
LKCA User's Manual in R

Public domain version 1.0 for R

Release date: 12/01/2016

Fuyuan Cao¹, Haini Li¹, Liqin Yu¹, Joshua Zhexue Huang², Jiye Liang¹

¹Shanxi University, Taiyuan, China

²Shenzhen University, Shenzhen, China

cfy@sxu.edu.cn, 851657890@qq.com, 1367545521@qq.com,
zx.huang@szu.edu.cn, lji@sxu.edu.cn

LKCA is a software toolbox for clustering algorithms. It provides the open-source package for use in R that implements the k-modes-type clustering algorithms for categorical data. It is designed to facilitate the development of new algorithms in this research.

Copyright(C) 2016 Fuyuan Cao, Haini Li, Liqin Yu, Joshua Zhexue Huang, Jiye Liang.

This program is a free software: it can be redistributed and/or modified under the terms of the GUN General Public License as published by the Free Software Foundation, either version 3 of the License or (at your option) any later version.

This program is distributed in the hope that it will be useful, but WITHOUT ANY WARRANTY; without even the implied warranty of MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE. See the GUN General Public License for more details.

.

CONTENTS

1 Overview of LKCA	1
1.1 Introduction	1
1.2 Architecture of LKCA	1
1.3 Core function overview	2
2 Setup in R	3
2.1 Getting and installing	3
2.2 Data format	3
3 Core functions	4
3.1 Hard and Fuzzy k-modes clustering algorithm	4
3.1.1 K-modes-S	4
3.1.2 K-modes-M	5
3.1.3 K-modes-D	6
3.1.4 F-K-modes-S	8
3.1.5 F-K-modes-M	9
3.1.6 F-K-modes-D	10
3.2 Hard and Fuzzy SV-k-modes clustering algorithm	11
3.2.1 SV-K-modes -S	11
3.2.2 SV-K-modes-M	12
3.2.3 SV-K-modes-D	13
3.2.4 F-SV-K-modes-S	14
3.2.5 F-SV-K-modes-M	16
3.2.6 F-SV-K-modes-D	17
3.3 Initial class center selection function	18
3.3.1 K-modes-k-initial-center	18
3.3.2 SV-K-modes-k-initial-center	19
3.4 Sub function	20
3.4.1 K-dis	20
3.4.2 K-center	21
3.4.3 SV-K-dis	22
3.4.4 SV-K-center	23

1 Overview of LKCA

1.1 Introduction

The k-means algorithm is well known for its efficiency in clustering the large data set, however, working only on the numeric data limits the use of the k-means and its variants on the categorical data. To solve this problem, Huang presented a k-modes algorithm, which extends the k-means algorithm by using a simple matching dissimilarity measure for categorical objects, modes instead of means for clusters, and a frequency-base method to update modes in the clustering process to minimize the clustering cost function.

In real applications, some objects maybe have more than one value for an attribute. For instance, many people may have several job titles and more than one hobby or email. Such a data representation is very widespread in questionnaire, banking, insurance, retails, and medical databases. To cluster the categorical data whose each object has more than one value for an attribute, Cao proposed a SV-k-modes algorithm.

As a kind of partitioning algorithm, the k-modes and its variants have been widely used in many applications because of its efficiency. However, to the best of our knowledge, there is no comprehensive open-source packages existing for implementing these algorithms. To achieve the goal, we develop the open-source library called LKCA (Library of k-modes-type Clustering Algorithm).

The main contribution of the LKCA library concludes two aspects. (1) It is the first comprehensive open-source library for clustering categorical data. (2) It is written in R that is easy to use and completely open source. We hope researchers can use the LKCA conveniently and share their algorithms through the framework.

1.2 Architecture of LKCA

The LKCA architecture is composed of four modules, that is, the k-modes algorithm, the fuzzy k-modes algorithm, the SV-k-modes algorithm and the fuzzy SV-k-modes algorithm, as shown in Figure 1. The four modules in the LKCA architecture are designed independently, and all codes follow the R standards.

In each module, LKCA provides three patterns to cluster categorical data, including single-threaded, multi-threaded and distributed computation. In the multi-threaded operation, it is provided with multiple CPU to execute multiple threads at the same time, which equivalently creates a set of functions running in parallel. Through the multi-thread operation, it will improve the overall processing performance. In addition, by using distributed computing technology, the task will be decomposed into a number of small parts, and assigned to multiple computers for processing, which can save the overall computing time, and greatly improve the computational efficiency.

The implementation of the clustering algorithms depends on these sub functions, including the distance function, finding cluster centers function, and initial cluster center selection function.

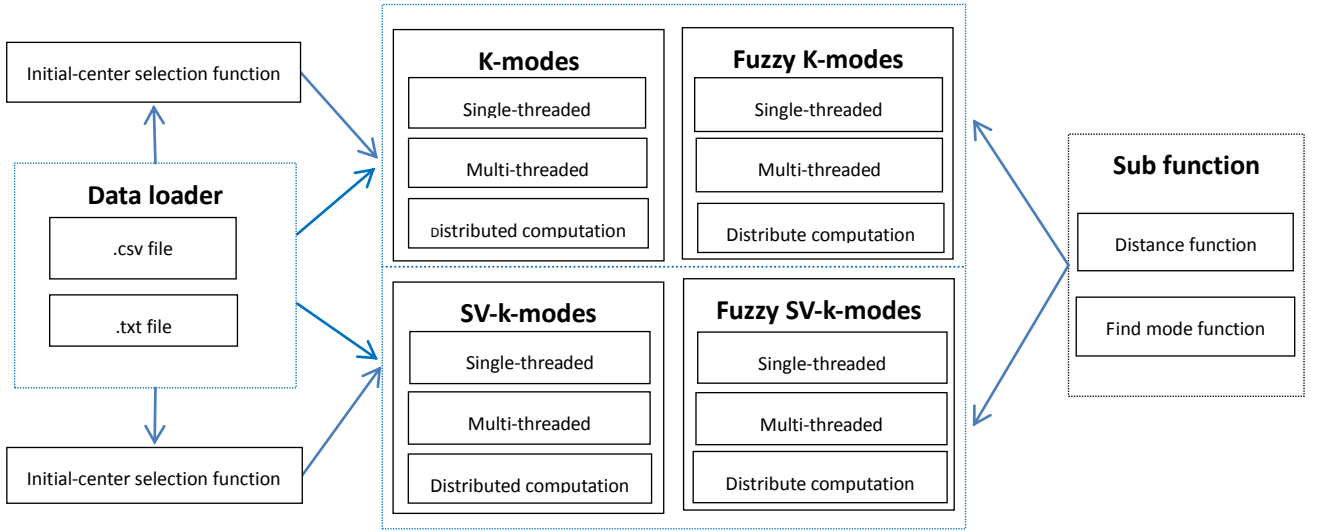


Figure 1 The architecture of LKCA

1.3 Core functions overview

The core functions in the LKCA library are listed in Table 1.

Table 1 Core functions

R Function	Corresponding Algorithm
K-modes-S	K-modes algorithm(Single-threaded)
K-modes-M	K-modes algorithm(Multi-threaded)
K-modes-D	K-modes algorithm(Distributed Computation)
F-K-modes-S	Fuzzy K-modes algorithm(Single-threaded)
F-K-modes-M	Fuzzy K-modes algorithm(Multi-threaded)
F-K-modes-D	Fuzzy K-modes algorithm(Distributed Computation)
K-dis	Distance function of K-modes algorithm
K-center	Find mode function of K-modes algorithm
SV-K-modes-S	SV-K-modes algorithm(Single-threaded)
SV-K-modes-M	SV-K-modes algorithm(Multi-threaded)
SV-K-modes-D	SV-K-modes algorithm(Distributed Computation)
F-SV-K-modes-S	Fuzzy SV-K-modes algorithm(Single-threaded)
F-SV-K-modes-M	Fuzzy SV-K-modes algorithm(Multi-threaded)
F-SV-K-modes-D	Fuzzy SV-K-modes algorithm(Distributed Computation)
SV-K-dis	Distance function of SV-K-modes algorithm
SV-K-center	Find mode function of SV-K-modes algorithm
K-modes-k-initial-center	Initial cluster centers of K-modes
SV-K-modes-k-initial-center	Initial cluster centers of SV-K-modes

2 Setup in R

2.1 Getting and installing

To run the package, it is required that (1) Windows 7 operating system or above the version; (2) R 2.14.0 or above to be installed; (3) RStudio. The package should be loaded to the folder of current path, as shown in Figure 2.

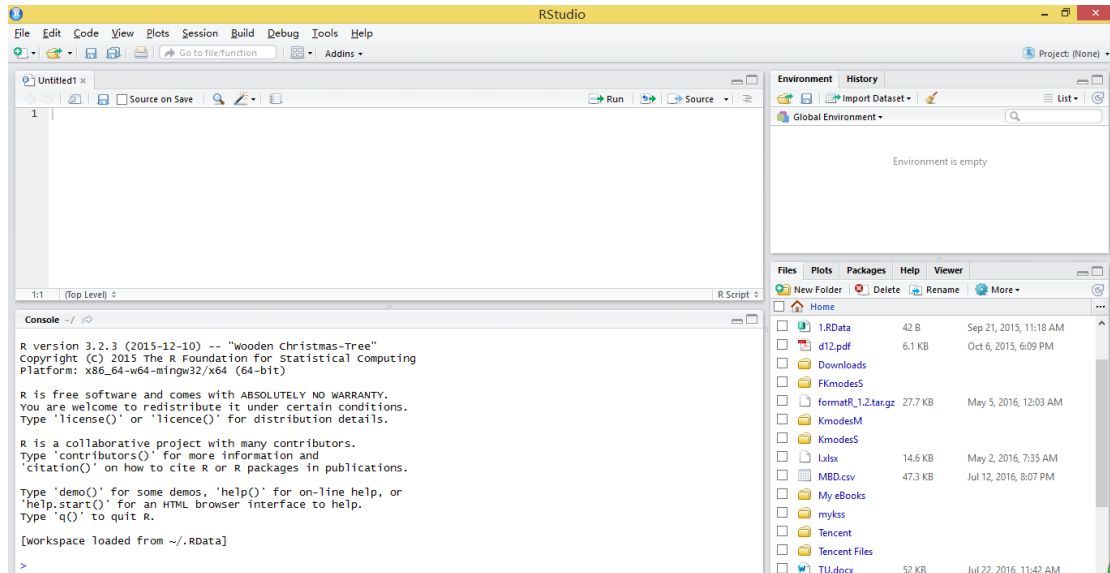


Figure 2 Loading the package

To use RStudio, you must install R in advance. In this paper, we use R to complete our work, as shown in Figure3. And we choose to use RStudio, which is a powerful and useful tool for R.

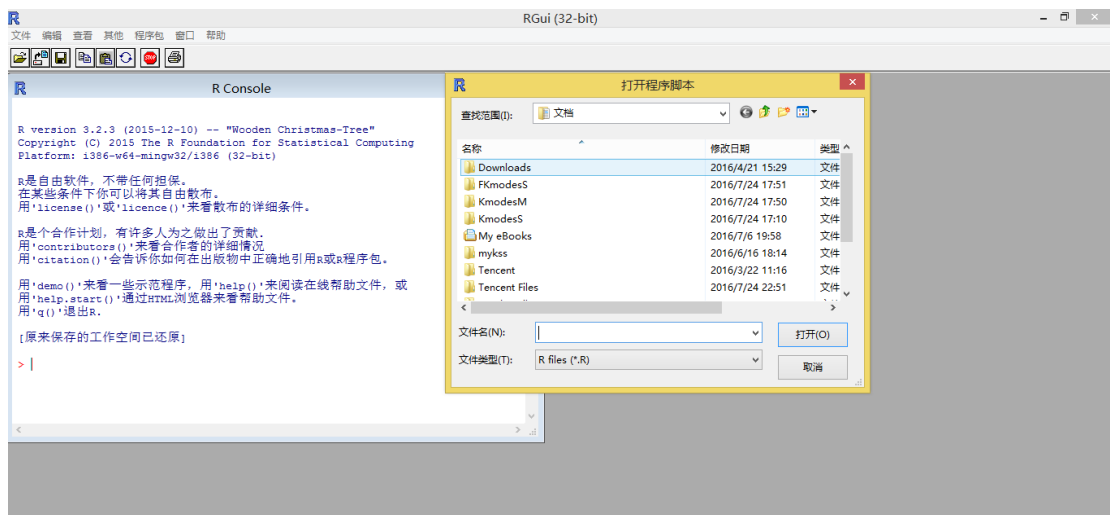


Figure 3 The Interface of R

2.2 Data format

It is required that a data set should be correctly imported to RStudio. Currently, R

supports the .csv, .txt file, et al..

An example in RStudio is given in Figure 4. The lines represent the number of instances, and the columns represent the number of attributes. “header=FALSE or TRUE” represents that whether the first line is used as the column name. If the first line is used as the column name, we will input “header=True”. If not, input “header=FALSE”.

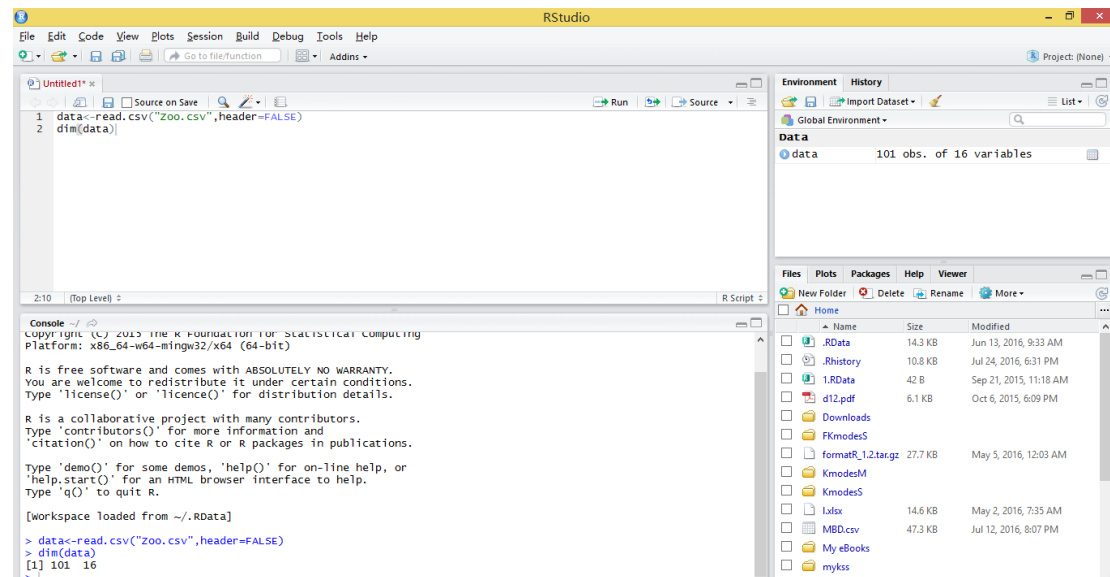


Figure 4 Example of Data format

It can be seen that the data set contains 101 data instances and 16 attributes. “header=FALSE” represents that the first line is not used as the column name.

3 Core functions

3.1 Hard and Fuzzy k-modes clustering algorithm

3.1.1 K-modes-S

Description

Implement the k-modes algorithm for single-threaded.

Usage

Hard-K-Mode (data, K, InitialCenters)

return (iter, cid, time)

Arguments

Inputs	data	Data set to be clustered
	K	Number of clusters
	InitialCenters	Initial cluster centers
Outputs	iter	Iterations
	cid	Clustering result
	time	Computational cost

Details

The k-modes algorithm extends the k-means algorithm by using a simple matching dissimilarity measure for categorical objects, modes instead of means for clusters, and a frequency-base method to update modes in the clustering process to minimize the clustering cost function.

Reference

Z. Huang, Extensions to the k-means algorithm for clustering large data sets with categorical values, Data Mining and Knowledge Discovery 2 (3) (1998) 283-304.

Example

```
>data=read.csv ( "Zoo.csv", header=FALSE )
>K=7
>InitialCenters=NULL
>library ( KmodesS )
>Hard-K-Mode ( data, K, InitialCenters )
```

Results

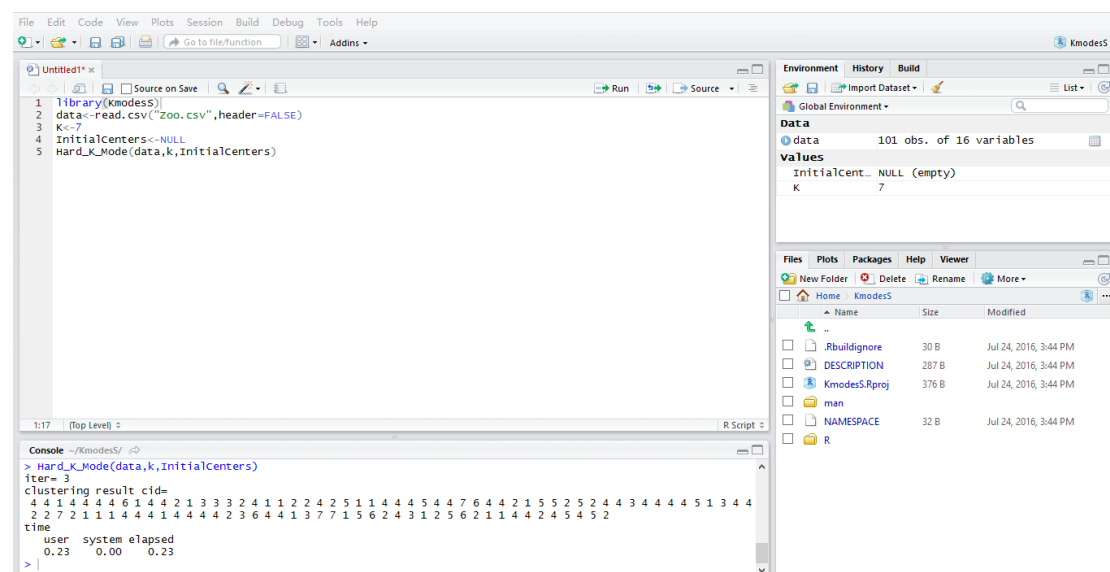


Figure 5 Example of the K-modes -S function

3.1.2 K-modes-M

Description

Implement the k-modes algorithm for multi-thread operation.

Usage

```
Hard-K-Mode-M ( data, K, InitialCenters )
return ( iter, cid, time )
```

Arguments

Inputs	data	Data set to be clustered
	K	Number of clusters
	InitialCenters	Initial cluster centers
Outputs	iter	Iterations
	cid	Clustering result
	time	Computational cost

Details

Multi-thread operation is used to reducing elapsed times by the multi-core machines and clusters of categorical data. The operation is provided with multiple CPU to execute multiple threads at the same time which equivalently creates a set of copies of KmodesS package running in parallel. Through the multi-thread operation, the overall processing performance can be improved.

Reference

Z. Huang, Extensions to the k-means algorithm for clustering large data sets with categorical values, Data Mining and Knowledge Discovery 2 (3) (1998) 283-304.

Example

```
>library ( KmodesM )
>data=read.csv ( " Zoo.csv ", header=FALSE )
>K=7
>InitialCenters=NULL
>Hard-K-Mode-M ( data, K, InitialCenters )
```

Results

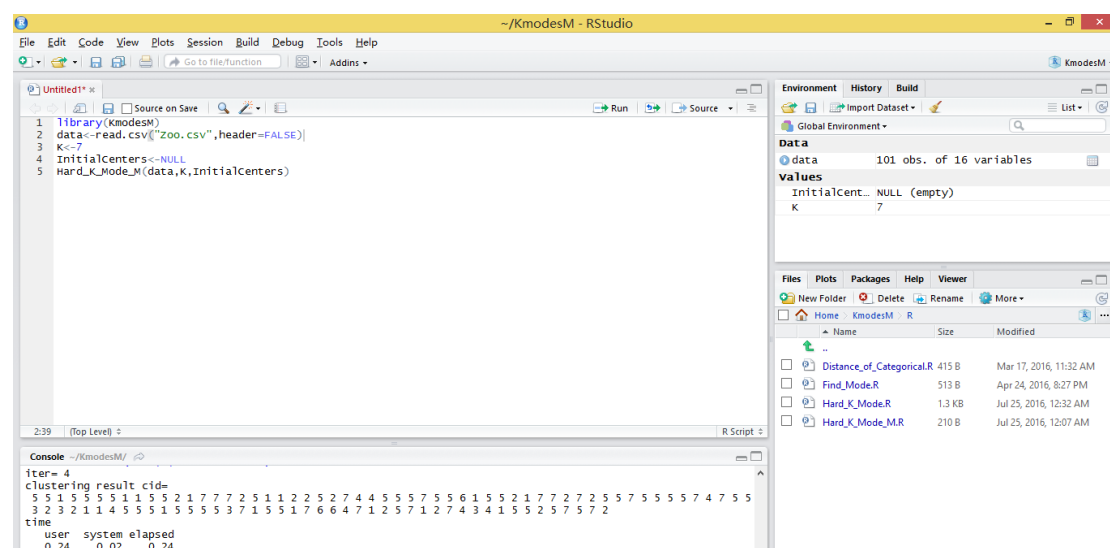


Figure 6 Example of the K-modes -M function

3.1.3 K-modes-D

Description

Implement the k-modes algorithm for distributed computing.

Usage

Hard-K-Mode-D (data, K, InitialCenters)

return (nj)

Arguments

Inputs	data	Data set to be clustered
	K	Number of clusters
	InitialCenters	Initial cluster centers
Outputs	nj	Clustering result

Details

The distributed computing of the k-modes algorithm relies on the RHadoop, which is a tool that connects R and Hadoop to realize the distributed computing of massive data.

Reference

Z. Huang, Extensions to the k-means algorithm for clustering large data sets with categorical values, Data Mining and Knowledge Discovery 2 (3) (1998) 283-304.

Example

```
>library ( KmodesD )
>data=read.csv ( " Zoo.csv ", header=FALSE )
>K=7
>InitialCenters=NULL
>Hard-K-Mode-D ( data, K, InitialCenters )
```

Results

```
$key
NULL

$val
[,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11] [,12] [,13] [,14]
[1,] 0 1 1 0 1 0 0 0 1 1 0 0 2 1
[2,] 1 0 0 1 0 0 1 1 1 1 0 0 4 1
[3,] 0 0 1 0 0 1 1 1 1 0 0 1 0 1
[4,] 0 0 0 1 0 1 1 1 1 1 0 1 0 1
[5,] 0 1 1 0 1 1 0 0 1 1 0 0 2 1
[6,] 0 0 1 0 0 1 1 0 1 1 0 0 0 1
[7,] 0 0 1 0 1 0 0 0 1 1 0 0 2 1
[,15] [,16]
[1,] 0 0
[2,] 0 1
[3,] 0 0
[4,] 0 1
[5,] 0 1
[6,] 0 0
[7,] 0 0
```

Figure 7 Example of the K-modes -D function

3.1.4 F-K-modes-S

Description

Implement the fuzzy k-modes algorithm for single-threaded.

Usage

Fuzzy-K-mode-S (data, K, InitialCenters, a)

return (iter, cid, time)

Arguments

Inputs	data	Data set to be clustered
	K	Number of clusters
	InitialCenters	Initial cluster centers
	a	Index
Outputs	cid	Clustering result
	time	Computational cost
	iter	Iterations

Details

The fuzzy k-modes algorithm aims to generate the fuzzy partition matrix from categorical data within the framework of the fuzzy k-means algorithm. The main result of the algorithm is to find the fuzzy cluster modes when the simple matching dissimilarity measure is used for categorical objects.

Reference

Z. Huang, M. K. Ng, A fuzzy k-modes algorithm for clustering categorical data, Fuzzy Systems, IEEE Transactions on 7 (4) (1999) 446-452.

Example

```
>data=read.csv ( "Zoo.csv", header=FALSE )
>K=7
>InitialCenters=NULL
>a=1.1
>library ( FKmodesS )
>Fuzzy-K-Mode-S ( data, K, InitialCenters, a )
```

Results

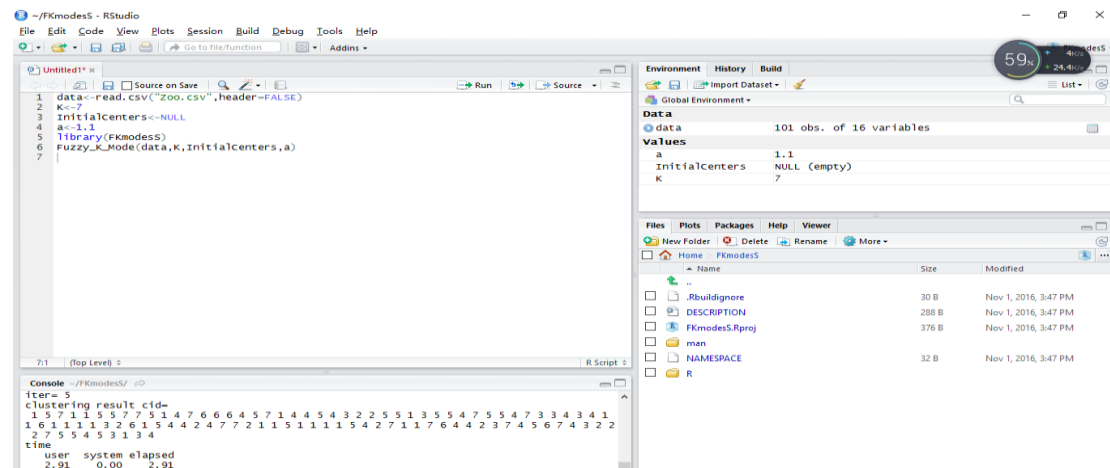


Figure 8 Example of the Fuzzy-K-Modes -S function

3.1.5 F-K-modes-M

Description

Implement the fuzzy k-modes algorithm for multi-thread operation.

Usage

Fuzzy-K-Mode-M (data, K, InitialCenters, a)
return(iter, cid, time)

Arguments

Inputs	data	Data set to be clustered
	K	Number of clusters
	InitialCenters	Initial cluster centers
	a	Index
Outputs	cid	Clustering result
	time	Computational cost
	iter	Iterations

Details

Please refer to the details of the F-K-modes-S mentioned above.

Reference

Z. Huang, M. K. Ng, A fuzzy k-modes algorithm for clustering categorical data, Fuzzy Systems, IEEE Transactions on 7 (4) (1999) 446-452.

Example

```
>data=read.csv ( "Zoo.csv", header=FALSE )
>K=7
>InitialCenters=NULL
>a=1.1
```

```
>library ( FKmodesM )
>Fuzzy-K-Mode-M ( data, K, InitialCenters, a )
```

Results

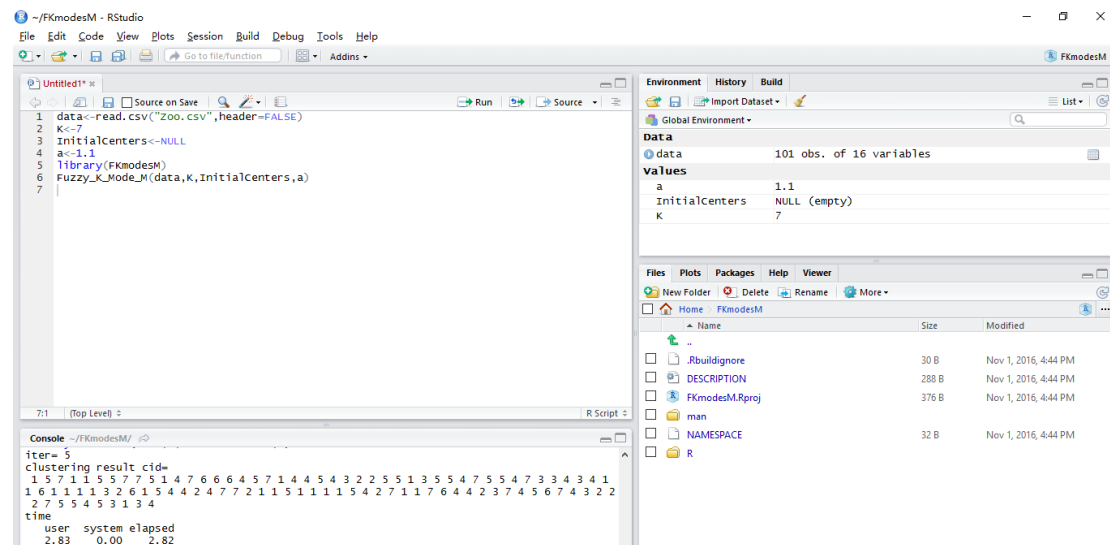


Figure 9 Example of the Fuzzy-K-modes -M function

3.1.6 F-K-modes-D

Description

Implement the fuzzy k-modes algorithm for distributed computing.

Usage

```
Fuzzy-K-Mode-D ( data, K, InitialCenters, a )
return ( iter, cid, time )
```

Arguments

Inputs	data	Data set to be clustered
	K	Number of clusters
	InitialCenters	Initial cluster centers
	a	index
Outputs	nj	Clustering result

Details

Please refer to the details mentioned above.

Reference

Z. Huang, M. K. Ng, A fuzzy k-modes algorithm for clustering categorical data, Fuzzy Systems, IEEE Transactions on 7 (4) (1999) 446-452.

Example

```
>data=read.csv ( " Zoo.csv ", header=FALSE)
```

```

>K=7
>InitialCenters=NULL
>a=1.1
>library ( FKmodesD )
>Fuzzy-K-Mode-D ( data, K, InitialCenters, a)

```

Results

```

$val
  [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11] [,12] [,13] [,14]
[1,]   0   0   1   0   0   1   1   1   1   1   0   0   4   0
[2,]   1   0   0   1   0   0   0   1   1   1   0   0   2   1
[3,]   0   0   1   0   0   1   1   1   1   0   0   1   0   1
[4,]   1   0   1   0   0   0   1   0   0   1   0   0   6   0
[5,]   1   0   0   1   0   0   1   1   1   1   0   0   4   1
[6,]   1   0   0   1   0   0   0   1   1   1   0   0   4   1
[7,]   1   0   0   1   0   0   0   1   1   1   0   0   4   1
  [,15] [,16]
[1,]    0    0
[2,]    0    0
[3,]    0    0
[4,]    0    0
[5,]    0    1
[6,]    0    1
[7,]    1    0
>

```

Figure 10 Example of the Fuzzy-K-modes -D function

3.2 Hard and Fuzzy SV-k-modes clustering algorithm

3.2.1 SV-K-modes -S

Description

Implement the SV-k-modes algorithm for single-threaded.

Usage

```

New-ratio-k-multi-modes ( Data, K, InitialCenters )
return ( time )

```

Arguments

Inputs	Data	Data set to be clustered
	K	Number of clusters
	InitialCenters	Initial cluster centers
Outputs	Time	Computational cost

Details

The SV-k-modes clustering algorithm is designed to cluster the objects with set-valued attributes. The algorithm can partition data with mixed single-valued and set-valued attributes and the k-modes algorithm is its special case.

Reference

F. Cao, Z. Huang, J. Liang, X. Zhao and Y. Meng, SV-k-modes: an algorithm for clustering categorical data with set-valued attributes, Neural Networks and Learning System, IEEE Transactions on (Under review).

Example

```
>library ( SVKS )
>Data=read.csv ( “ MBD.csv ” , header=FALSE )
>K=10
>InitialCenters=NULL
>New-ratio-k-multi-modes ( Data, K, InitialCenters )
```

Results

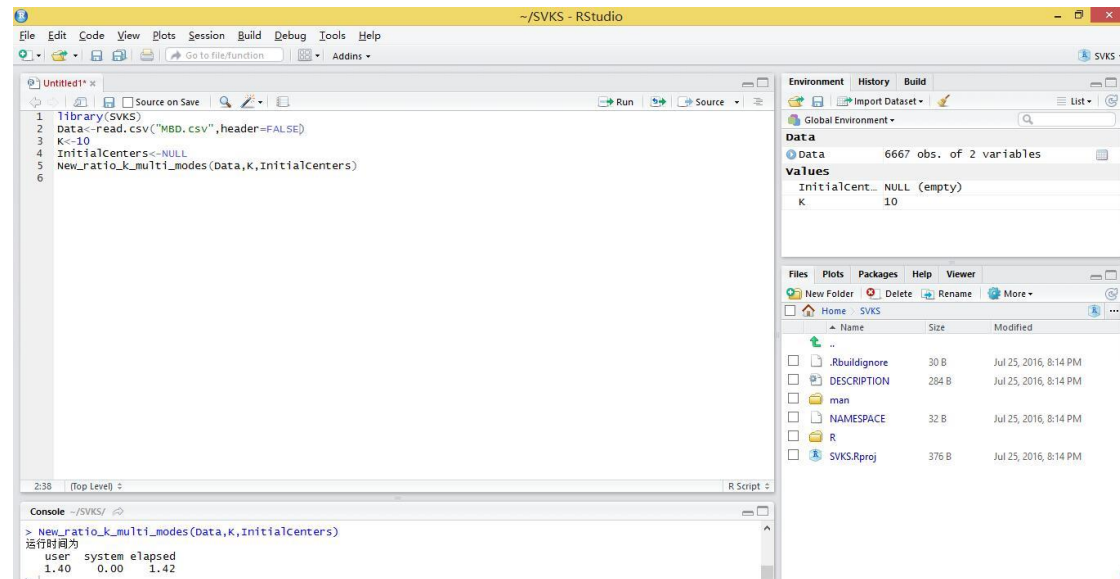


Figure 11 Example of the SV-K-modes -S function

3.2.2 SV-K-modes-M

Description

Implement the SV-k-modes algorithm for multi-threaded.

Usage

New-ratio-k-multi-modes-M (Data, K, InitialCenters)

Return (time)

Arguments

Inputs	Data	Data set to be clustered
	K	Number of clusters
	InitialCenters	Initial cluster centers
Outputs	time	Computational cost

Details

Please refer to the description mentioned above.

Reference

F. Cao, Z. Huang, J. Liang, X. Zhao and Y. Meng, SV-k-modes: an algorithm for clustering categorical data with set-valued attributes, Neural Networks and Learning System, IEEE Transactions on (Under review).

Example

```
>library ( SVKM )
>Data=read.csv ( " MBD.csv " , header=FALSE )
>K=10
>InitialCenters=NULL
>New-ratio-k-multi-modes ( Data, K, InitialCenters )
```

Results

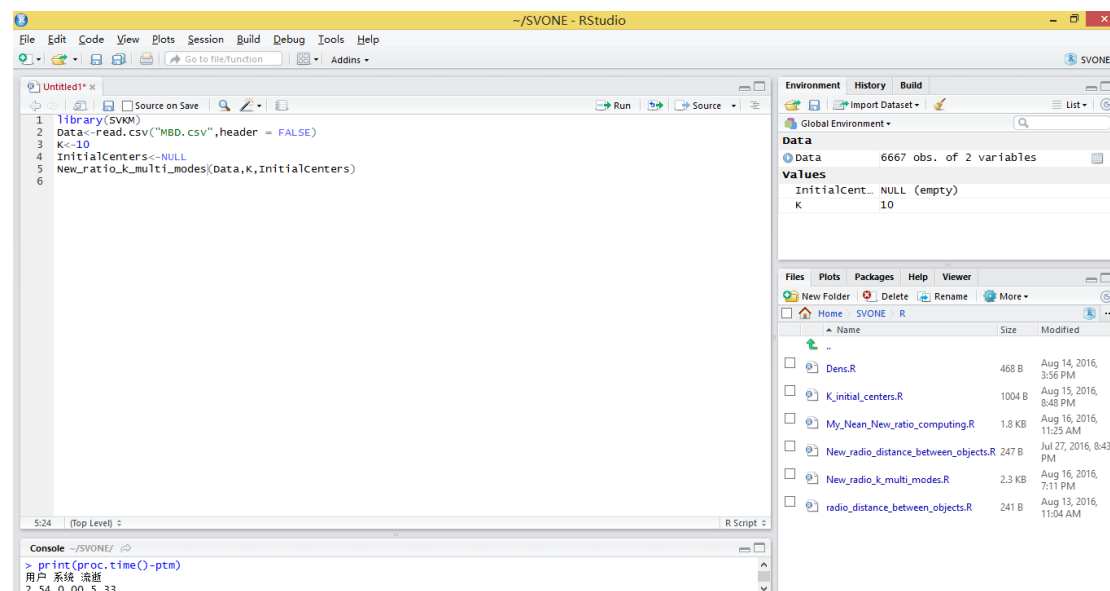


Figure 12 Example of the SV-K-modes -M function

3.2.3 SV-K-modes-D

Description

Implement the SV-k-modes algorithm for distribute computation.

Usage

New-ratio-k-multi-modes-D (Data, K, InitialCenters)

Return (nj)

Arguments

Inputs	Data	Data set to be clustered
	K	Number of clusters
	InitialCenters	Initial cluster centers
Outputs	nj	Clustering result

Details

Please refer to the details mentioned above.

Reference

F. Cao, Z. Huang, J. Liang, X. Zhao and Y. Meng, SV-k-modes: an algorithm for clustering categorical data with set-valued attributes, Neural Networks and Learning System, IEEE Transactions on (Under review).

Example

```
>library ( SVKD )
>Data=read.csv ( " MBD.csv " , header=FALSE )
>K=10
>InitialCenters=null
>New-ratio-k-multi-modes ( Data, K, InitialCenters )
```

Results

```
[1,] 1 5
[2,] 1 3
[3,] 1 8
[4,] 1 2
[5,] 1 7
[6,] 1 1
[7,] 1 9
[8,] 2 5
[9,] 2 3
[10,] 2 8
[11,] 2 2
[12,] 2 7
[13,] 2 1
[14,] 2 9
[15,] 2 9
[16,] 3 5
[17,] 3 3
[18,] 3 8
[19,] 3 2
[20,] 3 7
[21,] 3 1
[22,] 3 9
[23,] 3 9
```

Figure 13 Example of the SV-K-modes -D function

3.2.4 F-SV-K-modes-S

Description

Implement the fuzzy SV-k-modes algorithm for single-threaded.

Usage

Fuzzy-ratio-k-multi-modes (Data, K, InitialCenters, a)

return (iter, cid, time)

Arguments

Inputs	data	Data set to be clustered
	K	Number of clusters
	InitialCenters	Initial cluster centers
	a	Index
Outputs	cid	Clustering result
	time	Computational cost
	iter	Iterations

Details

Fuzzy SV-k-modes algorithm uses Jaccard coefficient to compute the distance between two objects and set-valued modes as the center of a cluster. A kind of update way of clustering prototype is developed for the fuzzy partition matrix. These extensions made the fuzzy SV-k-modes algorithm can cluster categorical data with mixed single-valued and set-valued attributes.

Reference

F. Cao, J. Huang, and J. Liang, A fuzzy SV-k-modes algorithm for clustering categorical data with set-valued attributes, Applied Mathematics and Computation 295 (15) (2017) 1-15.

Example

```
>library ( FSVKS )
>Data=read.csv ( " MBD.csv ", header=FALSE )
>K=10
>InitialCenters=NULL
>a=1.1
>Fuzzy-ratio-k-multi-modes ( Data, K, InitialCenters, a )
```

Results

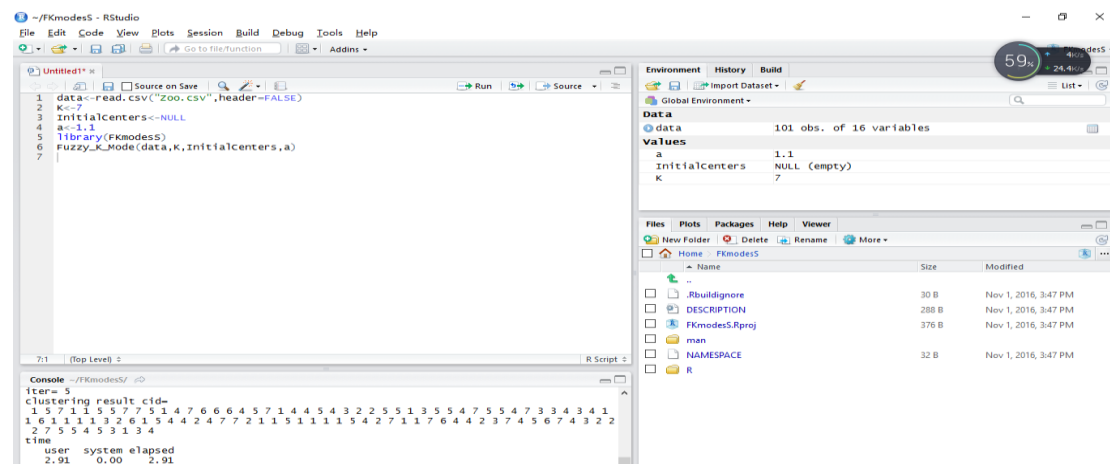


Figure 14 Example of the Fuzzy-SV-K-modes -S function

3.2.5 F-SV-K-modes-M

Description

Implement the fuzzy SV-k-modes algorithm for multi-threaded.

Usage

Fuzzy-ratio-k-multi-modes-M (Data, K, InitialCenters, a)

Return (time)

Arguments

Inputs	data	Data set to be clustered
	K	Number of clusters
	InitialCenters	Initial cluster centers
	a	Index
Outputs	cid	Clustering result
	time	Computational cost
	iter	Iterations

Details

Please refer to the description mentioned above.

Reference

F. Cao, J. Huang, and J. Liang, A fuzzy SV-k-modes algorithm for clustering categorical data with set-valued attributes, Applied Mathematics and Computation 295 (15) (2017) 1-15.

Example

```
>library ( FSVKM )  
>Data=read.csv ( " MBD.csv " ,header=FALSE )  
>K=10  
>InitialCenters=NULL  
>a=1.1  
Fuzzy-ratio-k-multi-modes-M ( Data, K, InitialCenters,a )
```

Results

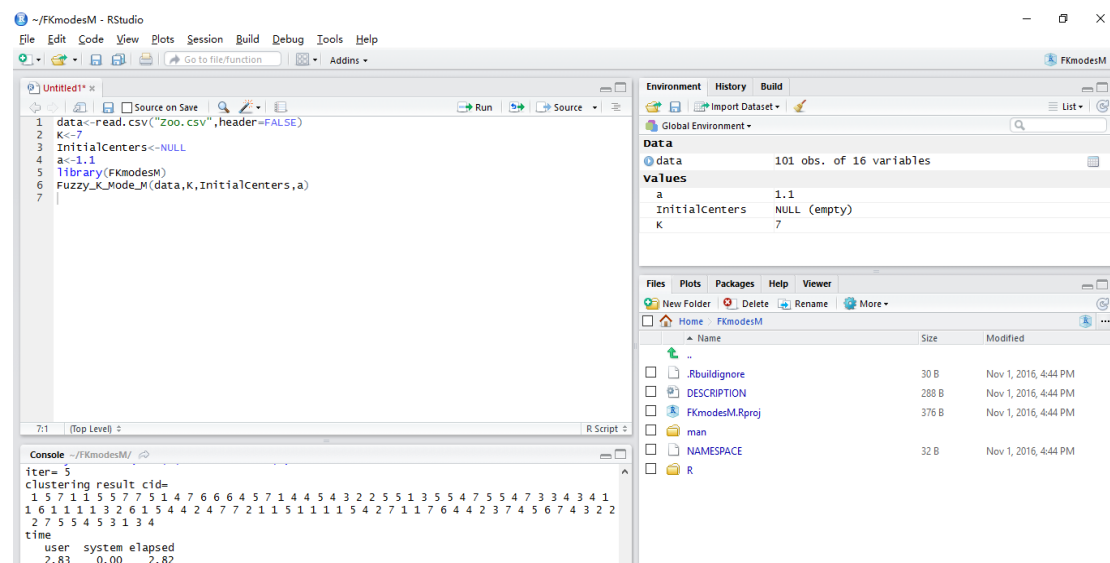


Figure 15 Example of the Fuzzy-SV-K-modes-M function

3.2.6 F-SV-K-modes-D

Description

Implement the fuzzy SV-k-modes algorithm for distribute computation.

Usage

Fuzzy-ratio-k-multi-modes-D (Data, K, InitialCenters)

Return (nj)

Arguments

Inputs	Data	Data set to be clustered
	K	Number of clusters
	InitialCenters	Initial cluster centers
	a	index
Outputs	nj	Clustering result

Details

Please refer to the description mentioned above.

Reference

F. Cao, J. Huang, and J. Liang, A fuzzy SV-k-modes algorithm for clustering categorical data with set-valued attributes, Applied Mathematics and Computation 295 (15) (2017) 1-15.

Example

```
>library ( FSVKD )
>Data=read.csv ( " MBD.csv " , header=FALSE )
>K=10
```

```
>InitialCenters=NULL
>a==1.1
>Fuzzy-ratio-k-multi-modes-D ( Data, K, InitialCenters, a)
```

Results

```
[1,] 1 1
[2,] 1 2
[3,] 1 3
[4,] 1 4
[5,] 1 5
[6,] 1 6
[7,] 2 7
[8,] 2 1
[9,] 2 8
[10,] 2 9
[11,] 2 3
[12,] 2 2
[13,] 3 10
[14,] 3 8
[15,] 3 11
[16,] 3 9
[17,] 3 4
[18,] 3 5
```

Figure 16 Example of the Fuzzy-SV-K-modes -D function

3.3 Initial class center selection function

3.3.1 K-modes-k-initial-center

Description

The initialization clustering algorithm aims at selecting the initial cluster centers for the k-modes algorithm.

Usage

```
k-initial-center<-function(data,K)
return(DDist)
```

Arguments

Input	data	A vector in data matrix
	K	Number of clusters
Output	DDist	Initial cluster centers

Details

In the initialization clustering algorithm, the first cluster center for categorical data is given by selecting the object with the maximal density. For the other initial centers, we need to consider the density of an object and the distance between the object and the selected initial cluster centers simultaneously.

Reference

F. Cao, J. Liang, L. Bai, A new initialization method for categorical data clustering, Expert Systems with Applications 36 (7) (2009) 10223-10228.

Z. Huang, Extensions to the k-means algorithm for clustering large data sets with categorical values, Data Mining and Knowledge Discovery 2 (3) (1998) 283-304.

Example

```
>Data<-read.csv("Zoo.csv",header=FALSE)
>K<-7
```

Result

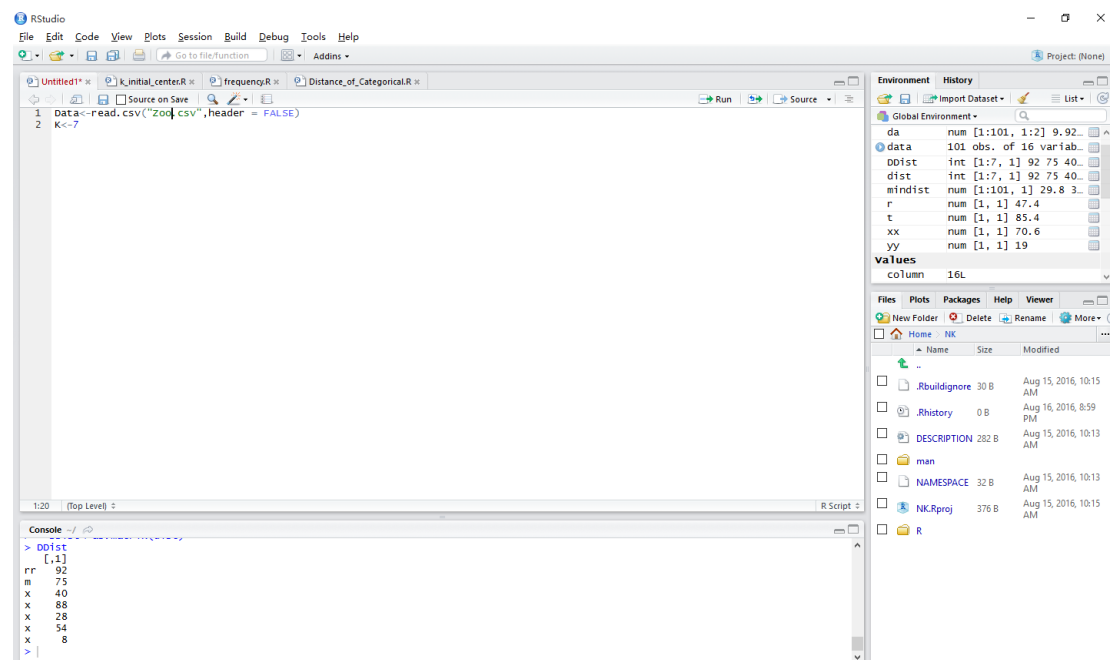


Figure 17 Example of the K-modes-k-initial-center function

3.3.2 SV-K-modes-k-initial-center

Description

The initialization algorithm aims to select the initial cluster centers for the SV-k-modes algorithm.

Usage

```
K-initial-centers<-function(Data,K)
Return(dist)
```

Arguments

Input	Data	A vector in data matrix
	K	Number of clusters
Output	dist	Initial cluster centers

Details

The initialization algorithm is to select the initial cluster centers for the SV-k-modes algorithm by extending the initialization method for the k-modes algorithm. When

more initial centers are selected, we need to consider the density of an object and the distance between the object and the selected initial cluster centers simultaneously.

Reference

F. Cao, Z. Huang, J. Liang, X. Zhao and Y. Meng, SV-k-modes: an algorithm for clustering categorical data with set-valued attributes, Neural Networks and Learning System, IEEE Transactions on (Under review).

Example

```
>Data=read.csv("MBD2.csv", header=FALSE)
>K=3
```

Result

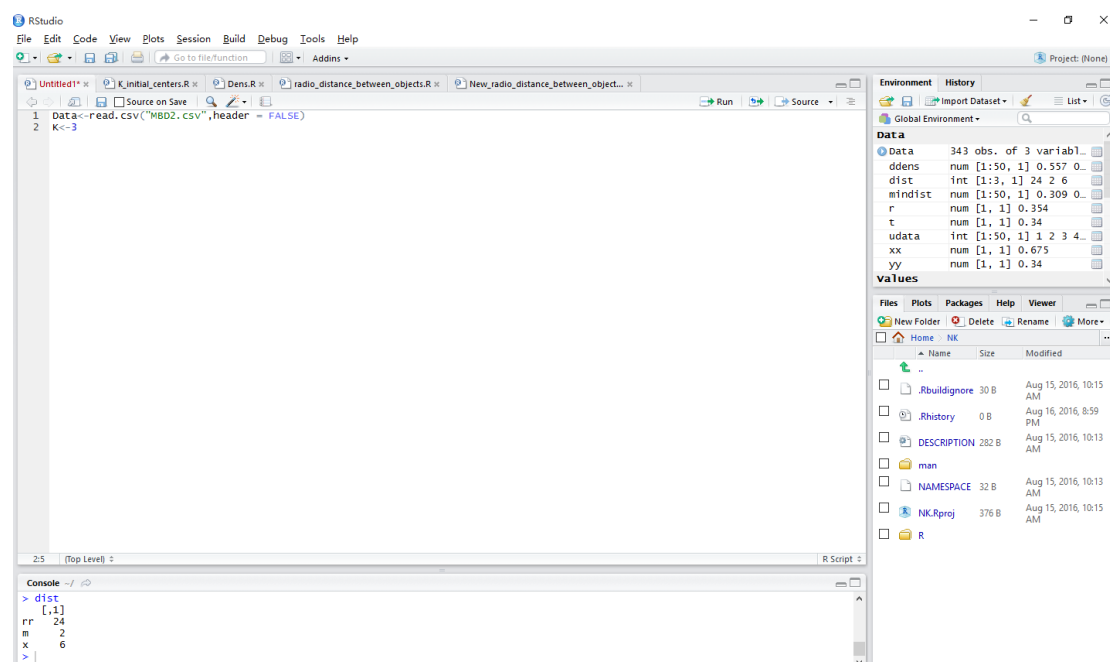


Figure 18 Example of the K-modes-k-initial-center function

3.4 Sub function

3.4.1 K-dis

Description

Calculate the distance between each object and the cluster center.

Usage

Distance-of-Categorical (Data1, Data2)

Return (distance)

Arguments

Inputs	Data1	A vector in data matrix
	Data2	A vector in data matrix
Output	distance	The distance between Data1 and Data2

Details

A simple 0 - 1 matching method is used to calculate the distance between the two different objects.

Example

```
>Data=read.csv ( " Zoo.csv " , header=FALSE )
>Distance-of-Categorical ( Data [ 1:10 , ] , Data2 [10:20 , ] )
```

Results

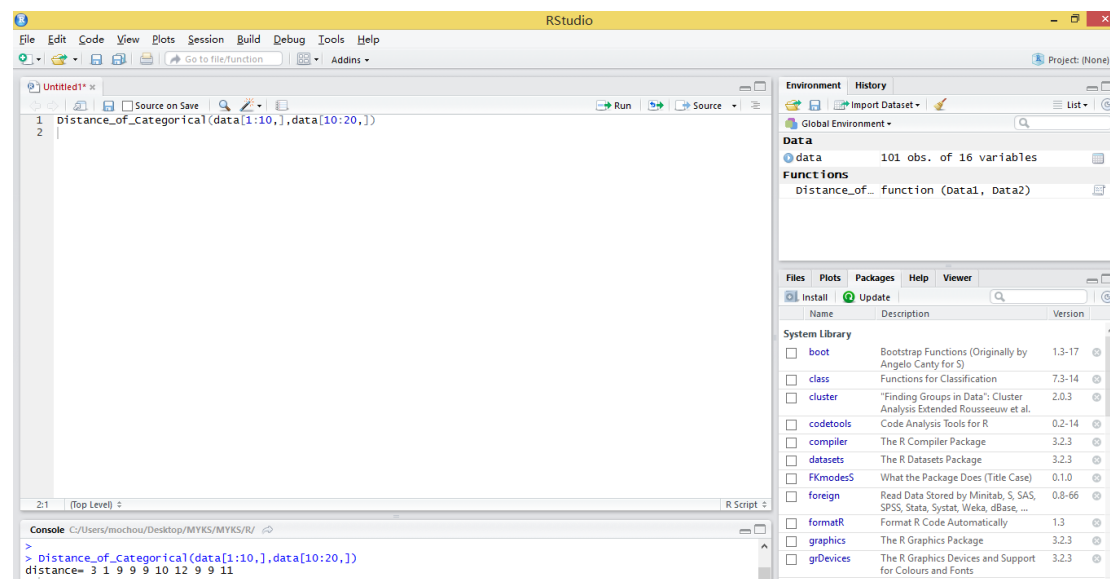


Figure 19 Example of the K-dis function

3.4.2 K-center

Description

Find the cluster center for each cluster.

Usage

```
Find-Mode ( Data )
Return ( NNew-Mode )
```

Arguments

Input	Data	A vector in data matrix
Output	NNew-Mode	The cluster centers of Data

Details

The center of each cluster is obtained by choosing the value whose frequency is the

highest on each attribute.

Example

```
>Data=read.csv ( " Zoo.csv ", header=FALSE )
>Find-Mode ( Data [ 1:30, ] )
```

Results

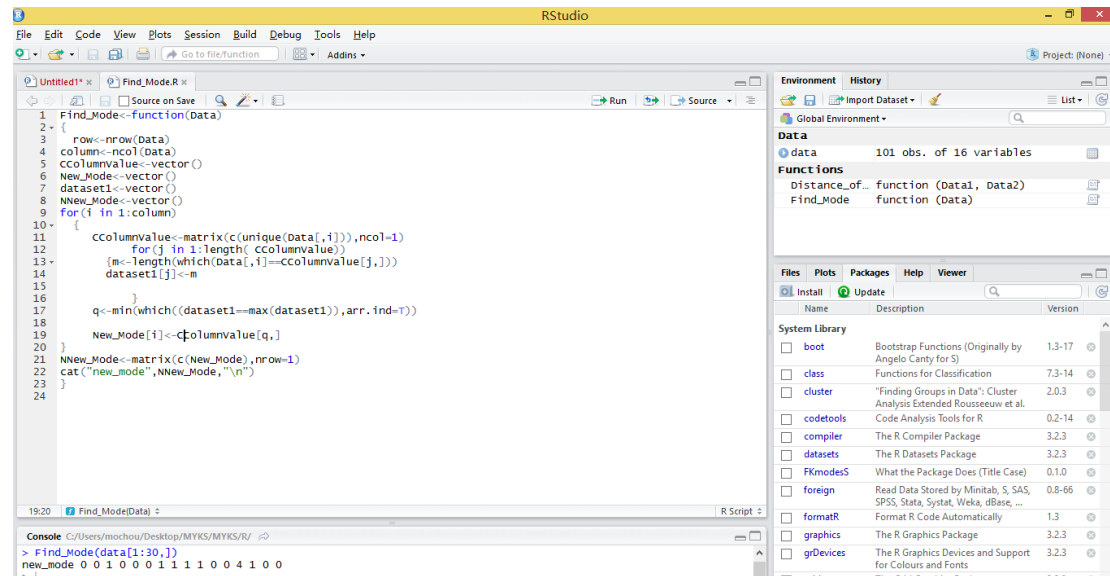


Figure 20 Example of the K-center function

3.4.3 SV-K-dis

Description

The distance function of SV-k-modes algorithm.

Usage

```
New-ratio-distance-between-objects ( DData1, DData2 )
Return ( distance )
```

Arguments

Inputs	DData1	A vector in data matrix
	DData2	A vector in data matrix
Output	distance	The distance between Data1 and Data2

Details

Calculate the distance between each object and the center of each class.

Reference

F. Cao, Z. Huang, J. Liang, X. Zhao and Y. Meng, SV-k-modes: an algorithm for clustering categorical data with set-valued attributes, Neural Networks and Learning System, IEEE Transactions on (Under review).

Example

```
>DData1=c(1,3,5,6,8,9,2)
>DData2=c(2,4,5,7,8,1,9)
>New-ratio-distance-between-objects(DData1,DData2)
```

Results

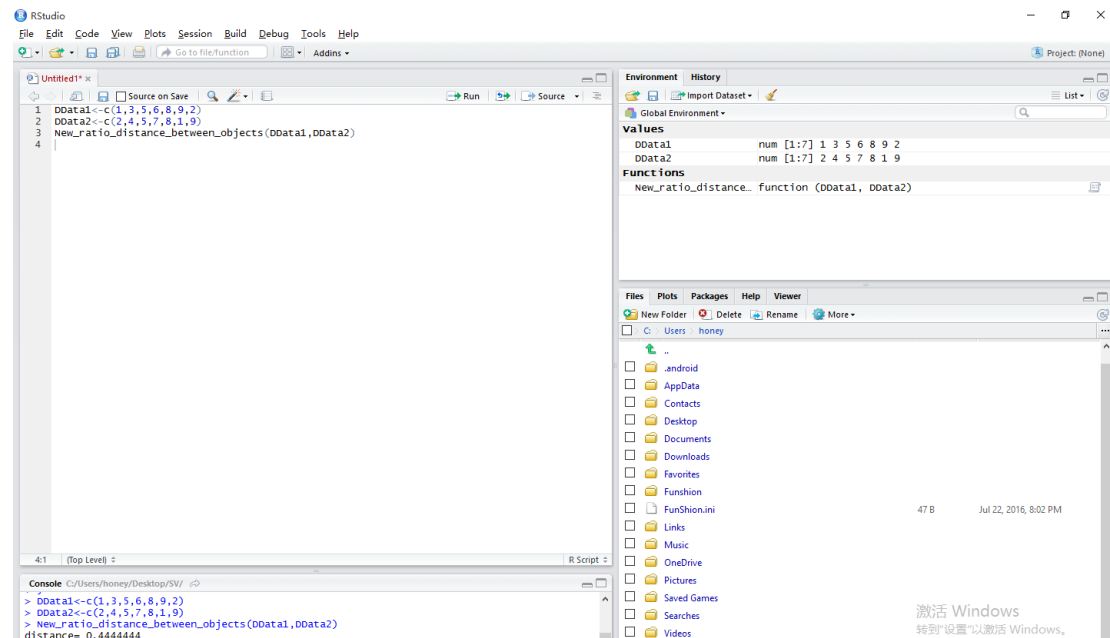


Figure 21 Example of the SV-K-dis function

3.4.4 SV-K-center

Description

The find mode function of SV-k-modes algorithm.

Usage

```
My-Mean-New-ratio-computing-modes(data)
return(Mode)
```

Arguments

Input	data	A vector in data matrix
Output	Mode	The cluster centers of data

Reference

F. Cao, Z. Huang, J. Liang, X. Zhao and Y. Meng, SV-k-modes: an algorithm for clustering categorical data with set-valued attributes, Neural Networks and Learning System, IEEE Transactions on (Under review).

Example

```
>Data=read.csv("MBD1.csv",header=FALSE)
```

>My-Nean-New-ratio-computing-modes(data)

Results

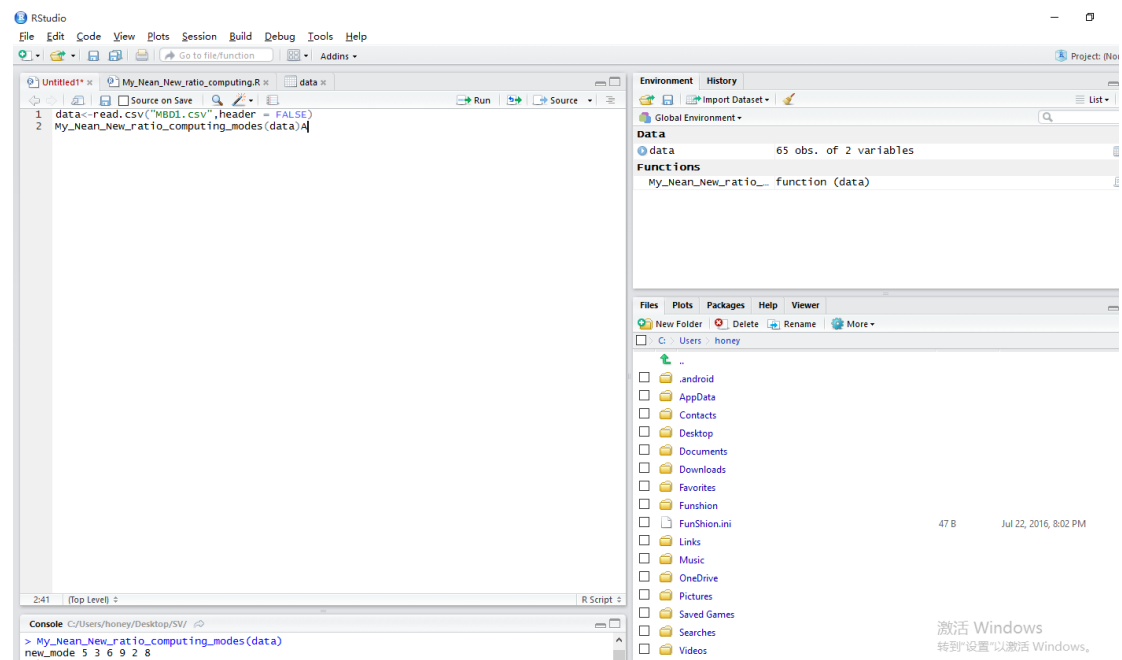


Figure 22 Example of the SV-K-center function