Package 'GridOnClusters'

April 6, 2020

Type Package

Title Joint Discretization of Data on a Grid that Preserves Clusters

Version 0.0.7

Date 2020-04-03

Author Jiandong Wang [aut],
Sajal Kumar [aut] (<https://orcid.org/0000-0003-0930-1582>),
Joe Song [aut, cre] (<https://orcid.org/0000-0002-6883-6547>)

Maintainer Joe Song `<joemsong@cs.nmsu.edu>`

Description Discretize multivariate continuous data using a grid that captures the joint distribution via preserving clusters in the original data. Joint grid discretization is applicable as a data transformation step before using other methods to infer association, function, or causality without assuming a parametric model.

Imports Rcpp, cluster, fossil, dqrng

Suggests Ckmeans.1d.dp, FunChisq, knitr, testthat (>= 2.1.0), rmarkdown

Depends R (>= 3.0)

License LGPL (>= 3)

Encoding UTF-8

LazyData true

LinkingTo Rcpp

RoxygenNote 7.1.0

NeedsCompilation yes

VignetteBuilder knitr

Repository CRAN

Date/Publication 2020-04-06 12:42:11 UTC

R topics documented:

discretize.jointly — *Discretize Multivariate Continuous Data by a Cluster-Preserving Grid*

Description

Discretize multivariate continuous data using a grid that captures the joint distribution via preserving clusters in the original data

Usage

```
discretize.jointly(data, k = c(2:10), cluster_label = NULL)
```

Arguments

`data` a matrix containing two or more continuous variables. Columns are variables, rows are observations.

`k` either the number or range of clusters to be found on `data`. The default is 2 to 10 clusters. If a range is specified, an optimal k in the range is chosen to maximize the average silhouette width. If `cluster_label` is specified, `k` is ignored.

`cluster_label` a vector of user-specified cluster labels for each observation in `data`. The user is free to choose any clustering. If unspecified, k-means clustering is used by default.

Value

A list that contains four items:

`D` a matrix that contains the discretized version of the original `data`. Discretized values are one(1)-based.

`grid` a list of vectors containing decision boundaries for each variable/dimension.

`clabels` a vector containing cluster labels for each observation in `data`.

`csimilarity` a similarity score between clusters from joint discretization `D` and cluster labels `clabels`. The score is the adjusted Rand index.

See Also

See Ckmeans.1d.dp for discretizing univariate continuous data.

Examples

```
# using a specified k
x = rnorm(100)
y = sin(x)
z = cos(x)
data = cbind(x, y, z)
discretized_data = discretize.jointly(data, k=5)$D

# using a range of k
x = rnorm(1000)
y = log1p(abs(x))
z = tan(x)
data = cbind(x, y, z)
discretized_data = discretize.jointly(data, k=c(3:10))$D

# using an alternate clustering method to k-means
library(cluster)
x = rnorm(1000)
y = log1p(abs(x))
z = sin(x)
data = cbind(x, y, z)

# pre-cluster the data using partition around medoids (PAM)
cluster_label = pam(x=data, diss = FALSE, metric = "euclidean", k = 5)$clustering
discretized_data = discretize.jointly(data, cluster_label = cluster_label)$D
```

Index