JCSE International Journal of Computer Sciences and Engineering Open Access

Research Paper Vol.-7, Issue-5, May 2019 E-ISSN: 2347-2693

Kmeans Clustering in R Studio

Shivani Chauhan [1*], Bharti Nagpal [2]

[1] Department of Computer Science, AIACTR, Delhi, India
[2] Department of Science and Technology, AIACTR, , Delhi, India

DOI: https://doi.org/10.26438/ijcse/v7i5.13601362 | Available online at: www.ijcseonline.org

Accepted: 22/May/2019, Published: 31/May/2019

Abstract— This paper describes about data mining and its techniques , the main focus for this paper is the clustering techniques. There are many different tools available for performing the clustering techniques , for this paper the clustering technique which is chosen is the Kmeans clustering and the tool which is used to perform this clustering technique is R studio. R studio is the tool which is used for both data mining as well as for the visual analytics also . The Kmeans clustering is the clustering technique which is used to cluster the data in which data items are categorised into the k groups of similarity. This tool not only provide interface to use data mining but it also give us a visual representation of the result generated by that data mining algorithm . The data set which has been used for performing Kmeans clustering consist of nine attributes and 583 observations.

Keywords—**:** Data Mining. Clustering , classification ,clustering , Kmeans , R studio .

I. Introduction

Data is an integral part of decision making or taking the new approaches for increasing the growth of an organisation . All the organisations mainly depends on mining of data so as to find the new approaches to increase there growth in the market.

Data mining is the process of searching the new patterns and rules in the data set . Its is also known as knowledge mining.

The steps that are involved in data mining are:

- Extraction , transform and load the data into the data warehouse.
- Storing and manage the data in multidimensional database system .
- Provide data access to the business analysts and information professionals .
- Analysis of data by application software.
- Presentation of the data in useful format such as pie chart, graphs and tables .

The functionalities of data mining tasks are classified into two types:

- Descriptive mining tasks :This task is used to characterize the general properties of data in the database.
- Predictive mining tasks: This task is mainly used on the current data in order to make predictions.

2. Data Mining Techniques

2.1 Classification

Classification is the type of data mining technique in which the data are classified according to there classes .This technique is used to retrieve important and relevant information about data , and meta data .

2.2 Clustering

Clustering is a technique of data mining in which the similar kind of data are group together. And that group is called the cluster.

The data grouping is based on some similarities between the data.

2.3 Regression

Regression is the data mining technique in which the relationship are identified and analysed between the variables. This techniques is basically used to identify the likelihood of a specific variable given in presence of another variable.

2.4 Association Rules

Association rule is helps to find the association between two or more items . This technique is basically used to find hidden patterns in the dataset.

3.Methodology Used

Kmeans Clustering

The Kmeans clustering is the clustering technique which is used to cluster the data in which data items are categorised into the k groups of similarity. Thus assign each data points into k groups based upon some similarities between them. The result of the kmeans algorithms are :

Centers or centroids which used to label these data.

- Each data points are assign to clusters which is known as labels

Here the averaging of data takes place which we called centroids .

4. Tool used

For the kmeans clustering i am using using the tool called Rstudio . R studio is the tool which is used for both data mining as well as for the visual analytics also . This tool not only provide interface to use data mining but it also give us a visual representation of the result generated by that data mining algorithm . It is a free open source tool which provide an integrated environment for R programming language . This tool is available for Windows, Mac and Linux. It is partially written in C++, java and javascript , it use Qt framework for its graphical user interface .

5.Dataset

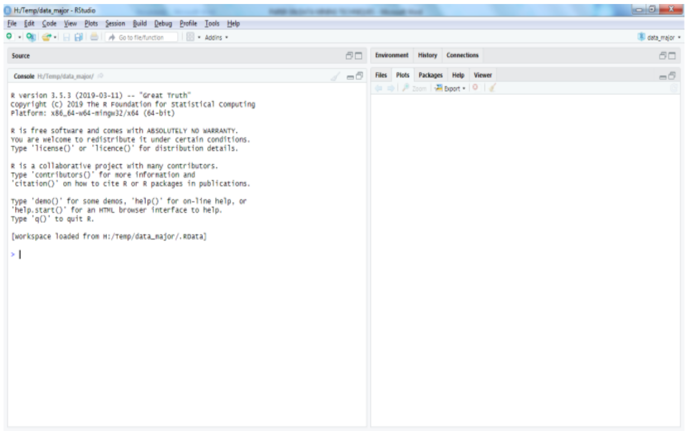
The data set which will be used in this paper is the Indian liver Patient dataset(ILPD) [4]. The data set contains 583 observations and 9 variable .

Table1. Attributes Description of ILPD

S.NO	ATTRIBUTE	TYPE
1.	AGE	NUMERICAL
2.	GENDER	NOMINAL
3.	TOTAL BILIRUNBIN	NUMERICAL
4.	DIRECT BILIRUNBIN	NUMERICAL
5.	ALKALINE PHOSPHOTASE	NUMERICAL
6.	ALAMINE AMINOTRANSFERASE	NUMERICAL
7.	ASPARTATE AMINOTRANSFERASE	NUMERICAL
8.	ALBUMIN AND GLOBULIN RATIO	NUMERICAL
9.	PROBLEM	NOMINAL

6. Working of kmeans in R studio

The working of k means clustering is very simple and easy as r studio is a simple tool which has many libraries in it .


H:/Temp/data_major - RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
data_major
Source
Console H:/Temp/data_major/
Environment History Connections
Files Plots Packages Help Viewer
Zoom
Export

```
R version 3.5.3 (2019-03-11) -- "Great Truth"
Copyright (C) 2019 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[workspace loaded from H:/Temp/data_major/.RData]

> |
```

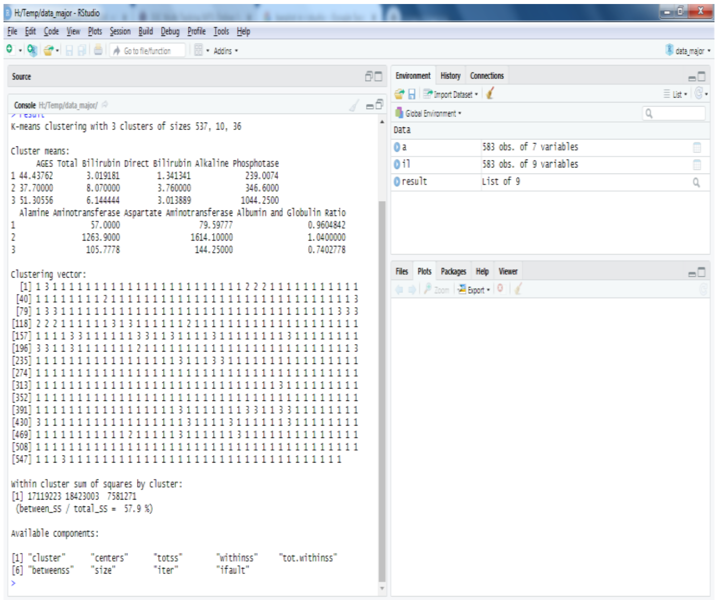
Fig1. This figure show the user interface of r studio

Step 1. Load the data in R studio through the import data set tab . The dataset which is to be imported should be present in excel file or in the text file .

Step2 . Create the copy of the data set and store it into another variable .

Step3 . Remove all the nominal categories present the copy of the data .

Step4 . Use the built in function kmeans to do the kmeans clustering to the data . And stores it into another variable.


H:/Temp/data_major - RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
data_major
Source
Console H:/Temp/data_major/
Environment History Connections
Import Dataset
List
Global Environment
Files Plots Packages Help Viewer
Zoom
Export

```
> result
K-means clustering with 3 clusters of sizes 537, 10, 36

Cluster means:
      AGES Total Bilirubin Direct Bilirubin Alkaline Phosphotase
1 44.43762        3.019181         1.341341             239.0074
2 37.70000        8.070000         3.760000             346.6000
3 51.30556        6.144444         3.013889            1044.2500
  Alamine Aminotransferase Aspartate Aminotransferase Albumin and Globulin Ratio
1                  57.0000                   79.59777                  0.9604842
2                1263.9000                 1614.10000                  1.0400000
3                 105.7778                  144.25000                  0.7402778

Clustering vector:
  [1] 1 3 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 1 1 1 1 1 1 1 1 1 1 1
 [40] 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 3
 [79] 1 3 3 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 3 3 3
[118] 2 2 2 1 1 1 1 1 1 3 1 3 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
[157] 1 1 1 1 3 3 1 1 1 1 1 1 1 3 3 1 1 3 1 1 1 1 3 1 1 1 1 1 1 1 1 1 3 1 1 1 1 1 1 1 1
[196] 3 3 1 1 3 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 3
[235] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 3 1 1 1 3 3 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
[274] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
[313] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 3 1 1 1 1 1 1 1 1 1
[352] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
[391] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 3 1 1 1 1 1 1 1 1 3 3 1 1 3 3 1 1 1 1 1 1 1
[430] 3 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 3 1 1 1 1 3 1 1 1 1 1 1 1 3 1 1 1 1 1 1 1
[469] 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 3 1 1 1 1 1 1 1 3 1 1 1 1 1 1 1 1 1 1 1 1 1
[508] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
[547] 1 1 1 3 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1

Within cluster sum of squares by cluster:
[1] 17119223 18423003  7581271
 (between_SS / total_SS =  57.9 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
[6] "betweenss"    "size"         "iter"         "ifault"
>
```

```
Data
a        583 obs. of 7 variables
il       583 obs. of 9 variables
result   List of 9
```

Fig 2 . Show the result of k means clustering

Step5 . Clusters the problems of the ILPD according to the three clusters . The problems are clustered into the three clusters as follows :

	1	2	3
A	211	2	7
B	165	2	20
C	161	6	9

Step6.Now show the barplot of problems present in clusters .

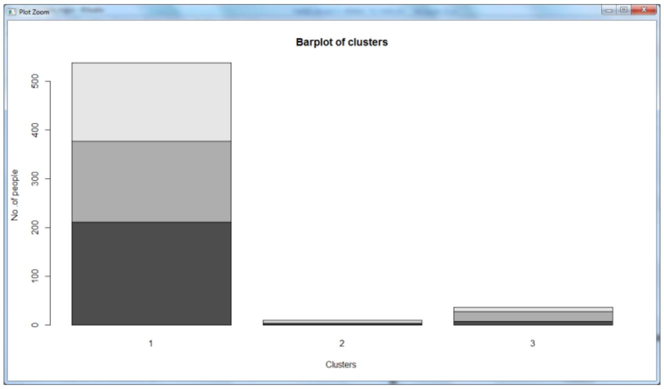


Fig 3.Shows the barplot of problems present in different clusters .

Step7. Now show the male and female presnt in the three clusters .

```
          1      2    3
 Female 130     0   12
 Male   407    10   24
```

Step6.Now show the barplot of gender present in clusters.

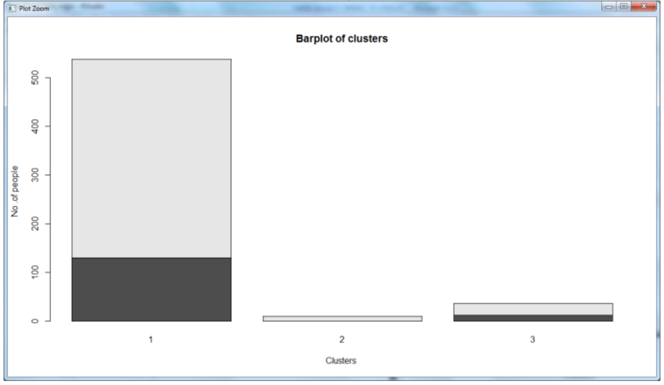


Fig 4.Shows the barplot of gender present in different clusters .

7. Conclusion

As the data is increasing day by day it will become impossible for us to analysis the whole lot of data , hence we require the tools which can mine the data and give us the visual representation of the data . Like the kmeans clustering we can also perform the classification technique and visual it results also .

References

[1]. Nikita Jain, Vishal Srivastava,” DATA MINING TECHNIQUES: A SURVEY PAPER” eISSN: 2319-1163 | pISSN: 2321-7308

[2]. Wahbeh, A.H., Al-Radaideh Q.A., Al-Kabi, M.N. and AlShawakfa E.M. 2010. A comparison study between Data Mining Tools over some classification methods. IJACSA, Special Issue on Artificial Intelligence, SAI Publisher, 2(8), pp. 18-26.

[3]. Auza, J. 2010. 5 of the Best Free and Open Source Data Mining Software. [Accessed Online March 2013] http://www.junauza.com/2010/11/free-data-miningsoftware.html.

[4]. Indian Liver Patient dataset https://archive.ics.uci.edu/ml/datasets/ILPD+(Indian+Liver+Patient+Dataset)

Authors Profile

Ms. Shivani Chauhan has Bachelors of Technology from Maharshi Dayanand University , India . She is currently pursing Masters in Technology specailized in Information Security from Guru Gobind Singh Indraprastha University, India .

Dr.Bharti Nagpal is Assistant Professor in AIACTR college, Delhi.She has 18 years of teaching experience . Her area of specialisation are:Information Security ,Data Mining,Big Data and Web Enginnering .