Package: datanugget (via r-universe)

September 15, 2024

Type Package

Title Create, and Refine Data Nuggets

Version 1.3.1

Date 2024-09-14

Author Yajie Duan [cre, ctb], Traymon Beavers [aut], Javier Cabrera [aut], Ge Cheng [aut], Kunting Qi [aut], Mariusz Lubomirski [aut]

Maintainer Yajie Duan <yajieritaduan@gmail.com>

Description Creating, and refining data nuggets. Data nuggets reduce a large dataset into a small collection of nuggets of data, each containing a center (location), weight (importance), and scale (variability) parameter. Data nugget centers are created by choosing observations in the dataset which are as equally spaced apart as possible. Data nugget weights are created by counting the number observations closest to a given data nugget center. We then say the data nugget 'contains' these observations and the data nugget center is recalculated as the mean of these observations. Data nugget scales are created by calculating the trace of the covariance matrix of the observations contained within a data nugget divided by the dimension of the dataset. Data nuggets are refined by 'splitting' data nuggets which have scales or shapes (defined as the ratio of the two largest eigenvalues of the covariance matrix of the observations contained within the data nugget) Reference paper: [1] Beavers, T. E., Cheng, G., Duan, Y., Cabrera, J., Lubomirski, M., Amaratunga, D., & Teigler, J. E. (2024). Data Nuggets: A Method for Reducing Big Data While Preserving Data Structure. Journal of Computational and Graphical Statistics, 1-21. [2] Cherasia, K. E., Cabrera, J., Fernholz, L. T., & Fernholz, R. (2022). Data Nuggets in Supervised Learning. \emph{In Robust and Multivariate Statistical Methods: Festschrift in Honor of David E. Tyler} (pp. 429-449). Cham: Springer International Publishing.

Depends R (>= 4.0), doSNOW (>= 1.0.16), foreach (>= 1.5.1), parallel (>= 4.0.5),Rfast(>= 2.0.7)

License GPL-2

Encoding UTF-8

NeedsCompilation no

Repository CRAN

Date/Publication 2024-09-14 16:30:02 UTC

Contents

`datanugget-package` *Data Nuggets*

Description

This package contains functions to create and refine data nuggets which serve as representative samples of large datasets. The functions which perform these processes are create.DN, refine.DN, and AC, respectively.

Author(s)

Traymon Beavers, Javier Cabrera, Mariusz Lubomirski

References

Beavers, T. E., Cheng, G., Duan, Y., Cabrera, J., Lubomirski, M., Amaratunga, D., & Teigler, J. E. (2024). Data Nuggets: A Method for Reducing Big Data While Preserving Data Structure. Journal of Computational and Graphical Statistics, 1-21.

Cherasia, K. E., Cabrera, J., Fernholz, L. T., & Fernholz, R. (2022). Data Nuggets in Supervised Learning. *In Robust and Multivariate Statistical Methods: Festschrift in Honor of David E. Tyler* (pp. 429-449). Cham: Springer International Publishing.

AC *Calculate Arithmetic Complexicity of the Algorithm That Creates Data Nuggets*

Description

This function creates the centers of data nuggets from a random sample.

Usage

```
AC(x,
   R,
   delete.percent,
   DN.num1,
   DN.num2)
```

Arguments

x	A data matrix (of class matrix, data.frame, or data.table) containing only entries of class numeric.
R	The number of observations to sample from the data matrix when creating the initial data nugget centers. Must be of class numeric within [100,10000].
delete.percent	The proportion of observations to remove from the data matrix at each iteration when finding data nugget centers. Must be of class numeric and within (0,1).
DN.num1	The number of initial data nugget centers to create. Must be of class numeric.
DN.num2	The number of data nuggets to create. Must be of class numeric.

Details

This function is used for calculating the arithmetic complexicity of the algorithm behind the create.DN function for the given parameter choices.

Value

my.AC	The arithmetic complexicity of the algorithm behind the create.DN function for the given parameter choices on a log10 scale.

Author(s)

Traymon Beavers, Javier Cabrera, Mariusz Lubomirski

References

Beavers, T. E., Cheng, G., Duan, Y., Cabrera, J., Lubomirski, M., Amaratunga, D., & Teigler, J. E. (2024). Data Nuggets: A Method for Reducing Big Data While Preserving Data Structure. Journal of Computational and Graphical Statistics, 1-21.

Cherasia, K. E., Cabrera, J., Fernholz, L. T., & Fernholz, R. (2022). Data Nuggets in Supervised Learning. *In Robust and Multivariate Statistical Methods: Festschrift in Honor of David E. Tyler* (pp. 429-449). Cham: Springer International Publishing.

Examples

```
X = cbind.data.frame(rnorm(10^6),
                     rnorm(10^6),
                     rnorm(10^6),
                     rnorm(10^6),
                     rnorm(10^6))

my.AC = AC(x = X,
           R = 5000,
           delete.percent = .1,
           DN.num1 = 10^4,
           DN.num2 = 2000)
```

`create.DN` *Create Data Nuggets*

Description

This function draws a random sample of observations from a large dataset and creates data nuggets, a type of representative sample of the dataset, using a specified distance metric.

Usage

```
create.DN(x,
          center.method = "mean",
          R = 5000,
          delete.percent = .1,
          DN.num1 = 10^4,
          DN.num2 = 2000,
          dist.metric = "euclidean",
          seed = 291102,
          no.cores = (detectCores() - 1),
          make.pbs = TRUE)
```

Arguments

x	A data matrix (of class matrix, data.frame, or data.table) containing only entries of class numeric.
center.method	The method used for choosing data nugget centers. Must be 'mean' or 'random'. 'mean' chooses the data nugget center to be the mean of all observations within that data nugget, while 'random' chooses the data nugget center to be some random observation within that data nugget.
R	The number of observations to sample from the data matrix when creating the initial data nugget centers. Must be of class numeric within [100,10000].
delete.percent	The proportion of observations to remove from the data matrix at each iteration when finding data nugget centers. Must be of class numeric and within (0,1).
DN.num1	The number of initial data nugget centers to create. Must be of class numeric.
DN.num2	The number of data nuggets to create. Must be of class numeric.
dist.metric	The distance metric used to create the initial centers of data nuggets. Must be 'euclidean' or 'manhattan'.
seed	Random seed for replication. Must be of class numeric.
no.cores	Number of cores used for parallel processing. If '0' then parallel processing is not used. Must be of class numeric.
make.pbs	Print progress bars? Must be TRUE or FALSE.

Details

Data nuggets are a representative sample meant to summarize Big Data by reducing a large dataset to a much smaller dataset by eliminating redundant points while also preserving the peripheries of the dataset. Each data nugget is defined by a center (location), weight (importance), and scale (internal variability). This function creates data nuggets using Algorithm 1 provided in the reference.

Value

An object of class datanugget:

Data Nuggets	DN.num by (ncol(x)+3) data frame containing the information for the data nuggets created (index, center, weight, scale).
Data Nugget Assignments	Vector of length nrow(x) containing the data nugget assignment of each observation in x.

Author(s)

Traymon Beavers, Javier Cabrera, Mariusz Lubomirski

References

Beavers, T. E., Cheng, G., Duan, Y., Cabrera, J., Lubomirski, M., Amaratunga, D., & Teigler, J. E. (2024). Data Nuggets: A Method for Reducing Big Data While Preserving Data Structure. Journal of Computational and Graphical Statistics, 1-21.

Cherasia, K. E., Cabrera, J., Fernholz, L. T., & Fernholz, R. (2022). Data Nuggets in Supervised Learning. *In Robust and Multivariate Statistical Methods: Festschrift in Honor of David E. Tyler* (pp. 429-449). Cham: Springer International Publishing.

Examples

```
## small example
X = cbind.data.frame(rnorm(10^3),
                     rnorm(10^3),
                     rnorm(10^3))

suppressMessages({

  my.DN = create.DN(x = X,
                    R = 500,
                    delete.percent = .1,
                    DN.num1 = 500,
                    DN.num2 = 250,
                    no.cores = 0,
                    make.pbs = FALSE)

})

my.DN$`Data Nuggets`
my.DN$`Data Nugget Assignments`



## large example
X = cbind.data.frame(rnorm(5*10^4),
                     rnorm(5*10^4),
                     rnorm(5*10^4),
                     rnorm(5*10^4),
                     rnorm(5*10^4))

my.DN = create.DN(x = X,
                  R = 5000,
                  delete.percent = .9,
                  DN.num1 = 10^4,
                  DN.num2 = 2000,
                  no.cores = 2)

my.DN$`Data Nuggets`
my.DN$`Data Nugget Assignments`
```

`create.DNcenters` *Create Data Nugget Centers*

Description

This function creates the centers of data nuggets from a random sample.

Usage

```
create.DNcenters(RS,
                 delete.percent,
                 DN.num,
                 dist.metric,
                 make.pb = FALSE)
```

Arguments

`RS`	A data matrix (data frame, data table, matrix, etc) containing only entries of class numeric.
`delete.percent`	The proportion of observations to remove from the data matrix at each iteration when finding data nugget centers. Must be of class numeric and within (0,1).
`DN.num`	The number of data nuggets to create. Must be of class numeric.
`dist.metric`	The distance metric used to create the initial centers of data nuggets. Must be 'euclidean' or 'manhattan'.
`make.pb`	Print progress bar? Must be TRUE or FALSE.

Details

This function is used for reducing a random sample to data nugget centers in the create.DN function. NOTE THAT THIS FUNCTION IS NOT DESIGNED FOR USE OUTSIDE OF THE create.DN FUNCTION.

Value

`DN.data`	DN.num by (ncol(RS)) data frame containing the data nugget centers.

Author(s)

Traymon Beavers, Javier Cabrera, Mariusz Lubomirski

References

Beavers, T. E., Cheng, G., Duan, Y., Cabrera, J., Lubomirski, M., Amaratunga, D., & Teigler, J. E. (2024). Data Nuggets: A Method for Reducing Big Data While Preserving Data Structure. Journal of Computational and Graphical Statistics, 1-21.

Cherasia, K. E., Cabrera, J., Fernholz, L. T., & Fernholz, R. (2022). Data Nuggets in Supervised Learning. *In Robust and Multivariate Statistical Methods: Festschrift in Honor of David E. Tyler* (pp. 429-449). Cham: Springer International Publishing.

`create_refine.DN` *Create and Refine Data Nuggets in one function*

Description

This function combines creating and refining data nuggets in one function. It's a wrapper function for `create.DN` and `refine.DN`.

Usage

```
create_refine.DN(x,
                 center.method = "mean",
                 R = 5000,
                 delete.percent = .1,
                 DN.num1 = 10^4,
                 DN.num2 = 2000,
                 dist.metric = "euclidean",
                 seed = 291102,
                 no.cores = (detectCores() - 1),
                 make.pbs = TRUE,
                 EV.tol = .9,
                 max.splits = 5,
                 min.nugget.size = 2,
                 delta = 2)
```

Arguments

`x`	A data matrix (of class matrix, data.frame, or data.table) containing only entries of class numeric.
`center.method`	The method used for choosing data nugget centers. Must be 'mean' or 'random'. 'mean' chooses the data nugget center to be the mean of all observations within that data nugget, while 'random' chooses the data nugget center to be some random observation within that data nugget.
`R`	The number of observations to sample from the data matrix when creating the initial data nugget centers. Must be of class numeric within [100,10000].
`delete.percent`	The proportion of observations to remove from the data matrix at each iteration when finding data nugget centers. Must be of class numeric and within (0,1).
`DN.num1`	The number of initial data nugget centers to create. Must be of class numeric.
`DN.num2`	The number of data nuggets to create. Must be of class numeric.
`dist.metric`	The distance metric used to create the initial centers of data nuggets. Must be 'euclidean' or 'manhattan'.

seed	Random seed for replication. Must be of class numeric.
no.cores	Number of cores used for parallel processing. If '0' then parallel processing is not used. Must be of class numeric.
make.pbs	Print progress bars? Must be TRUE or FALSE.
EV.tol	A value designating the percentile for finding the corresponding quantile that will designate how large the largest eigenvalue of the covariance matrix of a data nugget can be before it must be split. Must be of class numeric and within (0,1).
max.splits	A value designating the maximum amount of attempts that will be made to split data nuggets according to their largest eigenvalue before the algorithm breaks. Must be of class numeric.
min.nugget.size	A value designating the minimum amount of observations a data nugget created from a split must contain. Must be of class numeric and greater than 1.
delta	Ratio between the first and second eigenvalues of the covariance matrix of a data nugget to force its split. Default is 2.

Details

Data nuggets are a representative sample meant to summarize Big Data by reducing a large dataset to a much smaller dataset by eliminating redundant points while also preserving the peripheries of the dataset. Each data nugget is defined by a center (location), weight (importance), and scale (internal variability). This function combines creating and refining data nuggets in one function. It's a wrapper function for `create.DN` and `refine.DN`.

Value

An object of class datanugget:

Data Nuggets	DN.num by (ncol(x)+3) data frame containing the information for the data nuggets created (index, center, weight, scale).
Data Nugget Assignments	Vector of length nrow(x) containing the data nugget assignment of each observation in x.

Author(s)

Traymon Beavers, Javier Cabrera, Mariusz Lubomirski

References

Beavers, T. E., Cheng, G., Duan, Y., Cabrera, J., Lubomirski, M., Amaratunga, D., & Teigler, J. E. (2024). Data Nuggets: A Method for Reducing Big Data While Preserving Data Structure. Journal of Computational and Graphical Statistics, 1-21.

Cherasia, K. E., Cabrera, J., Fernholz, L. T., & Fernholz, R. (2022). Data Nuggets in Supervised Learning. *In Robust and Multivariate Statistical Methods: Festschrift in Honor of David E. Tyler* (pp. 429-449). Cham: Springer International Publishing.

Examples

```
## small example
X = cbind.data.frame(rnorm(10^3),
                     rnorm(10^3),
                     rnorm(10^3))

suppressMessages({

  my.DN = create_refine.DN(x = X,
                    R = 500,
                    delete.percent = .1,
                    DN.num1 = 500,
                    DN.num2 = 250,
                    no.cores = 0,
                    make.pbs = FALSE,
                    EV.tol = .9,
                    min.nugget.size = 2,
                    max.splits = 5,
                    delta = 2)

})

my.DN$`Data Nuggets`
my.DN$`Data Nugget Assignments`



## large example
X = cbind.data.frame(rnorm(5*10^4),
                     rnorm(5*10^4),
                     rnorm(5*10^4),
                     rnorm(5*10^4),
                     rnorm(5*10^4))

my.DN = create_refine.DN(x = X,
                    R = 5000,
                    delete.percent = .9,
                    DN.num1 = 10^4,
                    DN.num2 = 2000,
                    no.cores = 2,
                    EV.tol = .9,
                    min.nugget.size = 2,
                    max.splits = 5,
                    delta = 2)

my.DN$`Data Nuggets`
my.DN$`Data Nugget Assignments`
```

refine.DN *Refine Data Nuggets*

Description

This function refines the data nuggets found in an object of class datanugget created using the create.DN function.

Usage

```
refine.DN(x,
          DN,
          EV.tol = .9,
          max.splits = 5,
          min.nugget.size = 2,
          delta = 2,
          seed = 291102,
          no.cores = (detectCores() - 1),
          make.pbs = TRUE)
```

Arguments

x	A data matrix (data frame, data table, matrix, etc.) containing only entries of class numeric.
DN	An object of class data nugget created using the create.DN function.
EV.tol	A value designating the percentile for finding the corresponding quantile that will designate how large the largest eigenvalue of the covariance matrix of a data nugget can be before it must be split. Must be of class numeric and within (0,1).
max.splits	A value designating the maximum amount of attempts that will be made to split data nuggets according to their largest eigenvalue before the algorithm breaks. Must be of class numeric.
min.nugget.size	A value designating the minimum amount of observations a data nugget created from a split must contain. Must be of class numeric and greater than 1.
delta	Ratio between the first and second eigenvalues of the covariance matrix of a data nugget to force its split. Default is 2.
seed	Random seed for replication. Must be of class numeric.
no.cores	Number of cores used for parallel processing. If '0' then parallel processing is not used. Must be of class numeric.
make.pbs	Print progress bars? Must be TRUE or FALSE.

Details

Data nuggets can be refined by attempting to make all of the data nugget shapes as spherical as possible. This is achieved by designating an eigenvalue tolerance (EV.tol) which is used to give a lower threshold for a data nugget's deviation from sphericity, respectively.

If the largest eigenvalue of a data nugget's covariance matrix has a ratio greater than the quantile associated with the percentile given by EV.tol, this data nugget is split into two smaller data nuggets using K-means clustering.

However, if either of the two data nuggets created by this split have less than the designated minimum data nugget size (min.nugget.size), then the split is cancelled and the data nugget remains as is. This function refines data nuggets using Algorithm 2 provided in the reference.

Updated: When data nuggets are not spherical, with the ratio between the first and second eigenvalues of the covariance matrix of the data nugget is greater than `delta` (its default value is 2), the data nugget is split.

Value

An object of class datanugget:

`Data Nuggets` DN.num by (ncol(x)+3) data frame containing the information for the data nuggets created (index, center, weight, scale).

`Data Nugget Assignments`
Vector of length nrow(x) containing the data nugget assignment of each observation in x.

Author(s)

Traymon Beavers, Javier Cabrera, Mariusz Lubomirski

References

Beavers, T. E., Cheng, G., Duan, Y., Cabrera, J., Lubomirski, M., Amaratunga, D., & Teigler, J. E. (2024). Data Nuggets: A Method for Reducing Big Data While Preserving Data Structure. Journal of Computational and Graphical Statistics, 1-21.

Cherasia, K. E., Cabrera, J., Fernholz, L. T., & Fernholz, R. (2022). Data Nuggets in Supervised Learning. *In Robust and Multivariate Statistical Methods: Festschrift in Honor of David E. Tyler* (pp. 429-449). Cham: Springer International Publishing.

Examples

```
## small example
X = cbind.data.frame(rnorm(10^3),
                     rnorm(10^3),
                     rnorm(10^3))

suppressMessages({

  my.DN = create.DN(x = X,
                    R = 500,
```

```
                    delete.percent = .1,
                    DN.num1 = 500,
                    DN.num2 = 250,
                    no.cores = 0,
                    make.pbs = FALSE)

  my.DN2 = refine.DN(x = X,
                     DN = my.DN,
                     EV.tol = .9,
                     min.nugget.size = 2,
                     max.splits = 5,
                     no.cores = 0,
                     make.pbs = FALSE)

})

my.DN2$`Data Nuggets`
my.DN2$`Data Nugget Assignments`



  ## large example
  X = cbind.data.frame(rnorm(5*10^4),
                       rnorm(5*10^4),
                       rnorm(5*10^4),
                       rnorm(5*10^4),
                       rnorm(5*10^4))

  my.DN = create.DN(x = X,
                    R = 5000,
                    delete.percent = .9,
                    DN.num1 = 10^4,
                    DN.num2 = 2000,
                    no.cores = 2)

  my.DN2 = refine.DN(x = X,
                     DN = my.DN,
                     EV.tol = .9,
                     min.nugget.size = 2,
                     max.splits = 5,
                     no.cores = 2)

  my.DN2$`Data Nuggets`
  my.DN2$`Data Nugget Assignments`
```

Index