based Medication Clustering

Version 0.0.1

Description Agglomerative hierarchical clustering with a bespoke distance measure based on medication similarities in the Anatomical Therapeutic Chemical Classification System, medication timing and medication amount or dosage. Tools for summarizing, illustrating and manipulating the cluster objects are also available.

License MIT + file LICENSE

Encoding UTF-8

RoxygenNote 7.2.3

Imports dplyr, fuzzyjoin, magrittr, purrr, Rfast, rlang, stats, stringr, tibble, tidyr, tidyselect, Rcpp (>= 1.0.8)

Suggests rmarkdown, knitr, testthat (>= 3.0.0)

BugReports https://github.com/Laksafoss/tame/issues

Config/testthat/edition 3

Depends R (>= 4.2)

LazyData true

LinkingTo Rcpp

NeedsCompilation yes

Author Anna Laksafoss [aut, cre] (<https://orcid.org/0000-0002-9898-2924>)

Maintainer Anna Laksafoss <adls@ssi.dk>

Repository CRAN

Date/Publication 2023-02-23 14:00:05 UTC

R topics documented:

`complications` *A Simulated Data Set About Pregnancy Complications*

Description

We use this data set in all the examples in the package.

Usage

```
complications
```

Format

An object of class `data.frame` with 149 rows and 8 columns.

`eczema` *A Simulated Data Set About Eczema*

Description

A Simulated Data Set About Eczema

Usage

```
eczema
```

Format

An object of class `data.frame` with 50644 rows and 7 columns.

`employ` *Employ a Clustering to New Data*

Description

Employ a clustering to new data

Usage

```
employ(
  object,
  new_data,
  only = NULL,
  additional_data = NULL,
  assignment_method = "nearest_cluster",
  parallel = FALSE,
  ...
)
```

Arguments

`object`	A `medic` clustering object for which employment is desired.
`new_data`	A data frame in which to look for variables with
`only`	<data-masking> Expressions that return a logical value, and are defined in terms of the variables in `object` and/or `additional_data` and specifies which clusterings should be employed to the new data.
`additional_data`	A data frame with additional data that may be (left-)joined onto the `parameters` in `object`. This is often used in conjuction with `only` to select specific clusterings based on `additional_data`.
`assignment_method`	A character naming the employment method. The default assignment method `"nearest_cluster"` matches people in `new_data` to their nearest cluster in the chosen clusterings from `object`. As finding exact matches (the next assignment method) is contained within this strategy the `"exact_only"` matches are also reported in additional columns in the output. The assignment method `"exact_only"` only matches a person from `new_data` to a cluster if they are a perfect match to anyone in `object`. Thus, people from `new_data` are not guaranteed assignment to a cluster.
`parallel`	A logical or an integer. If `FALSE`, the default, no parallelization is done. If `TRUE` or an integer larger than 2L parallelization is implemented via parLapply from the **parallel** package. When `parallel` is `TRUE` the number of clusters is set to detectCores - 1, and when `parallel` is an integer then the number of clusters is set to `parallel`. For more details on the parallelization method see parallel::parLapply.
`...`	Additional arguments affecting the employment procedure.

Value

`employ` returns a `medic` object.

Examples

```
part1 <- complications[1:100,]
part2 <- complications[101:149,]

clust <- medic(part1, id = id, atc = atc, k = 3)

# Nearest cluster matching
employ(clust, part2)

# Only exact matching
employ(clust, part2, assignment_method = "exact_only")
```

---

enrich *Enrich Clustering Parameter*

Description

Enrich the parameter information in a clustering with user-defined data.

Usage

```
enrich(object, additional_data = NULL, by = NULL)
```

Arguments

`object`	A `medic` object for enrichment.
`additional_data`	A data frame with additional data that may be (left-)joined onto the `parameters` in `object`.
`by`	A character vector of variables to join by. This variables is passed to the by term in a `dplyr::left_join()` and inherits its behavior: If `NULL`, the default, the join will perform a natural join, using all variables in common across the `parameters` and `additional_data`. To join by different variables on `parameters` and `additional_data`, use a named vector. For example, `by = c("k" = "cluster_size")` will match `parameters$k` to `additional_data$cluster_size`. To join by multiple variables, use a vector with length > 1. For example, `by = c("k", "summation_method")` will match `parameters$k` to `additional_data$k` and `parameters$summation_method` to `additional_data$summation_method`. Use a named vector to match different variables in `parameters` and `additional_data`. For example, `by = c("k" = "cluster_size", "summation_method" = "sm")` will match `parameters$k` to `additional_data$cluster_size` and `parameters$summation_method` to `additional_data$sm`.

Details

The `enrich()` function is a joining function used for enriching the clustering characteristics with user-defined data. This function is used in all of the investigative functions with a `additional_data` statement such as `frequencies()` and `amounts()`.

Value

An object of class *medic*.

Examples

```
clust <- medic(
   complications,
   id = id,
   atc = atc,
   timing = first_trimester:third_trimester,
   k = 3:5
)

new_parameters <- data.frame(k = 3:5, size = c("small", "small", "large"))

enrich(clust, new_parameters)
```

---

is.medic *Test if an object is a medic-object*

Description

Test if an object is a medic-object

Usage

```
is.medic(object)
```

Arguments

`object`	Any object.

Value

`TRUE` is the object inherits from the `medic` class and has the required elements.

Examples

```
clust <- medic(complications, id = id, atc = atc, k = 3)
is.medic(clust)
```

medic *Medication clustering (based on ATC and timing)*

Description

The medic method uses agglomerative hierarchical clustering with a bespoke distance measure based on medication ATC codes similarities, medication timing and medication amount or dosage.

Usage

```
medic(
  data,
  k = 5,
  id,
  atc,
  timing,
  base_clustering,
  linkage = "complete",
  summation_method = "sum_of_minima",
  alpha = 1,
  beta = 1,
  gamma = 1,
  p = 1,
  theta = (5:0)/5,
  parallel = FALSE,
  return_distance_matrix = FALSE,
  set_seed = FALSE,
  ...
)
```

Arguments

data	A data frame containing all the variables for the clustering.
k	a vector specifying the number of clusters to identify.
id	<tidy-select> An unquoted expression naming the variable in data describing person id.
atc	<tidy-select> An unquoted expression naming the variable in data containing ATC codes.
timing	<tidy-select> An unquoted expression naming the variable or variables in data describing medication timing. Variable names can be used as if they were positions in the data frame, so expressions like x:y can be used to select a range of variables. Moreover, pattern matching selection helpers such as starts_with or num_range may also be used to select timing variables.
base_clustering	<tidy-select> An unquoted expression naming the variable in data that gives an initial clustering to start the medic from or NULL.

`linkage` The agglomeration method to be used in the clustering. This should be (an unambiguous abbreviation of) one of "ward.D", "ward.D2", "single", "complete", "average" (= UPGMA), "mcquitty" (= WPGMA), "median" (= WPGMC) or "centroid" (= UPGMC). See stats::hclust for more information. For a discussion of linkage criterion choice see *details* below.

`summation_method` The summation method used in the distance measure. This should be either "double_sum" or "sum_of_minima". See *details* below for more information.

`alpha` A number giving the tuning of the normalization. See *details* below for more information.

`beta` A number giving the power of the individual medication combinations. See *details* below for more information.

`gamma` A number giving the weight of the timing terms. See *details* below for more information.

`p` The power of the Minkowski distance used in the timing-specific distance. See *details* below for more information.

`theta` A vector of length 6 specifying the tuning of the ATC measure. See *details* below for more information.

`parallel` A logical or an integer. If `FALSE`, the default, no parallelization is done.

If `TRUE` or an integer larger than 2L parallelization is implemented via parLapply from the **parallel** package. When `parallel` is `TRUE` the number of clusters is set to detectCores - 1, and when `parallel` is an integer then the number of clusters is set to `parallel`. For more details on the parallelization method see parallel::parLapply.

`return_distance_matrix` A logical.

`set_seed` A logical or an integer.

`...` Additional arguments not currently in use.

Details

The `medic` method uses agglomerative hierarchical clustering with a bespoke distance measure based on medication ATC codes and timing similarities to assign medication pattern clusters to people.

Two versions of the distance measure are available:

The *double sum*:

$$d(p_i, p_j) = N_\alpha(M_i \times M_j) \sum_{m \in M_i} \sum_{n \in M_j} ((1 + D_\theta(m, n))(1 + \gamma T_p(t_{im}, t_{jn})) - 1)^\beta.$$

and the *sum of minima*:

$$d(p_i, p_j) = \frac{1}{2}(N_\alpha(M_i) \sum_{m \in M_i} \min_{n \in M_j} ((1+D_\theta(m,n))(1+\gamma T_p(t_{im}, t_{jn}))-1)^\beta + N_\alpha(M_j) \sum_{n \in M_j} \min_{m \in M_i} ((1+D_\theta(m,n))(1+\gamma T_p($$

Normalization:

$$N_\alpha(x) = |x|^{-\alpha}$$

If the normalization tuning, `alpha`, is 0, then no normalization is preformed and the distance measure becomes highly dependent on the number of distinct medications given. That is, people using more medication will have larger distances to others. If the normalization tuning, `alpha`, is 1 - the default - then the summation is normalized with the number of terms in the sum, in other words, the average is calculated.

ATC distance:

The central idea of this method, namely the ATC distance, is given as

$$D_\theta(x, y) = \sum_{i=1,\ldots,5} 1\{xandymatchonleveli, butnotleveli + 1\}\theta_i$$

The ATC distance is tuned using the vector `theta`.

Note that two ATC codes are said to match at level i when they are identical at level i. E.g. the two codes N06AB01 and N06AA01 match on level 1, 2, and 3 as they are both "N" at level 1, "N06" at level 2, and "N06A" at level 3, but at level 4 they differ ("N06AB" and "N06AA" are not the same).

Timing distance:

The timing distance is a simple Minkowski distance:

$$T(x, y) = (\sum_{t \in T} |x_t - y_t|^p)^{1/p}.$$

When `p` is 1, the default, the Manhattan distance is used.

Value

An object of class *medic* which describes the clusters produced the hierarchical clustering process. The object is a list with components:

data the inputted data frame `data` with the cluster assignments appended at the end.

clustering a data frame with the person id as given by `id`, the `.analysis_order` and the clusters found.

variables a list of the variables used in the clustering.

parameters a data frame with all the inputted clustering parameters and the corresponding method names. These method names correspond to the column names for each cluster in the `clustering` data frame described right above.

key a list of keys used internally in the function to keep track of simplified versions of the data.

distance_matrix the distance matrices for each method if `return_distance_matrix` is `TRUE` otherwise `NULL`.

call the matched call.

See Also

summary.medic for summaries and plots.

employ for employing an existing clustering to new data.

enrich for enriching the meta data in the `medic` object with additional data.

bind for binding together two comparable lists of clusterings.

Examples

```
# A simple clustering based only on ATC
clust <- medic(complications, id = id, atc = atc, k = 3)

# A simple clustering with both ATC and timing
clust <- medic(
  complications,
  id = id,
  atc = atc,
  timing = first_trimester:third_trimester,
  k = 3
)
```

---

parameters_constructor

Internal option constructor

Description

Given the input of the `medic` this function checks the input and constructs a data frame with the analysis parameters specified by the user.

Usage

```
parameters_constructor(
  data,
  id,
  k = 5,
  atc,
  timing,
  base_clustering,
  linkage = "complete",
  summation_method = "sum_of_minima",
  alpha = 1,
  beta = 1,
  gamma = 1,
  p = 1,
  theta = (5:0)/5,
```

```
  ...
)
```

Arguments

data	A data frame containing all the variables for the clustering.
id	<tidy-select> An unquoted expression naming the variable in data describing person id.
k	a vector specifying the number of clusters to identify.
atc	<tidy-select> An unquoted expression naming the variable in data containing ATC codes.
timing	<tidy-select> An unquoted expression naming the variable or variables in data describing medication timing. Variable names can be used as if they were positions in the data frame, so expressions like x:y can be used to select a range of variables. Moreover, pattern matching selection helpers such as starts_with or num_range may also be used to select timing variables.
base_clustering	<tidy-select> An unquoted expression naming the variable in data that gives an initial clustering to start the medic from or NULL.
linkage	The agglomeration method to be used in the clustering. This should be (an unambiguous abbreviation of) one of "ward.D", "ward.D2", "single", "complete", "average" (= UPGMA), "mcquitty" (= WPGMA), "median" (= WPGMC) or "centroid" (= UPGMC). See stats::hclust for more information. For a discussion of linkage criterion choice see *details* below.
summation_method	The summation method used in the distance measure. This should be either "double_sum" or "sum_of_minima". See *details* below for more information.
alpha	A number giving the tuning of the normalization. See *details* below for more information.
beta	A number giving the power of the individual medication combinations. See *details* below for more information.
gamma	A number giving the weight of the timing terms. See *details* below for more information.
p	The power of the Minkowski distance used in the timing-specific distance. See *details* below for more information.
theta	A vector of length 6 specifying the tuning of the ATC measure. See *details* below for more information.
...	Additional arguments not currently in use.

Value

A data.frame with the parameters for clustering.

Examples

```
parameters_constructor(
   data = complications,
   k = 3,
   id = id,
   atc = atc
)
```

`refactor` *Refactor Cluster Levels*

Description

Refactor the levels of the chosen clusters.

Usage

```
refactor(object, ..., inheret_parameters = TRUE)
```

Arguments

`object`	A `medic` object.
`...`	<data-masking> Name-value pairs. `...` is passed to `dplyr::mutate`, and therefor inherits its behavior: The name gives the name of the new clustering in the output. The value can be: • A vector of length 1, which will be recycled to the correct length. • A function of another clustering. When a recording uses the name of an existing clustering, this new clustering will overwrite the existing one.
`inheret_parameters`	A logical. If `TRUE` a new clustering overwriting an existing clustering inherits the `parameters` of the old.

Value

A `medic` object with relevant clusterings refactored.

Examples

```
clust <- medic(complications, id = id, atc = atc, k = 3:4)

# Refactor one clustering
refactor(
  clust,
  `cluster_1_k=4` = dplyr::recode(`cluster_1_k=4`, IV = "III")
```

```
)

# Refactor all clusterings
refactor(
  clust,
  dplyr::across(
    dplyr::everything(),
    ~dplyr::recode(., IV = "III")
  )
)
```

---

summary.medic *Summary of medic object*

Description

Make cluster characterizing summaries.

Usage

```
## S3 method for class 'medic'
summary(
  object,
  only = NULL,
  clusters = NULL,
 outputs = c("frequencies", "medications", "amounts", "trajectories", "interactions"),
  additional_data = NULL,
  ...
)

## S3 method for class 'summary.medic'
print(x, ...)

## S3 method for class 'summary.medic'
plot(x, by, facet, ...)
```

Arguments

object	An object for which a summary is desired.
only	<data-masking> Expressions that return a logical value, and are defined in terms of the variables in object and/or additional_data. The default NULL selects all clusterings in object.
clusters	<tidy-select> An unquoted expression naming the cluster or clusters in object one wants to see summaries of. Names can be used as if they were positions in the data frame, so expressions like I:IV can be used to select a range of clusters. The default NULL selects all clusters in the chosen clusterings of object.

`outputs` A character vector naming the desired characteristics to output. The default names all possible output types.

`additional_data`
A data frame with additional data that may be (left-)joined onto the `parameters` in `object`. This is often used in conjuction with `only` to select specific clusterings based on `additional_data`.

`...` Additional arguments passed to the internal summary function.

- `cluster_wise` an option in the `medications()` function.
- `m` an option in the `medications()` function. A numeric restricting the number of distinct ATC codes plotted within each cluster. That is, the (at most) `m` most frequent ATC codes within that cluster is given a color.
- `q` an option in the `medications()` function. A numeric between 0 and 1 restricting the minimal ATC codes frequency displayed within each cluster.
- `count_grouper` an option in the `amounts()` function. A function for grouping counts. As a standard it groups counts as 1 medication, 2 medications, and 3+ medications.
- `atc_groups` A data.frame specifying the ATC groups to summaries by. The data.frame must have two columns: (1) `regex` giving regular expressions specifying the wanted ATC groups and (2) `atc_groups` the name of this ATC grouping. As a standard the anatomical level (first level) of the ATC codes is used.

`x` A `summary.medic` object for printing or plotting.

`by` <data-masking>

`facet` <data-masking>

Value

A list of clustering characteristics of class `summary.medic` is returned. It can contain any of the following characteristics:

Frequencies:
The number of individuals assigned to each cluster and the associated frequency of assignment.

Medications:
The number of individuals with a specific ATC code within a cluster. Moreover, it calculates the percentage of people with this medication assigned to this cluster and the percent of people within the cluster with this medication.

Amounts:
The number of ATC codes an individual has, and then outputs the number of individuals within a cluster that has that many ATC codes. Moreover, various relevant percentages or calculated. See Value below for more details on these percentages.

Trajectories:
The number of unique timing trajectories in each cluster, and the average timing trajectories in each cluster.

Interactions:
The number of people with unique timing trajectory and ATC group, as given by `atc_groups`, in each cluster.

Methods (by generic)

- `print(summary.medic)`: Print method for medic-objects
- `plot(summary.medic)`: Plot method for medic-objects

Examples

```
clust <- medic(complications, id = id, atc = atc, k = 3:5)
```

Index