Package ‘blink’

October 6, 2020

Type Package
Title Record Linkage for Empirically Motivated Priors
Version 1.1.0
Depends R (>= 3.0.2), stringdist, plyr
Imports stats, utils
Suggests knitr, rmarkdown
Encoding UTF-8
VignetteBuilder knitr
Description An implementation of the model in Steorts (2015) <DOI:10.1214/15-
BA965SI>, which performs Bayesian entity resolution for categorical and text data, for any dis-
tance function defined by the user. In addition, the precision and recall are in the package to al-
low one to compare to any other comparable method such as logistic regression, Bayesian addi-
tive regression trees (BART), or random forests. The experiments are reproducible and illus-
trated using a simple vignette. LICENSE: GPL-3 + file license.
License GPL-3
LazyData TRUE
RoxygenNote 7.1.1.9000
NeedsCompilation no
Author Rebecca Steorts [aut, cre]
Maintainer Rebecca Steorts <beka@stat.duke.edu>
Repository CRAN
Date/Publication 2020-10-06 09:50:02 UTC

R topics documented:

check.IDs ................................................................. 2
identity.RLdata500 ..................................................... 2
links ................................................................. 3
links.compare .......................................................... 4
mms ................................................................. 4
Description

Check whether 2 records which are estimated to be linked have the same IDs

Usage

check_IDs(recpair, identity_vector)

Arguments

recpair A record pair
identity_vector A vector of the unique ids

Value

Whether or not two records which are estimated to be linked have the same unique ids

Examples

id <- c(1,2,3,4,5,1,7,8,9,10,11,12,13,14,15,16,17,18,19,20)
rec1 <- 6
rec2 <- 1
check_IDs(recpair=c(rec1,rec2),identity_vector=id)

Identity

Unique identifier for RLdata500 that corresponds to the record number format A vector that contains the codeid

Usage

identity.RLdata500
**Format**

An object of class numeric of length 500.

**Description**

Function that returns the shared MPMMS (except with an easier condition to code than JASA paper). Function to make a list of vectors of estimated links by "P(MPMMS)>0.5" method Note: The default settings return only MPMMSs with multiple members.

**Usage**

```r
links(lam.gs = lam.gs, include.singles = FALSE, show.as.multiple = FALSE)
```

**Arguments**

- `lam.gs` The estimated linkage structure with a default of 10 iterations
- `include.singles` Do not include the singleton records
- `show.as.multiple` Always return MPMMSs that have more than one member

**Value**

Returns the shared MPMMS

**Examples**

```r
lam.gs <- matrix(c(1,1,2,2,3,3,5,6,4,3,4,5,3,2,4,1,2,3,4,2),ncol=20, nrow=4)
links(lam.gs)
```
links.compare

*This function takes a set of pairwise links and identifies correct, incorrect, and missing links (correct = estimated and true, incorrect = estimated but not true, missing = true but not estimated)*

Description

This function takes a set of pairwise links and identifies correct, incorrect, and missing links (correct = estimated and true, incorrect = estimated but not true, missing = true but not estimated)

Usage

```
links.compare(est.links.pair, true.links.pair, counts.only = TRUE)
```

Arguments

- `est-links.pair` The number of estimated links
- `true-links.pair` The number of true links
- `counts-only` State whether or not counts only is true or false

Value

Gives a vector of the estimated and true links, estimated but not true links, and the true but not estimated links

Examples

```
id <- c(1,2,3,4,5,1,7,8,9,10,11,12,13,14,15,16,17,18,19,20)
lam.gs <- matrix(c(1,1,2,3,5,6,4,3,4,5,3,2,4,1,2,3,4,2),ncol=20, nrow=4)
est.links <- links(lam.gs)
true.links <- links(matrix(id,nrow=1))
est.links.pair <- pairwise(est.links)
links.compare(est.links.pair, true.links=id)
```

mms

*Function to compute a record’s Maximal Matching Set (MMS) based on a single linkage structure*

Description

Function to compute a record’s Maximal Matching Set (MMS) based on a single linkage structure

Usage

```
mms(lambda, record)
```
**Arguments**

- **lambda**: The linkage structure
- **record**: A vector of records

**Value**

Computes a record's MMS

**Examples**

```r
lambda <- matrix(c(1,1,2,2,3,3),ncol=3)
record <- c(1,10,3,5,20,2)
mms(lambda=lambda, record=record)
```

---

**mpmms**

*Function to compute a record’s MPMMS based on a Gibbs sampler.*

*Note: It returns a list of the MPMMS ($mpmms$) and its probability ($prob$)*

**Description**

Function to compute a record’s MPMMS based on a Gibbs sampler. Note: It returns a list of the MPMMS ($mpmms$) and its probability ($prob$)

**Usage**

```r
mpmms(lam.gs, record)
```

**Arguments**

- **lam.gs**: The gibbs sampler
- **record**: A specific record

**Value**

Returns a list of the MPMSS and the associated probabilities.

**Examples**

```r
lam.gs <- matrix(c(1,1,2,2,3,3,6,4,3,4,5,3,2,4,1,2,3,4,2),ncol=20, nrow=4)
record <- c(1,3,1,3,1,3,1,3,1,3,1,3,1,3,1,3,1,3,1,3,1,3,1,3)
mpmms(lam.gs=lam.gs, record=record)
```
pairwise  

Function to take links list that may contain 3-way, 4-way, etc. and reduce it to pairwise only (e.g., a 3-way link 12-45-78 is changed to 2-way links: 12-45, 12-78, 45-78)

Description

Function to take links list that may contain 3-way, 4-way, etc. and reduce it to pairwise only (e.g., a 3-way link 12-45-78 is changed to 2-way links: 12-45, 12-78, 45-78)

Usage

`pairwise(.links)`

Arguments

`.links`  
A vector of records that are linked to one another

Value

Returns two ways links of records

Examples

```r
id <- c(1,2,3,4,5,1,7,8,9,10,11,12,13,14,15,16,17,18,19,20)
lam.gs <- matrix(c(1,1,2,2,3,3,5,6,4,3,4,5,3,2,4,1,2,3,4,2),ncol=20, nrow=4)
est.links <- links(lam.gs)
est.links.pair <- pairwise(est.links)
```

rl.gibbs  

Gibbs sampler for empirically motivated Bayesian record linkage

Description

Gibbs sampler for empirically motivated Bayesian record linkage

Usage

```r
rl.gibbs(
  file.num = file.num,
  X.s = X.s,
  X.c = X.c,
  num.gs = num.gs,
  a = a,
  b = b,
  c = c,
)```
RLdata500

```r
d = d,
M = M
)

Arguments

file.num  The number of the file
X.s  A vector of string variables
X.c  A vector of categorical variables
num.gs  Total number of gibb iterations
a  Shape parameter of Beta prior
b  Scale parameter of Beta prior
c  Positive constant
d  Any distance metric measuring the latent and observed string
M  The true value of the population size

Value

lambda.out The estimated linkage structure via Gibbs sampling

Examples

data(RLdata500)
X.c <- as.matrix(RLdata500[c("by","bm","bd")][1:3,])
p.c <- ncol(X.c)
X.s <- as.matrix(RLdata500[c(1,3)][1:3,])
p.s <- ncol(X.s)
file.num <- rep(c(1,1,1),c(1,1,1))
d <- function(string1,string2){adist(string1,string2)}
lam.gs <- rl.gibbs(file.num,X.s,X.c,num.gs=2,a=.01,b=100,c=1,d, M=3)
```

Description

Data on synthetic generation of German names with 500 total records and 10 percent duplication.

Usage

RLdata500

Format

A data frame with five variables: fname_c1,lname_c1, by, codebm, bd.
Index

* datasets
  identity.RLdata500, 2
  RLdata500, 7

check_IDs, 2

identity.RLdata500, 2

links, 3
links.compare, 4

mms, 4
mpmms, 5

pairwise, 6

rl.gibbs, 6
RLdata500, 7