Package 'deductive'

October 13, 2022

Maintainer Mark van der Loo <mark.vanderloo@gmail.com>

License GPL-3

Title Data Correction and Imputation Using Deductive Methods

LazyData no

Type Package

LazyLoad yes

Description Attempt to repair inconsistencies and missing values in data records by using information from valid values and validation rules restricting the data.

Version 1.0.0

Depends R (>= 3.2.0)

URL https://github.com/data-cleaning/deductive

BugReports https://github.com/data-cleaning/deductive/issues

Imports methods, lintools, validate, stringdist

Suggests tinytest (>= 0.9.5)

RoxygenNote 7.1.1

NeedsCompilation yes

Author Mark van der Loo [cre, aut], Edwin de Jonge [aut]

Repository CRAN

Date/Publication 2021-03-29 15:00:06 UTC

R topics documented:

`correct_typos` *Correct typos in restricted numeric data*

Description

Attempt to fix violations of linear (in)equality restrictions imposed on a record by replacing values with values that differ from the original values by typographical errors.

Usage

```
correct_typos(dat, x, ...)

## S4 method for signature 'data.frame,validator'
correct_typos(dat, x, fixate = NULL, eps = 1e-08, maxdist = 1, ...)
```

Arguments

`dat`	An R object holding numeric (integer) data.
`x`	An R object holding linear data validation rules
`...`	Options to be passed to `stringdist` which is used to determine the typographic distance between the original value and candidate solutions. By default, the optimal string alignment distance is used, with all weights equal to one.
`fixate`	`[character]` vector of variable names that may not be changed
`eps`	`[numeric]` maximum roundoff error
`maxdist`	`[numeric]` maximum allowd typographical distance

Value

`dat`, with values corrected.

Details

The algorithm works by proposing candidate replacement values and checking whether they are likely to be the result of a typographical error. A value is accepted as a solution when it resolves at least one equality violation. An equality restriction `a.x=b` is considered satisfied when `abs(a.x-b)<eps`. Setting `eps` to one or two units of measurement allows for robust typographical error detection in the presence of roundoff-errors.

The algorithm is meant to be used on numeric data representing integers.

References

- The first version of the algorithm was described by S. Scholtus (2009). Automatic correction of simple typing errors in numerical data with balance edits. Statistics Netherlands, Discussion Paper 09046
- The generalized version of this algorithm that is implemented for this package is described in M. van der Loo, E. de Jonge and S. Scholtus (2011). Correction of rounding, typing and sign errors with the deducorrect package. Statistics Netherlands, Discussion Paper 2011019

Examples

```
library(validate)

# example from section 4 in Scholtus (2009)

v <-validate::validator(
   x1 + x2 == x3
 , x2 == x4
 , x5 + x6 + x7 == x8
 , x3 + x8 == x9
 , x9 - x10 == x11
 )


dat <- read.csv(textConnection(
"x1, x2 , x3  , x4 , x5 , x6, x7, x8 , x9   , x10 , x11
1452, 116, 1568, 116, 323, 76, 12, 411,  1979, 1842, 137
1452, 116, 1568, 161, 323, 76, 12, 411,  1979, 1842, 137
1452, 116, 1568, 161, 323, 76, 12, 411, 19979, 1842, 137
1452, 116, 1568, 161,   0,  0,  0, 411, 19979, 1842, 137
1452, 116, 1568, 161, 323, 76, 12,   0, 19979, 1842, 137"
))
cor <- correct_typos(dat,v)
dat - cor
```

`deductive` *Deductive Data Correction and Imputation*

Description

Use data validation restrictions to estimate missing values or trace and repair certain errors.

`impute_lr` *Impute values derived from linear (in)equality restrictions.*

Description

Partially filled records $\boldsymbol{x}$ under linear (in)equality restrictions may reveal unique imputation solutions when the system of linear inequalities is reduced by substituting observed values. This function applies a number of fast heuristic methods before deriving all variable ranges and unique values using Fourier-Motzkin elimination.

Usage

```
impute_lr(dat, x, ...)

## S4 method for signature 'data.frame,validator'
impute_lr(dat, x, methods = c("zeros", "piv", "implied"), ...)
```

Arguments

dat	an R object carrying data
x	an R object carrying validation rules
...	arguments to be passed to other methods.
methods	What methods to use. Add 'fm' to also compute variable ranges using fourier-motzkin elimination (can be slow and may use a lot of memory).

Note

The Fourier-Motzkin elimination method can use large amounts of memory and may be slow. When memory allocation fails for a ceratian record, the method is skipped for that record with a message. This means that there may be unique values to be derived but it is too computationally costly on the current hardware.

Examples

```
v <- validate::validator(y ==2,y + z ==3, x +y <= 0)
dat <- data.frame(x=NA_real_,y=NA_real_,z=NA_real_)
impute_lr(dat,v)
```

Index