

The SeqKnn Package

May 9, 2008

Version 1.0.1

Date 2008-5-7

Title Sequential KNN imputation method

Author Ki-Yeol Kim and Gwan-Su Yi, CSBio lab., Information and Communications University

Maintainer Taeho Hwang <fanteo@icu.ac.kr>

Depends R (>= 2.0.0)

Description This function estimates missing values sequentially from the gene that had least missing rate in microarray data

License GPL (>= 2)

URL <http://csbio.icu.ac.kr>

R topics documented:

SeqKNN	1
khan05	2
nnmiss	3

Index	5
--------------	----------

SeqKNN	<i>Sequential KNN imputation method</i>
--------	---

Description

This function estimates missing values sequentially from the gene that has least missing rate in microarray data, using weighted mean of k nearest neighbors. This function requires 'nnmiss'.

Usage

```
SeqKNN(data, k)
```

Arguments

data	matrix or dataframe, 1 row corresponds to 1 gene, 1 column to 1 sample,colnames and rownames can be used
k	number of nearest neighbors

Details

'SeqKNN' separates the dataset into incomplete and complete set that has or has not missing values respectively. The genes in incomplete set are imputed by the order of missing rate. Missing value is filled by the weighted mean value of corresponding column of the nearest neighbor genes in complete set. Once all missing values in a gene are imputed, the imputed gene is moved into the complete set and used for the imputation of the rest of genes in incomplete set. In this process, all missing values in one gene can be imputed simultaneously from the selected neighbor genes in complete set. This reduces execution time from previously developed KNN method that selects nearest neighbors for each imputation.

Author(s)

Ki-Yeol Kim and Gwan-Su Yi

References

Ki-Yeol Kim, Byoung-Jin Kim, Gwan-Su Yi (2004.Oct.26) "Reuse of imputed data in microarray analysis increases imputation efficiency", BMC Bioinformatics 5:160.

Examples

```
## Not run:  
data(khan05)  
imputedData<-SeqKNN(khan05,10)  
## End(Not run)
```

khan05

Khan et al.'s Small Round Blood Cell Tumor(SRBCT) data

Description

SRBCT dataset has 2308 genes and 63 experimental conditions, 8 Burkitt Lymphoma (BL), 23 Ewing Sarcoma (EWS), 12 neuroblastoma (NB), and 20 rhabdomyosarcoma (RMS) (Khan et al., 2001).

Usage

```
data(khan05)
```

Arguments

`khan05` data frame generated 5 percent missing entries randomly from original SRBCT data.

References

Javed Khan, Jun S. Wei, Markus Ringner, Lao H. Saal, Marc Ladanyi, Frank Westermann, Frank Berthold, Manfred Schwab, Cristina R. Antonescu, Carsten Peterson, and Paul S. Meltzer (2001). Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature Medicine*, Volume 7, Number 6, June

Examples

```
data(khan05)
imputedKhan<-SeqKNN(khan05,10)
```

`nnmiss`

Selects k nearest neighbors and calculates weighted mean of them

Description

A function to select k nearest neighbors using Euclidean distance, and estimate missing value with weighted mean of selected neighbors.

Usage

```
nnmiss(x, xmiss, ismiss, K)
```

Arguments

`x` data frame which contains only complete cases
`xmiss` data frame which contains incomplete cases
`ismiss` data frame with logical value(TRUE or FALSE) of `xmiss`
`K` number of nearest neighbors

Details

Appropriate number of k is 10-20. However, we need to control k smaller in case missing rate is high, especially k is larger than the size of complete set.

Examples

```
## Not run:
data(khan05)
x <- as.matrix(khan05)
N <- dim(x)
p <- N[2]
N <- N[1]
nas <- is.na(drop(x %*% rep(1, p)))
xcomplete <- x[!nas, ]          ## complete set
xbad <- x[nas, , drop = FALSE]  ## incomplete set
xnas <- is.na(xbad)
xbadhat <- xbad
xbadhat[1,]<-nnmiss(xcomplete, xbad[1,], xnas[1,], 10)
## End(Not run)
```

Index

*Topic **classes**
 nnmiss, 3
 SeqKNN, 1
*Topic **datasets**
 khan05, 2

khan05, 2

nnmiss, 3

SeqKNN, 1