Package 'CuCubes'

December 9, 2016

Title MultiDimensional Feature Selection (MDFS)

Version 0.1.0

URL https://featureselector.uco.uwb.edu.pl/pub/cucubes/

Description Functions for MultiDimensional Feature Selection (MDFS):
* calculating multidimensional information gains,
* finding interesting tuples for chosen variables,
* scoring variables,
* finding important variables,
* plotting selection results.
CuCubes is also known as CUDA Cubes and it is a library that allows fast CUDA-accelerated computation of information gains in binary classification problems.
This package wraps CuCubes and provides an alternative CPU version as well as helper functions for building MultiDimensional Feature Selectors.

Depends R (>= 3.2.1)

License GPL-3

Encoding UTF-8

LazyData true

RoxygenNote 5.0.1

NeedsCompilation yes

Author Radosław Piliszek [aut, cre],
Andrzej Sułecki [aut],
Paweł Tabaszewski [aut],
Krzysztof Mnich [aut],
Witold Rudnicki [cph]

Maintainer Radosław Piliszek <r.piliszek@uwb.edu.pl>

Repository CRAN

Date/Publication 2016-12-09 13:13:25

R topics documented:

ComputeInterestingTuples 2
ComputeMaxInfoGains 3
madelon 4
MDFS 4
plot.MDFS 5
RelevantVariables 6
RelevantVariables.MDFS 6

Index **8**

`ComputeInterestingTuples` *Interesting tuples*

Description

Interesting tuples

Usage

```
ComputeInterestingTuples(acceleration.type = "scalar", dimensions = 1,
  divisions = 1, discretizations = 1, seed = 0, range = 1,
  pseudo.count = 0.001, reduce.method = "max", ig.thr,
  interesting.vars = c(), data, decision)
```

Arguments

`acceleration.type`	acceleration type ('scalar' for none, 'avx'/'avx2' for use of the AVX/AVX2 instruction set respectively)
`dimensions`	number of dimensions
`divisions`	number of divisions
`discretizations`	number of discretizations
`seed`	seed for PRNG used during discretizations
`range`	discretization range (from 0.0 to 1.0)
`pseudo.count`	pseudo count
`reduce.method`	discretization reduce method (either "max" or "mean")
`ig.thr`	IG threshold above which the tuple is interesting
`interesting.vars`	variables for which to check the IGs (none = all)
`data`	input data where columns are variables and rows are observations
`decision`	decision variable as a boolean vector of length equal to number of observations

Value

none (the function prints results)

`ComputeMaxInfoGains` *Max information gains*

Description

Max information gains

Usage

```
ComputeMaxInfoGains(acceleration.type = "scalar", dimensions = 1,
  divisions = 1, discretizations = 1, seed = 0, range = 1,
  pseudo.count = 0.001, reduce.method = "max", data, decision)
```

Arguments

`acceleration.type`	acceleration type ('scalar' for none, 'avx'/'avx2' for use of the AVX/AVX2 instruction set respectively, 'cuda' for CUDA)
`dimensions`	number of dimensions
`divisions`	number of divisions
`discretizations`	number of discretizations
`seed`	seed for PRNG used during discretizations
`range`	discretization range (from 0.0 to 1.0)
`pseudo.count`	pseudo count
`reduce.method`	discretization reduce method (either "max" or "mean")
`data`	input data where columns are variables and rows are observations
`decision`	decision variable as a boolean vector of length equal to number of observations

Value

numeric vector with max information gain for each input variable

Examples

```
ComputeMaxInfoGains(data = madelon$data, decision = madelon$decision,
  discretizations = 1, range = 0, divisions = 22, dimensions = 1)
```

madelon *An artificial dataset called MADELON*

Description

An artificial dataset containing data points grouped in 32 clusters placed on the vertices of a five dimensional hypercube and randomly labeled 0/1.

Usage

```
madelon
```

Format

A list of two elements:

data 2000 by 500 matrix of 2000 objects with 500 features

decision vector of 2000 decisions (labels 0/1)

Details

The five dimensions constitute 5 informative features. 15 linear combinations of those features are added to form a set of 20 (redundant) informative features. There are 480 distractor features called 'probes' having no predictive power.

Included is the original training set with label -1 changed to 0.

Source

https://archive.ics.uci.edu/ml/datasets/Madelon

MDFS *Build MultiDimensional Feature Selector from IGs*

Description

Build MultiDimensional Feature Selector from IGs

Usage

```
MDFS(IGs, dimensions, divisions, response_divisions = 1, IG_bits = TRUE,
  IG_doubled = FALSE, ignore_lowest = length(IGs)%/%10,
  variable_number = length(IGs), calc_variable_number = TRUE,
  mode_1D = "exp", min_variable_number = variable_number%/%2,
  max_ignore_lowest = variable_number%/%3, max_iterations = 20,
  acceptable_error = 0.05)
```

Arguments

IGs	max conditional information gains
dimensions	number of dimensions
divisions	number of divisions
response_divisions	number of response divisions (i.e. categories-1)
IG_bits	input is in binary log (as opposed to natural log)
IG_doubled	input is doubled (to follow the chi-squared distribution)
ignore_lowest	number of variables with the lowest IG to ignore (ignored if computed)
variable_number	number of irrelevant variables (ignored if computed)
calc_variable_number	whether to compute the number of neglected and irrelevant variables
mode_1D	"exp" - exponential distribution, "lin" - linear function of chi-squared, "raw" - raw chi-squared
min_variable_number	minimum number of irrelevant variables
max_ignore_lowest	maximum number of ignored variables
max_iterations	maximum number of iterations in variable number calculation
acceptable_error	acceptable error level for distribution parameter

Value

MDFS (list-based S3 class object) with the following named elements: "IGs" is a vector of information gains (input copy) "order" is a vector of ordinal numbers (order of variables by decreasing score) "chi.squared" is a vector of chi-squared p-values "p.values" is a vector of eventual p-values "scores" is a list of two vectors FDR and FWER with FDR and FWER scores respectively "lo.sq.dev." is a vector of square deviations used to calculate the number of ignored variables "hi.sq.dev." is a vector of square deviations used to calculate the number of irrelevant variables "ign.lowest" is a number of ignored variables "var.number" is a number of irrelevant variables "dist.param." is an exponential distribution parameter or linear coefficient "err.param." is a square error of the parameter

---

plot.MDFS *Plot MDFS details*

Description

Plot MDFS details

Usage

```
## S3 method for class 'MDFS'
plot(x, plots = c("I", "p", "FDR"), ...)
```

Arguments

x	an MDFS object
plots	plots to plot (I for max IG, p for p-values, FDR, FWER)
...	ignored

RelevantVariables *Find indices of relevant variables*

Description

Find indices of relevant variables

Usage

```
RelevantVariables(fs, ...)
```

Arguments

fs	feature selector
...	arguments passed to methods

Value

indices of important variables

RelevantVariables.MDFS *Find indices of relevant variables from MDFS*

Description

Find indices of relevant variables from MDFS

Usage

```
## S3 method for class 'MDFS'
RelevantVariables(fs, level = 0.05, score = "FDR", ...)
```

Arguments

`fs`	an MDFS object
`level`	statistical significance level
`score`	score to use
`...`	ignored

Value

indices of relevant variables

Index

∗Topic **datasets**
madelon, 4

ComputeInterestingTuples, 2
ComputeMaxInfoGains, 3

madelon, 4
MDFS, 4

plot.MDFS, 5

RelevantVariables, 6
RelevantVariables.MDFS, 6