Global Journal of Computer Science and Technology: C
Software & Data Engineering
Volume 16 Issue 1 Version 1.0 Year 2016
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Cluster Analysis of Medical Research Data using R

By Lavanya Pamulaparty, Dr. C.V Guru Rao & Dr. M. Sreenivasa Rao

JNT University, India

Abstract- Cluster analysis divides the data into groups that are meaningful, useful or both. It is also used as a starting point for other purposes of data summarization. This paper discuss some very basic algorithms like K-means, Fuzzy C-means, Hierarchical clustering to come up with clusters, and use R data mining tool. The results are tested on the datasets namely Online News Popularity, Iris Data Set and from UCI data repository and mi RNA dataset for medical data analysis. All datasets was analyzed with different clustering algorithms and the figures we are showing is the working of them in R data mining tool. Every algorithm has its uniqueness and antithetical behavior.

Keywords: *k-means algorithm, fuzzy c-means algorithm, hierarchical clustering algorithm, r tool.*

GJCST-C Classification : *H.3.3 I.5.3*

Cluster Analysis of Medical Research Data using R

Lavanya Pamulaparty [α], Dr. C.V Guru Rao [σ] & Dr. M. Sreenivasa Rao [ρ]

Abstract- Cluster analysis divides the data into groups that are meaningful, useful or both. It is also used as a starting point for other purposes of data summarization. This paper discuss some very basic algorithms like K-means, Fuzzy C-means, Hierarchical clustering to come up with clusters, and use R data mining tool. The results are tested on the datasets namely Online News Popularity, Iris Data Set and from UCI data repository and mi RNA dataset for medical data analysis. All datasets was analyzed with different clustering algorithms and the figures we are showing is the working of them in R data mining tool. Every algorithm has its uniqueness and antithetical behavior.

Keywords: k-means algorithm, fuzzy c-means algorithm, hierarchical clustering algorithm, R tool.

I. Introduction

Cluster analysis divides data into meaning full groups (clusters) which share common characteristics i.e. same cluster are similar to each other than those in other clusters. It is the study of automatically finding classes. A web page especially news articles which are flooded in the internet have to be grouped. The clustering of these different groups is a step forward towards the automation process, which requires many fields, including web search engines, web robots and data analysis.

Any new web page goes through numerous phases including data acquisition, preprocessing, Feature extraction, classification and post processing into the database. Cluster analysis can be regarded as a form of the classification which creates a labeling of objects with class labels. However it derives these labels only from the data. Data mining functionalities are the Characterization and discrimination, mining frequent patterns, association, correlation, classification and prediction, cluster analysis, outlier analysis and evolution analysis [1].

Clustering is a vivid method. The solution is not exclusive and it firmly depends upon the analysts' choices. Clustering always provides groups or clusters, even if there is no predefined structure. While applying cluster analysis we are contemplating that the groups exist. But this speculation may be false. The outcome of clustering should never be generalized. [9].

Author α : Department of CSE, Methodist college of Engg. & Tech., OU, Hyderabad. e-mail: lavanya.post@gmail.com

Author σ : Department of CSE, S R Engineering College, JNT University, Warangal.

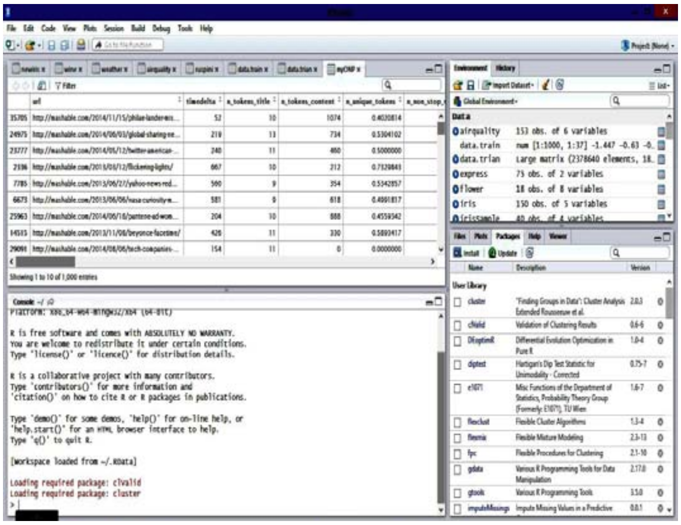
Author ρ : Department of CSE, School of IT, JNT University, Hyderabad.

II. R Tool

R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and Mac OS [12]

R is public domain software primarily used for statistical analysis and graphic techniques [10]. A core set of packages is included with the installation of R, with more than 7,801 additional packages (as of January 2016[update]) available at the Comprehensive R Archive Network (CRAN), Bio conductor, Omegahat, Git Hub, and other repositories.[14] R tool provides a wide class of statistical that includes classical statistical tests, linear and nonlinear modeling, classification, time-series analysis, clustering and various graphical functions.[13]

R uses collections of packages to perform different functions [11]. CRAN project views provide numerous packages to different users according to their taste. R package contain different functions for data mining approaches. This paper compares various clustering algorithms on datasets using R which will be useful for researchers working on medical data and biological data as well. For this IDE, R Studio is used refer the below Figure 1.



	url	timedelta	n_tokens_title	n_tokens_content	n_unique_tokens
35705	http://mashable.com/2014/11/15/philae-lander-mis...	52	10	1074	0.4020814
24975	http://mashable.com/2014/06/03/global-sharing-ne...	219	13	734	0.5304102
23777	http://mashable.com/2014/05/12/twitter-american-...	240	11	460	0.5000000
2106	http://mashable.com/2013/03/12/flickering-lights/	667	10	212	0.7320843
7785	http://mashable.com/2013/06/27/yahoo-news-red...	560	9	354	0.5342857
6673	http://mashable.com/2013/06/06/nasa-curiosity-m...	581	9	618	0.4901817
25963	http://mashable.com/2014/06/18/pantene-ad-wom...	204	10	888	0.4559342
14515	http://mashable.com/2013/11/08/beyonce-facetime/	426	11	330	0.5893417
29091	http://mashable.com/2014/08/06/tech-companies-...	154	11	0	0.0000000

Showing 1 to 10 of 1,000 entries

```
Console ~/
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 ~/.RData]

Loading required package: clValid
Loading required package: cluster
>
```

```
Data
airquality    153 obs. of 6 variables
data.train    num [1:1000, 1:37] -1.447 -0.63 -0...
data.trian    Large matrix (2378640 elements, 18...
express       75 obs. of 2 variables
flower        18 obs. of 8 variables
iris          150 obs. of 5 variables
irisSample    40 obs. of 4 variables
```

	Name	Description	Version
	User Library		
□	cluster	"Finding Groups in Data": Cluster Analysis Extended Rousseeuw et al.	2.0.3
□	clValid	Validation of Clustering Results	0.6-6
□	DEoptimR	Differential Evolution Optimization in Pure R	1.0-4
□	diptest	Hartigan's Dip Test Statistic for Unimodality - Corrected	0.75-7
□	e1071	Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien	1.6-7
□	flexclust	Flexible Cluster Algorithms	1.3-4
□	flexmix	Flexible Mixture Modeling	2.3-13
□	fpc	Flexible Procedures for Clustering	2.1-10
□	gdata	Various R Programming Tools for Data Manipulation	2.17.0
□	gtools	Various R Programming Tools	3.5.0
□	imputeMissings	Impute Missing Values in a Predictive	0.0.1

Fig 1: R tool Studio

III. Clustering Algorithms

a) K-Means

The term "k-means" was first used by James Macqueen in 1967 [2], though the idea goes back to

1957 [4]. The K-means algorithm is the most commonly used and simplest method among all partitional clustering algorithms. As Harington and Wong1979 mentioned it is an iterative method which minimizes the sum of the squares for a given number of Clusters.

Here's how the algorithm works [5]:

1. Select k points as initial centroids.
2. Repeat
3. From K clusters by assigning each point to its closest centroid.
4. Recompute the centroid of each cluster.
5. Until centroids do not change.

K-Means reaches a state in which no points are shifting from one cluster to another e.g. repeating until only 1% of the points change clusters.

For measuring the quality of the clustering we measure Sum of the squared error (SSE) or scatter.

$$\text{SSE}=\sum_{i=1}^{k}\sum_{x\in c_i} dist(c_i\,,\,\text{x})^2$$

Where dist is standard Euclidean distance between two objects in Euclidean space. The centroid (mean) of the ith cluster that minimizes the SSE is defined as

$$\sum_{i=1}^{k}\sum_{x\in c_i} dist(c_i\,,\,\text{x})^2$$

The advantage of this method is highly scalable of the huge sum of data sets with O(n * k * r) where r is the number of rounds, where n represent number of data items, k represent numbers of clusters [14]. It has user defined constant K and Runtime is totally dependent on the initial pick of centroids.

b) K-Means Implementation using R

For this analysis we have considered Online News Popularity datasets which consists of articles published by Mash able (www.mashable.com) [4]. Instances are 39797 and Number of Attributes is 61. As the results of the k means are undeterministic, we have followed the practice of running multiple rounds of k means so performed on various k values as k=3, k=5 and k=10. The best iteration is one who minimizes the average distance of each point to its assigned centroid.

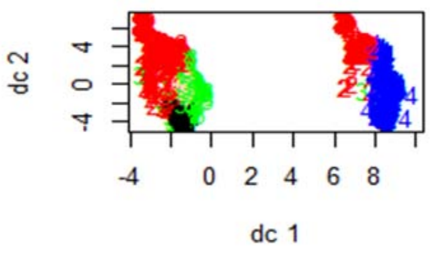


Fig 2 : K means plot with k=5

The Figure 2 shows the results of clustering the Online News Popularity datasets with 39645 number of News articles URL's. The above results show that there is overlapping of clusters. So preprocessing is required to address this problem and omit the NA values also. Then the following is the code after data cleaning [15].

```
>onpdat<- kmeans (myONP [, c
("n_tokens_title","n_tokens_content")], centers=5,
nstart=10)
> Onpdata
        K-means clustering with 5 clusters of sizes 15,
67, 195, 319, and 404
Cluster means:
  n_tokens_title n_tokens_content
1      11.46667       2648.9333
2      10.29851       1469.2985
3      10.41026        897.5897
4      10.50470        492.4044
5      10.28713        211.3886
Clustering vector:
35705  24975  23777  2336  7785  6673  25963  14515
29091  18807  27116  37480  14360  29375  35316   8015
24621  4744 10096 14587
  3    3    4    5    4    4    3    5    5    5    5    4    5
3    5    4    4    5    4    3
  506  14445  32852  12857  18210  22647  18642   7034
31249  25246  29996   4077  27331  15531  31001  24434
29564 20883 20002 29804
  3    4    4    4    5    3    4    5    5    5    3    5    4
3    5    5    5    5    2    5
  881  18018  27648  26153  18032  32512  16539   9241
2668    3755  11938  19576  24987  15355  34454  11081
17326 12545 24563  9737
  4    2    4    3    5    3    3    5    5    2    5    5    5
4    5    5    4    5    3    1
Within cluster sum of squares by cluster:
[1] 3437391 3417672 3385646 3069653 3279165
 (between_SS / total_SS =  92.1 %)
Available components
[1] "cluster"      "centers"      "totss"        "withinss"
"tot.withinss" "betweenss"   "size"
[8] "iter"       "ifault"
> summary(onpdata)
 Length    Class    Mode
cluster      1000    -none- numeric
centers        10    -none- numeric
totss           1     -none- numeric
withinss        5    -none- numeric
tot.withinss     1   -none- numeric
betweenss        1   -none- numeric
size            5      -none- numeric
iter            1      -none- numeric
ifault          1      -none- numeric
library(MASS)
> parcoord(data.train,onpdata$cluster)
>confuseTable.km<-
table(myONP$n_tokens_title,onpdata$cluster)
> confuseTable.km
      1  2  3  4  5
  3   0  0  0  0  1
  5   0  0  0  1  2
  6   0  1  1  8 13
  7   1  8 10  7 26
  8   1  4 29 34 44
  9   2 12 25 48 53
  10  2 12 39 67 81
  11  1 10 35 55 77
  12  2  7 22 53 50
  13  4  9 24 24 19
  14  1  3  7 12 25
  15  0  1  2  6  9
  16  0  0  0  4  3
  17  0  0  0  0  1
  18  1  0  1  0  0
> library(flexclust)
> randIndex(confuseTable.km)
        ARI
0.002285344
```

The results are showing the information about cluster means, clustering vector, sum of square by cluster and available components. The fpc package is used to draw the discriminant projection plot using Plotcluster function (Fig3).

The result of plotting the class returned by function application is shown below.

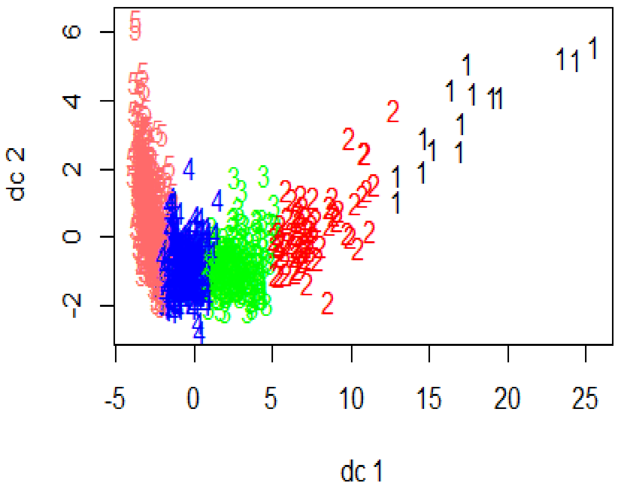


Fig. 3 : Preprocessed K-means plot k=5

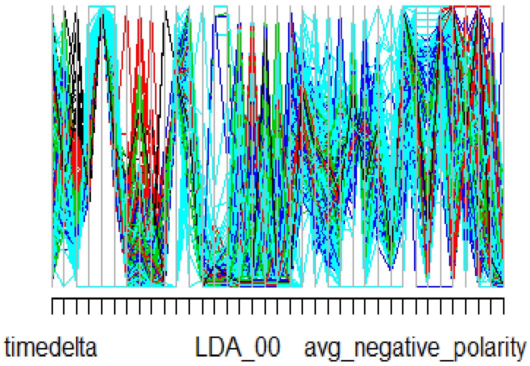


Fig 4 : Parallel coordinators plot

The Figure 4 shows the parallel coordinators plot to see the variables contributed in each cluster.

c) *Fuzzy C-Means*

Fuzzy c means clustering (FCM), each data point has a fraction of membership to each cluster. This algorithm works iteratively until no further clustering is possible. The membership fraction that minimizes the expected distance to each centroid has to be calculated.

$$J_m = \sum_{i=1}^{N}\sum_{j=1}^{C} u_{ij}^{m}\left\|x_i - c_j\right\|^2$$

The algorithm is very similar to K-Means, except that a matrix (row is each data point, column is each centroid, and each cell is the degree of membership) is used.

1. Initialize the membership matrix U
2. Repeat step (3), (4) until converge
3. Compute location of each centroid based on the weighted fraction of its member data point's location.

$$c_j = \frac{\sum_{i=1}^{N} u_{ij}^{m} \cdot x_i}{\sum_{i=1}^{N} u_{ij}^{m}}$$

4. Update each cell as follows

$$u_{ij} = \frac{1}{\sum_{k=1}^{C}\left(\frac{\left\|x_i - c_j\right\|}{\left\|x_i - c_k\right\|}\right)^{\frac{2}{m-1}}}$$

Notice that the parameter m is the degree of fuzziness. The output is the matrix with each data point assigned a degree of membership to each centroids.

d) *Fuzzy C-Means Implementation Using R*

The data repositories used in this paper are The Iris Repository [27]. They are obtained from (http://kdd.ics.uci.edu/). The data set is the fragments of iris flower which is clustered based on the degree assigned by a membership.

The following is the code

```
> library(e1071)
> result <- cmeans(iris[,-5], 3, 100, m=2, method="cmeans")
> plot(iris[,1], iris[,2], col=result$cluster)
> points(result$centers[,c(1,2)], col=1:3, pch=8, cex=2)
> result$membership[1:3,]
         1                      2                   3
[1,] 0.001072018 0.002304389 0.9966236
[2,] 0.007498458 0.016651044 0.9758505
[3,] 0.006414909 0.013760502 0.9798246
> table(iris$Species, result$cluster)
              1  2  3
setosa       0  0 50
versicolor  3 47  0
virginica   37 13  0
```

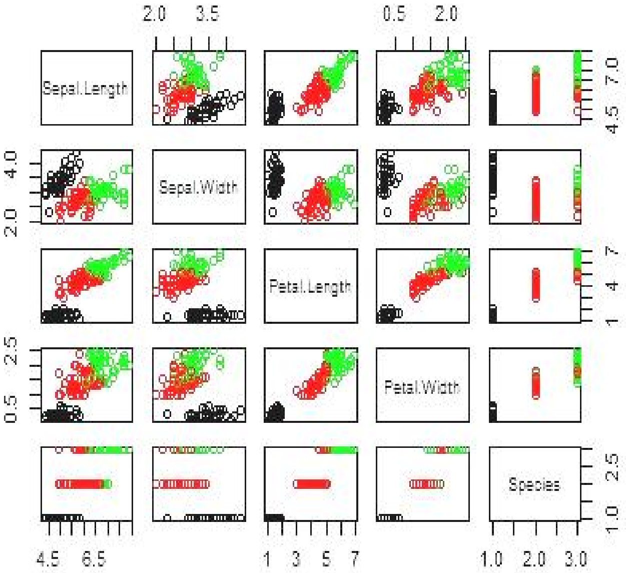


Fig. 5 : Fuzzy C means clustering plot

The observation of the above results and of the Online News Popularity datasets are FCM is mainly dependent on the initial clustering and the computation time is very high for the large data sets. The weight and accuracy are inversely proportional. It is Sensible to noise and membership degree for outliers or noisy points is very low.

e) *Hierarchical Clustering*

Hierarchical Clustering a series of partitions takes place, which may run from a single cluster containing all objects to n clusters each containing a single object. This clustering builds a cluster hierarchy or, in other words, a tree of clusters, also known as a dendrogram [7]. Hierarchical algorithms can be further categorized into two kinds [3]

i. *Agglomerative clustering*: This clustering starts with *n* clusters and iteratively merges the number of clusters which are most similar objects or clusters, respectively, until only one cluster is remaining (*n* → 1).This requires the defining of closest proximity.

ii. *Divisive clustering:* This clustering starts with one cluster and iteratively splits a cluster until singleton clusters of individual points remain, so that the heterogeneity is reduced as far as possible (1 → *n*).This requires the decision of splitting at each step.

The Hierarchical Clustering algorithm [6] below takes an n×n distance matrix d input and increasingly gives n different partitions of the data as the tree it outputs result. The largest partition has n single-element clusters, with every element forming its own cluster. The second-largest partition aggregates the two closest clusters from the largest partition, and thus has -n 1 clusters. In general, the ith partition combines the two closest clusters from the (i − 1)th partition and has (n − i + 1) clusters. Because of the additional complexity of keeping data in a sorted list or heap, so the time required is O(m^{2}log m) and Space required is O(m^2). In this approach, it compares all pairs of data points and merges the one with the closest distance.

Algorithm

```
1: Compute the proximity matrix if necessary
2: repeat
3: Merge the closest two clusters.
4: Update the proximity matrix to reflect the proximity between the new cluster and initial cluster
5: Until only one cluster remains.
```

The Proximity (C$_{i,}$Cj) of clusters C$_i$ and C$_{j,}$ which are of the size m$_i$ and m$_j$,respectively is expressed as

$$Proximity\,(c_i\,,c_j) \;=\; \frac{\sum_{\substack{x\in c_i\\ y\in c_j}} Proximity\;(x,y)}{m_i * m_j}$$

The data set considered is micro RNA expressions. It is actually collected from Fresh paired tumor and control samples from the PAC (Periampullary Carcinoma) patients undergoing Whipple's pan creaticoduodenectomys Data Mining in Health Informatics[15] is an emerging discipline, concerned with developing methods for exploring the unique type of data that come from Health Care database management system. We have also considered the Iris dataset. The code for the implementation is given as follows [16] Figure 6 shows the results of clustering.

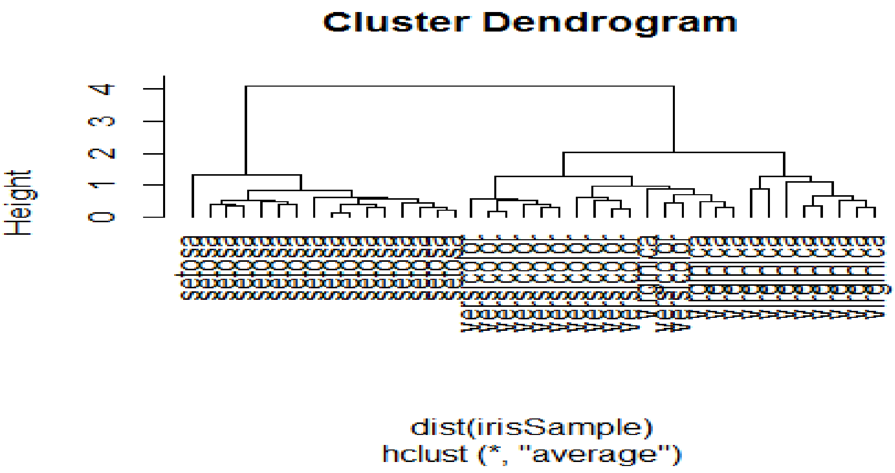


Fig. 6 : Hierarchical Clustering

The Authors [14] have also performed hierarchical clustering on PAC tumors dataset which are distinct from counterpart normal pancreas, normal duodenum, and normal distal CBD and normal ampulla. Unsupervised hierarchical cluster analysis of miRNA Expression profiles were clustered of PAC tumors into pancreatobiliary (n ¼ 23) and intestinal subtypes (n ¼ 17), while normal pancreas (n ¼ 22), normal duodenum (n ¼ 6), normal distal CBD (n ¼ 6) normal ampulla (n ¼ 6) are clustered as different entities (Figure 7).

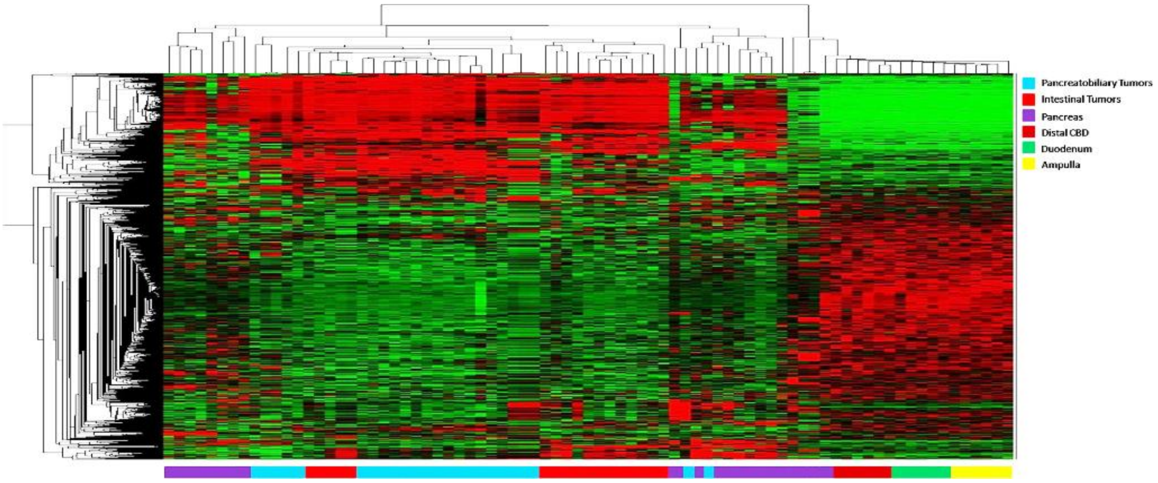


Fig. 7 : The microRNA expression profiles of PAC

IV. Conclusion

We have perceived a comprehensive scan of the k-means, Fuzzy C means and Hierarchical clustering methods using medical research datasets and iris dataset. Using R clustering, Statistical Computing and graphics are represented. All the clustering techniques show ambiguity in clustering noisy data and outliers. The Hierarchical clustering shows good results for small data sets and Fuzzy C means for the voluminous amount of data. K means technique has faster performance but finding the appropriate k value is a big challenge especially in medical research data sets. In continuation to this work we would like to improve clustering analysis by considering the time and accuracy for large data sets using R tool statistics.

V. Acknowledgements

The authors would like to thank Dr. K.S.S Uma Mahesh (M.D Biochemistry) and his lab for sharing their data of PAC tumor samples (microRNA).

References Références Referencias

1. Han J. and Kamber M.: "Data Mining: Concepts and Techniques," Morgan Kaufmann Publishers, San Francisco, 2000.
2. Pang-Ning Tan, Michael Steinbach and Vipin Kumar. Introduction to Data Mining. Addison Wesley; US edition. May 12, 2005. (HC)
3. Data Mining with R: learning by case studies Luis Torgo(kmeans)

4. K. Fernandes, P. Vinagre and P. Cortez. A Proactive Intelligent Decision Support System for Predicting the Popularity of Online News. Proceedings of the 17th EPIA 2015 - Portuguese Conference on Artificial Intelligence, September, Coimbra, Portugal.
5. Jiawei Han and Micheline Kamber, "Data Mining: Concepts and Techniques", Morgan Kaufmann Publishers, second Edition, (2006).
6. Doug Fisher, Optimization and Simplification of Hierarchical Clustering, KDD(HC)
7. Manish Verma, Mauly Srivastava, Neha Chack, Atul Kumar Diswar and Nidhi Gupta, "A Comparative Study of Various Clustering Algorithms in Data Mining", International Journal of Engineering Research and Applications (IJERA) Vol. 2, Issue 3, May-Jun 2012, pp.1379-1384
8. Frank, A. & Asuncion, A. UCI Machine Learning Repository (http:/ / archive. ics. uci. edu/ ml). Irvine, CA: University of California, School of Information and Computer Science 2010.
9. B¨OHM, C., KAILING, K., KRIEGEL, H.-P., AND KR¨OGER, P. 2004. Density connected clustering with local subspace preferences. In Proceedings of the 4th International Conference on Data Mining (ICDM).
10. Robert Gentleman Rafael A. Irizarry Vincent J. Carey Sandrine Dudoit Wolfgang Huber Editors Bioinformatics and Computational Biology Solutions Using R and Bioconductor, Springer
11. R and Data Mining: Examples and Case Studies 1 Yanchang Zhao
12. https://www.r-project.org/
13. Satish Kumar et al Analysis Clustering Techniques in Biological Data with R, International Journal of Computer Science and Information Technologies (IJCSIT), Vol. 6 (2) , 2015
14. KSS Uma Mahes, Kuruva MM , Mitnala S, Rupjyoti T, Venkat RG, Botlagunta S et al. MicroRNA profiling in periampullary carcinoma. Pancreatology 2014; 14: 36-47
15. Abdul Nazeer, K. A., Sebastian M. P., and. Madhu Kumar S.D., 2011 A Heuristic k-Means Algorithm with Better Accuracy and Efficiency for Clustering Health Informatics Data, Journal of Medical Imaging and Health Informatics(American Scientific Publisher) Vol. 1, 66–71.