IJMLNCE
JOURNAL

International Journal of Machine Learning and Networked Collaborative Engineering

Journal Homepage: http://www.mlnce.net/home/index.html

DOI : https://doi.org/10.30991/IJMLNCE.2019v03i01.004

Deterministic Machine Learning Cluster Analysis of Research Data: using R Programming

Yagyanath Rimal

Faculty of Science and Technology

Pokhara University, Nepal

Email: rimal.yagya@gmail.com

https://orcid.org/0000-0003-1045-7728

Abstract

This review paper clearly discusses the compression between various types of cluster analysis of different data sets were explained sufficiently. Although there is large gap between the way of analysis of collected data and its cluster categorization research data using r programming. Its primary purpose is to explain the simplest way of clustering analysis whose data structure were wide scattered using R software whose outputs were sufficiently explain with various inter-mediate output and graphical interpretation to reach the conclusion of analysis. Therefore, this paper presents easiest way of clustering when data sets with large dimensions with multivariate analysis of iris , utilities, mclust and dbscan data sets from internet and its strengths for data analysis using R programming.

Keywords

Data Analytic Machine Learning, K mean, Hierarchical, Dbscan

1. Introduction

Data science is not only data collection as a database, but it is an interdisciplinary science of statistics, data analysis and automatic learning of the scientific methods of process algorithms. The data collection, data processing and data analysis with different algorithms and, visualization of data which require the knowledge of calculation, mathematics, statistics and interpretation in each step to analyze research data. However, the data available in today's world in any structured, unstructured and semi-structured format are raw data. Raw data includes the integration of data and the selection of required data. The author [1] claimed that data cleaning takes 50 percent to 80 percent of a data scientist's work is significant which needs some methods and algorithms to analyzed and then a visual report is produced from data base. The data column is known as data structured. The data scientist creates many unstructured data formats. At one point, the data scientist integrated data from different sources [2]. Then selection of part for analysis, so that the data scientist finally makes the normalization of data with various transformation records. On many cases the data analyzer omits missing data, has been cleaned up while some computer scientists do so. This process needs machine learning for establishing the statistical requirements of the system for automatic processing. Therefore, modern data scientists need four pillars of statistical, computer skills, communication and visualization and competencies of many domain [3]. R programming with supervised learning has the best tools for data analysis; The data scientist can use many other data analysis processes. Machine learning can be applied in the analysis of neural networks, deep learning and artificial intelligence networks. The machine learning always requires information of requirement, since formal entry into the system produces the program and is defined as output according to the requirements of the system. Therefore, machine learning is the system that takes input and output as input, so the computer system automatically produces the model based on some previous parameters. For the simplest example, x entry and y is exit of given table, so find that the prediction at 4 is 40 with a relation having y = x * 10. Based on static observation, machine learning predicts its relationship. Similarly, example when x is input z is another output, so it predicts that its relation (14 + 4) between x and z is 26. So, the machine learning will be predicting the value of z when x is 500. This could be easily calculated with y = mx + c where m is slope 4, which is calculated with the relation y2□y1=x2□x1 and c is the intercept with the constant value when the line of mean intersection of the slope on y is 10. Therefore, the prediction when x becomes 500, the value of z becomes 2010 (4X + 10) relationship.

```
> fitk=kmeans(irs,3)# here we are using 3 cluster of species
> fitk
```

Table 1: Data

X	Y	Z
1	10	14
2	20	18
3	20	22
4	?	?

The data science work flow includes the process of defining cyclical problems, collecting data, developing models, implementing models, improving and monitoring performance. According to [7] the R, Python and Weka are the most important technologies used in science in modern world. Therefore, machine learning strategies first, learns the relationship and then predicts the relationship [4]. First data processing phase is data processing and then the processing data were translated into relationship for other data prediction. This process is known as the conversion of data into a statistical analysis design. The second step is to send the model, the third step is to evaluate the performance that is always compared with the output data in a rigorous process, which is accurately evaluated against errors. There were some techniques to improve performance are classification, grouping and regression in machine learning [1]. Machine learning technologies use R an open source programming language developed in 1993 by Ross Ilha and Robert Gentleman with dynamic programming [5]. which supports large complete functional package is installed in the r environment for a special package. R is a binary package, which does not require compilation, but must be connected to the r console with a different library for special packages. Clustering is one of the most important methods for extracting data to discover knowledge in multidimensional data. The purpose of grouping is to identify models or groups of similar objects. Cluster analysis is a group of data objects based on the information found in objects and data groups. The grouping means increasing the grouping distance among research data sets with various another field relationship. The big data analysis, supply chains network analysis, data interpretation predicts the future are the key application of data clustering. In many times running organization can change plans when the results are changed and success is achieved after data analysis [6]. Hierarchical grouping is also known as nested groupings that are groupings to form a tree, which uses local heuristics to form a nesting hierarchy of nested datasets. The most notable exception of the hierarchical grouping of a single link is distorted into convex and hyper spherical groupings [7]. The advantages of hierarchical clustering are low efficiency, as it has a time complexity of O(n), unlike the linear complexity of K-Means and GMM.

Partitioned cluster is simply the division of grouped data objects that do not overlap, so that each object is exactly in a subset. The exclusive grouping assigns each object to a single grouping. The stacking grouping used to reflect the fact that grouping widespread in this grouping. The hierarchical cluster analysis uses a set of differences in the object being grouped with the objectives of compact groups of appropriately the same diameter. The well ranked cluster mapped with distance between any two points within a group. The prototype cluster is often measuring with continuous data attributes, since clutters tend to be globular.

The density-based cluster is used when groups are irregular and with noise in data sets values is grouped up. Therefore, grouping, is the process of identifying the model in groups for effective segregation into groups of large data groups with various relationship among them. The most significant clustering is K means which group the data based on prototypes that try to define the number of groups (k) that are represented as centroids when the samples were more than 100. The agglomerated hierarchical grouping is closed groupings until a single and global grouping is maintained. Similarly, Dbscan is a density-based clustering algorithm that produces a particular grouping, in which the algorithm determines the number of clusters [7].

```
> data("iris")
Sepal.Length Sepal.Width Petal.Length Petal.Width   Species
1          5.1         3.5          1.4         0.2     setosa
2          4.9         3.0          1.4         0.2     setosa
3          4.7         3.2          1.3         0.2     setosa
> plot(iris)
```

2. Using R Programing

For detail analysis here researcher tries to the explanation with iris data sets with records of 150 observations in 5the categories.

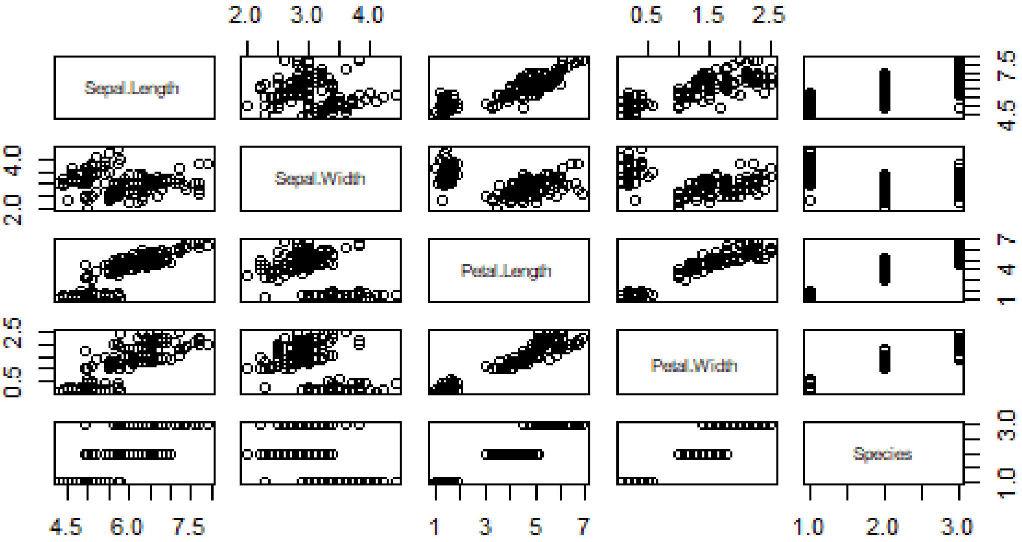

Fig1: Edgar Adelson’s Pair Scatter Plot

Fig1 shows the multidimensional figure of five variables of all the recorders were shown. mostly the last field species clearly demonstrates three categories most. Speal length, sepal width petal length petal width were numeric continuous variable records. Where there were some records were scattered in some cluster but how many clusters were fit those variables is primary of researcher judgement. The scaling and normalization are always required when the data set have large variation in data structure. If in case some data were in cm and some data were in m or km. in such case, we need mean and standard deviation of those data should be calculated. But there the last datasets species have three categories of species name should be excluding categorical variable form the rest of data sets using scale function.

```
> irs=scale(iris[,-5])  all row and columns except 5th
> irs
  Sepal.Length Sepal.Width Petal.Length   Petal.Width
  [1,]  -0.89767388  1.01560199  -1.33575163 -1.3110521482
   ……………………………………………………
  [150,]  -1.13920048 -0.13153881  -1.33575163 -1.3110521482
```

The kmean clustering is the process of finding k means center of data where each data is centered from k mean using kmean function. Each center searches the distance between observed value were grouped into based on data relationship to each cluster. All the data points were falls into k means of clusters. If we set out 4 k means it developed 4 clusters data groups from research data.

```
> fitk=kmeans(irs,3)# here we are using 3 cluster of species
> fitk
```

K-means clustering with 3 clusters of sizes 47, 53, 50 is the number of observations within each cluster. Cluster means:

```
Sepal.Length Sepal.Width Petal.Length Petal.Width
1   1.13217737  0.08812645    0.9928284   1.0141287
2  -0.05005221 -0.88042696    0.3465767   0.2805873
3  -1.01119138  0.85041372   -1.3006301  -1.2507035
```

```
Clustering vector:
[1] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 33 3 1
…..
1 1 1 2 1 1 1 1 1 1 2 2 1 1 1 1 2  1 2 1 2 1 1 2 1 1 1 1 1 1 2 2 1 1 1 2 1 1 1 2 1 1 1 2 1 1 2
Within cluster sum of squares by cluster:
[1] 47.45019                    44.08754               47.35062
(between_SS / total_SS =  76.7 %)
```

```
> irs=scale(iris[,-5])  all row and columns except 5th
> irs
  Sepal.Length Sepal.Width Petal.Length   Petal.Width
  [1,]  -0.89767388  1.01560199  -1.33575163 -1.3110521482
   ……………………………………………………
  [150,]  -1.13920048 -0.13153881  -1.33575163 -1.3110521482
```

Which implies how well the cluster dispersed among and within the data sets the 76.7% higher the percentage means higher the dispersed among data sets. Available components:

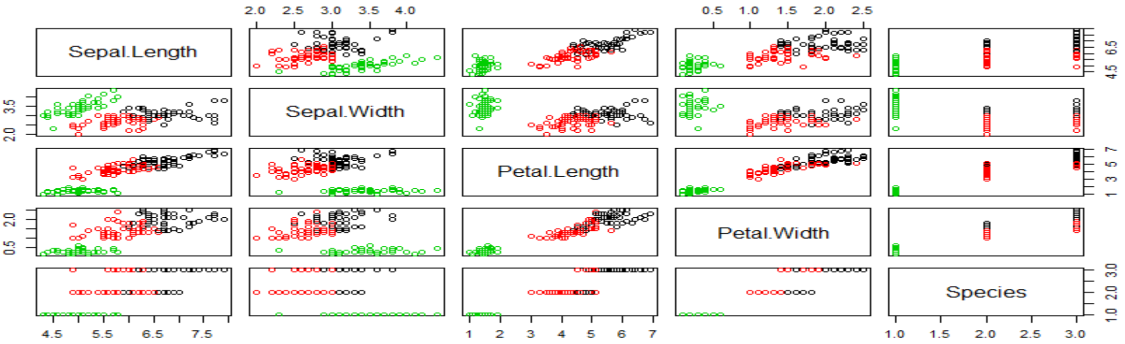

Fig2: Scatter Plot With All Cluster

```
[1] "cluster" "centers"  "totss" "withinss" "tot.withinss" "betweenss" "size" "iter"  "ifault"
> str(fitk) display the structure.    List of 9
 $ cluster     : int [1:150] 3 3 3 3 3 3 3 3 3 3
 $ centers     : num [1:3, 1:4] 1.1322 -0.0501 -1.0112 0.0881 -0.8804
attr(*, "dimnames")=List of 2
$ : chr [1:3] "1" "2" "3"
$ : chr [1:4] "Sepal.Length" "Sepal.Width" "Petal.Length" "Petal.Width"
$ totss          : num 596
$ withinss     : num [1:3] 47.5 44.1 47.4
$ tot.withinss: num 139
$ betweenss   : num 457
$ size         : int [1:3] 47 53 50
$ iter          : int 3
$ ifault       : int 0
attr(*, "class")= chr "kmeans"
>  plot(iris,col=fitk$cluster)
```

The above table demonstrates the three cluster with different colors the nearest data were in same colors than others data structure which separate from another. Here is great question how many clusters does it optimum fits for those records. To do so the empty list of k is designed and repeat the same procedure of kmean into list process is carried out.

```
k=list()
for(i in 1 :10){
k[[i]]=kmeans(irs,i)}
k
```

Which produces the 10 integration and it produced the sum of square and total sum of square of each reputation as below the k mean fit where there was gentle slope of data flow.

```
> plot(1:10,betweenss_totss,type="b",ylab="between ss/total SS",xlab="center(k)")
```

```
[[1]]
K-means clustering with 1 clusters of sizes 150.Cluster means:
Sepal.Length  Sepal.Width Petal.Length   Petal.Width
-9.793092e-16       4.503805e-16       5.107026e-17       -6.217249e-17
Clustering vector:
[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] 596
(between_SS / total_SS =   0.0 %)
Available components:
[1] "cluster" "centers"  "totss" "withinss" "tot.withinss" "betweenss"  "size"  "iter" "ifault"
[[10]]
K-means clustering with 10 clusters of sizes 21, 16, 12, 17, 13, 17, 16, 12, 11, 15
Cluster means:
   Sepal.Length Sepal.Width Petal.Length  Petal.Width
1    -0.9666815  0.92820079  -1.29259152 -1.217343093
………………
10    0.9942845  0.35790793   1.04344975  1.452740239
Clustering vector:
[1]  1  4  4  4  1  8  1  1  4  4  8  1  4  4  8  8  8  1  8  8  1  1  1  1  1  4  1  1  1  4  4  1   4  4  1 1
  ………..
 9  6 10  2  9  2 10  9  2  7  2  9  9  9  2  3  2  9 10  7  7 10 10 10  2 10 10 10  2  7 10  7
Within cluster sum of squares by cluster:
[1]  3.397867  5.606072  1.858918  5.163861  2.347768  9.671551  5.324952  3.954505 10.2
03534  4.597857
(between_SS / total_SS =  91.3 %)
Available components:
[1] "cluster"   "centers"   "totss" "withinss"  "tot.withinss" "betweenss" "size" "iter"  "ifault"
> betweenss_totss=list()
> for(i in 1:10){
betweenss_totss[[i]]=k[[i]]$betweenss/k[[i]]$totss }
```

This process calculates the 10 ratios of total sum of square in between_totss of k index whose output is further plot and produce with type is b means both dots line type in between y lab and x lab.

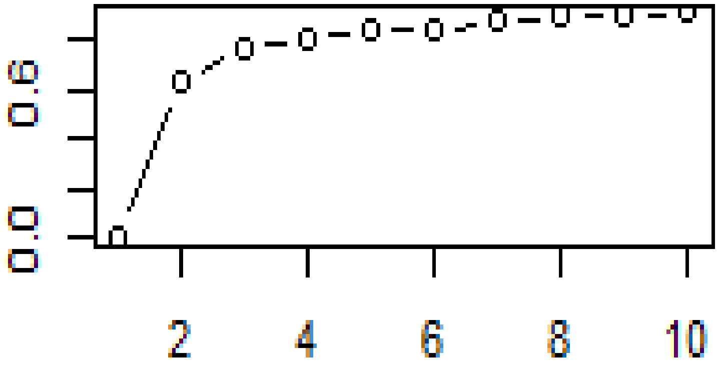

Fig:3 Best fit Cluster Plot

This picture shows the best fit of cluster in k means on x axis and between sum of total sum of square is in y axis, when first sharp drop-down point list between two points were neglected when there is normality between adjoin two points is the best fitted line area of k group. Here the two to four k mean considered as the best cluster groups is best fitted. Similarly, the 4 k mean cluster plot could be easily plot as.

```
>for(i in 1:4){ plot(iris,col=k[[i]]$cluster) }
```

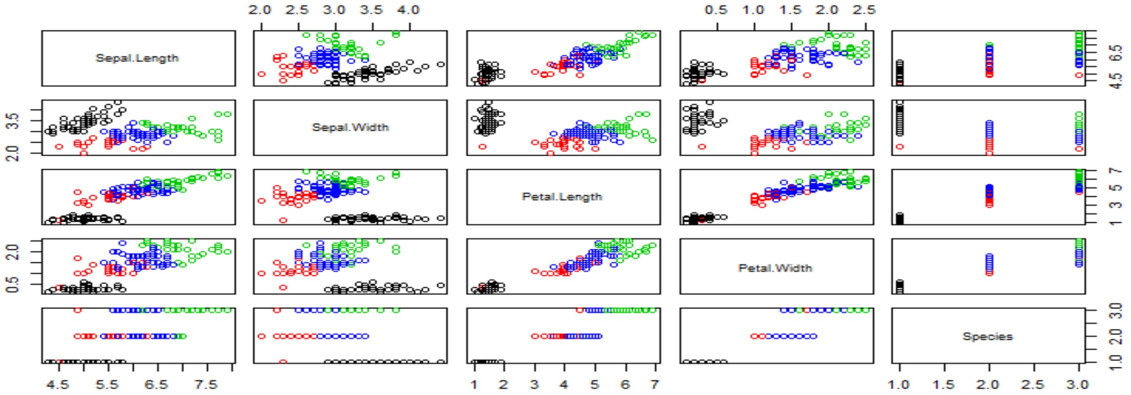

Fig 4: K Means Four Fitted Cluster Plot

```
>for(i in 1:3){  plot(iris,col=k[[i]]$cluster) }
```

This demonstrate four k means fitted model cluster produced but there is highly overlapped one another which is suite for the iris data sets where red, black and green points were mixed so this is not best fitted cluster of this data sets.

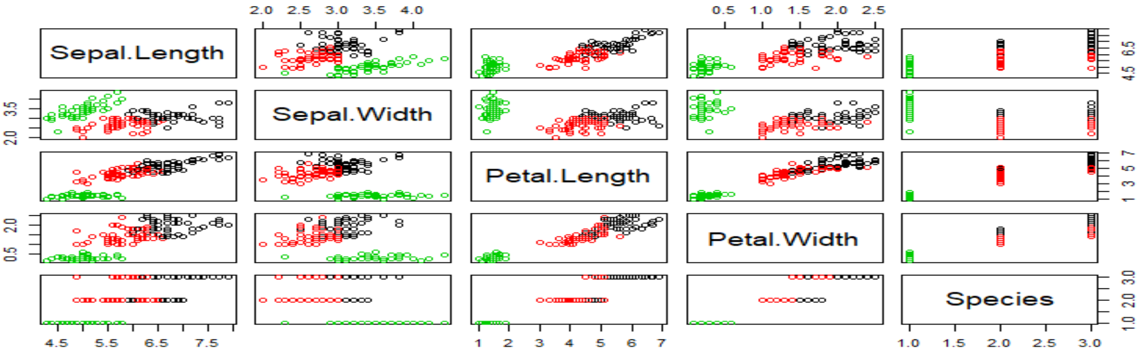

Fig 5: K Mean Three Cluster Scatter Plot

This cluster is quite best but there was less overlap than above figure when clustering the principle behind that the intra cluster dependency should consider as less one another in many cases the red and black cluster groups were found mixture in fixed side although.

```
>for(i in 1:2){ plot(iris,col=k[[i]]$cluster) }
```

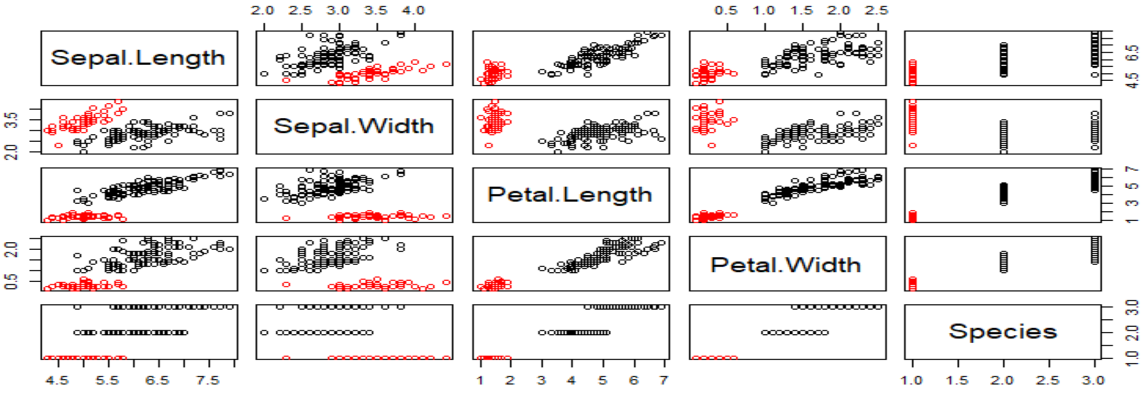

Fig: 6 K Mean Two Cluster of Scatter Plot

Figure conclude that the most fitted k mean is 2 which is generalized the figure and data structure of specific species data correlation among various data types on the data sets. Where each group are completely separated another group is best perfect than other 4 and three k mean.

3. Hierarchal Clustering

The hierarchal clustering places each data point in hierarchal order of its weight which measures each item in hierarchal **pattern [8].** It also measures how each point are similar to each other in tree format. The distance measures of every point of data sets using dist. function are merged typically to formed hierarchal tree pattern rather than one cluster to another cluster.

```
>d=dist(irs)
>fith=hclust(d,"ward.D2")
>plot(fith)
```

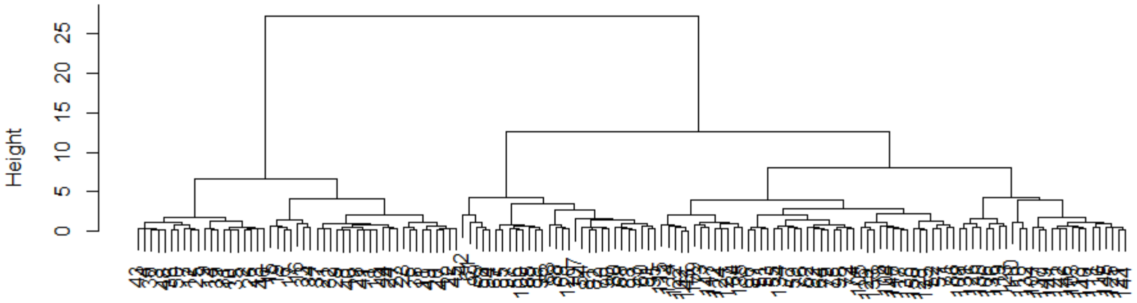

Fig: 7 Cluster Hierarchal Plot

From the above figure each similar observation is in lower cluster then the higher cluster was again connected with above cluster finally formed largest at top. Demonstrate higher cluster in top as parent cluster of another smaller child. The data scientist has to design where to cut to get appropriate data proportion. The lower down the cluster cut the more the cluster we get from the iris data sets. To fit the data structure in fixed cluster cut tree function is used.>rect.hclust(fith,k=3,border="red")

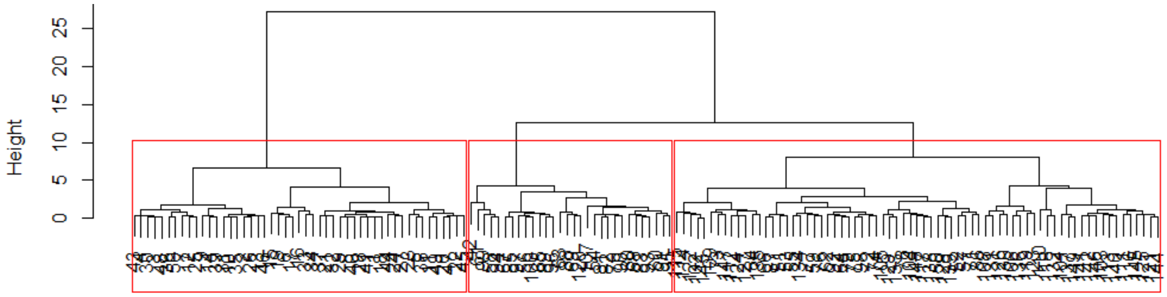

Fig: 8 Hierarchal cluster with Border Plot Dendrogram

From the above figure the three clusters will be easily selected when we cut at the red line point.

```
> cluster=cutree(fith,3)
> cluster
  [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 1 1 1 2 1 1 1 1
 1 1 1 1 3 3 3 2 3 2 3 2 3 2 2 3 2 3 2 3 2 2  2 2 3 3 3 3 3 3 3 3 3 2 2 2 2 3 2 3 3 2 2 2 2 3 2 2
 2 2 2 3 2 2 3 3  3 3 3 3 2 3 3 3 3 3 3 3 3 3 3 3 3 2 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
> plot(iris,col=cluster)
```

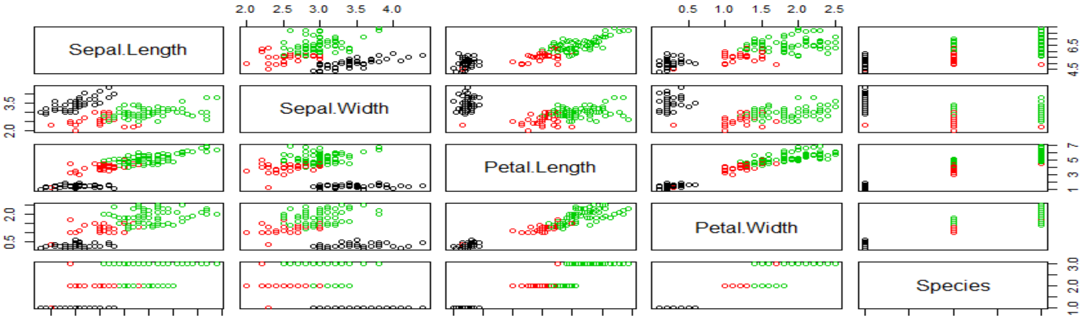

Fig: 9 Scattered Plot of Three Cluster

```
> cluster=cutree(fith,3)
> cluster
  [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 1 1 1 2 1 1 1 1
 1 1 1 1 3 3 3 2 3 2 3 2 3 2 2 3 2 3 2 3 2 2  2 2 3 3 3 3 3 3 3 3 3 2 2 2 2 3 2 3 3 2 2 2 2 3 2 2
 2 2 2 3 2 2 3 3  3 3 3 3 2 3 3 3 3 3 3 3 3 3 3 3 3 2 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
> plot(iris,col=cluster)
```

Fig: 9(a) Scattered Plot of Three Cluster

```
> cluster=cutree(fith,3)
> cluster
  [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 1 1 1 2 1 1 1 1
 1 1 1 1 3 3 3 2 3 2 3 2 3 2 2 3 2 3 2 3 2 2  2 2 3 3 3 3 3 3 3 3 3 2 2 2 2 3 2 3 3 2 2 2 2 3 2 2
 2 2 2 3 2 2 3 3  3 3 3 3 2 3 3 3 3 3 3 3 3 3 3 3 3 2 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
> plot(iris,col=cluster)
```

Fig: 9(b) Scattered Plot of Three Cluster

```
> cluster=cutree(fith,3)
> cluster
 [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 1 1 1 2 1 1 1 1
 1 1 1 1 3 3 3 2 3 2 3 2 3 2 2 3 2 3 2 3 2 2  2 2 3 3 3 3 3 3 3 3 3 2 2 2 2 3 2 3 3 2 2 2 2 3 2 2
 2 2 2 3 2 2 3 3  3 3 3 3 2 3 3 3 3 3 3 3 3 3 3 3 3 2 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
> plot(iris,col=cluster)
```

Fig: 9(c) Scattered Plot of Three Cluster

```
> cluster=cutree(fith,3)
> cluster
 [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 1 1 1 2 1 1 1 1
 1 1 1 1 3 3 3 2 3 2 3 2 3 2 2 3 2 3 2 3 2 2  2 2 3 3 3 3 3 3 3 3 3 2 2 2 2 3 2 3 3 2 2 2 2 3 2 2
 2 2 2 3 2 2 3 3  3 3 3 3 2 3 3 3 3 3 3 3 3 3 3 3 3 2 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
> plot(iris,col=cluster)
```

Fig: 9(d) Scattered Plot of Three Cluster

```
> cluster=cutree(fith,3)
> cluster
 [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 1 1 1 2 1 1 1 1
 1 1 1 1 3 3 3 2 3 2 3 2 3 2 2 3 2 3 2 3 2 2  2 2 3 3 3 3 3 3 3 3 3 2 2 2 2 3 2 3 3 2 2 2 2 3 2 2
 2 2 2 3 2 2 3 3  3 3 3 3 2 3 3 3 3 3 3 3 3 3 3 3 3 2 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
> plot(iris,col=cluster)
```

Fig: 9(e) Scattered Plot of Three Cluster

```
> cluster=cutree(fith,3)
> cluster
 [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 1 1 1 2 1 1 1 1
 1 1 1 1 3 3 3 2 3 2 3 2 3 2 2 3 2 3 2 3 2 2  2 2 3 3 3 3 3 3 3 3 3 2 2 2 2 3 2 3 3 2 2 2 2 3 2 2
 2 2 2 3 2 2 3 3  3 3 3 3 2 3 3 3 3 3 3 3 3 3 3 3 3 2 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
> plot(iris,col=cluster)
```

Fig: 9(f) Scattered Plot of Three Cluster

This figure demonstrates the three-clustering using hierarchal method is exactly similar in above k mean figure however there were some mixed overlap too when we set with three clusters

3.1. Model Based Clustering

Model based clustering always use various best models in various ellipsoidal, varying volume shape and orientation with mclust library.

```
> library(mclust)
> fitm=mclust(irs)
fitting ======================| 100%
>plot(fitm)
Model-based clustering plots:  1: BIC   2: classification  3: uncertainty 4: density Selection: 1
```

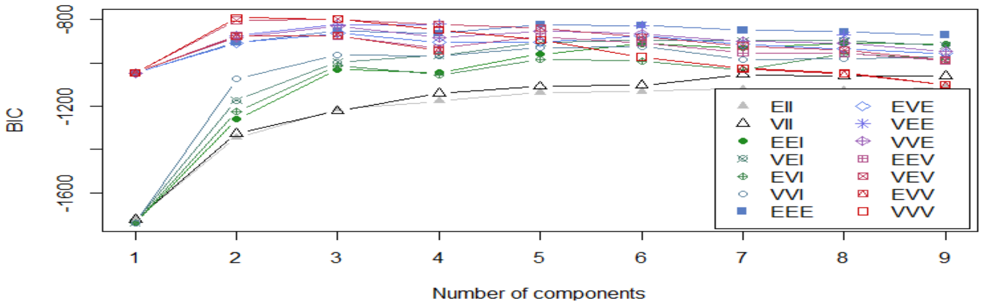

Fig: 10 Baysiean Cluster:

The Baysiean information cluster indicates the plot describes in x axis there were number of clusters where as in y axis based on criteria measurement. It further explained the how much variability of model fits he data the higher the value is the best fit the model. The above red cluster is the best model fitted and measured the at two with VVV cluster, describes that the adding the additional data points with cluster doesn't affects the cluster relationship therefore the best two cluster mostly describes the fitted in the data sets.

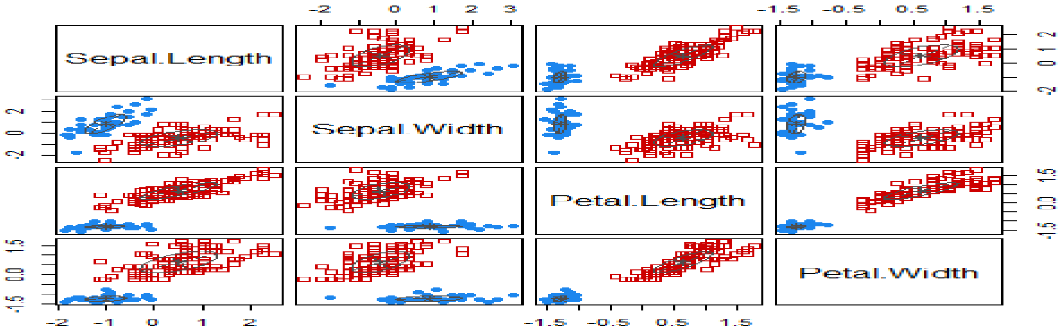

Fig: 11 Model-based Clustering Plots

The classification cluster describes the two different cluster is best fitted where two species where the significant on data sets is the best fitted.

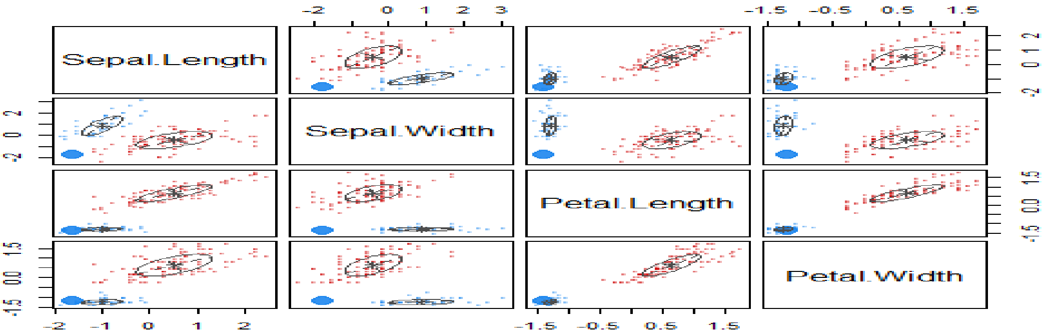

Fig: 12 Uncertainty Cluster Plot

This plot describes the uncertainty plot of subject of data structure the blue points does not fall inside the large circle where there were some cluster does no exactly fit the cluster were many uncertainty points values.

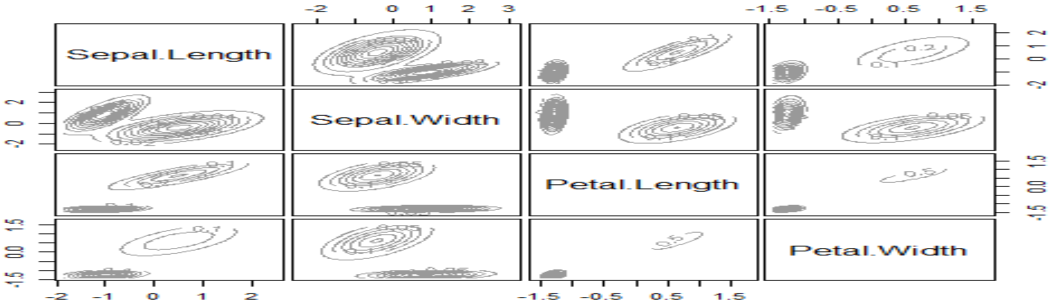

Fig: 13 Density Based Cluster Plot

The density-based cluster is final model of cluster which describes the two bivariate normal distribution of data to formed its circle.

4. Density Based Clustering

Density based clustering is another way of grouping which measures the cluster in which date were mostly falls in region where mostly data values lies [9]. When another cluster again formed when there were mostly likely values lies which can easily be calculated with package dbscan.

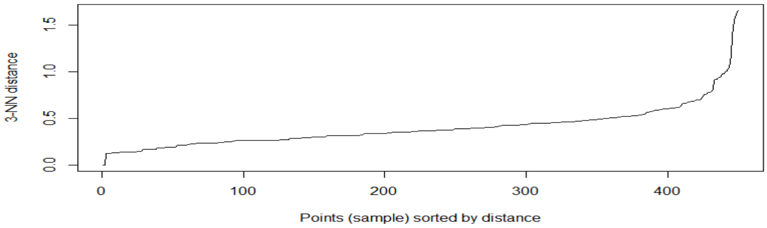

Fig: 14 The Line Graph

```
> install.packages("dbscan")
> library(dbscan)
> kNNdistplot(irs, k=3)
```

This figure describes the knndist distance with three groups of species.
> abline(h=0.7,col="red",lty=2)

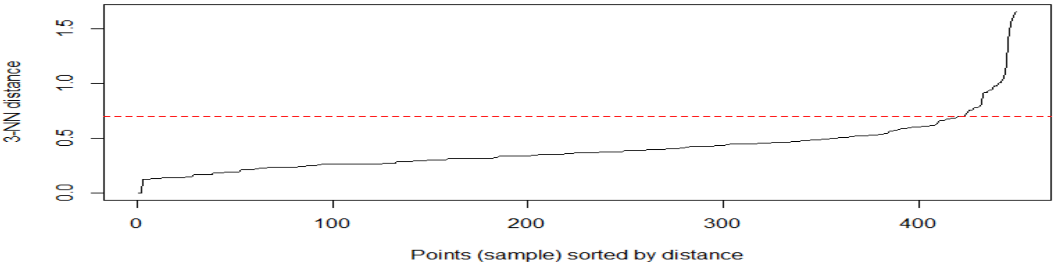

Fig: 15 Abline Plot with Cuts

This figure cuts abline at 0.7 points with.

```
> fitd=dbscan(irs,eps=0.7,minPts = 5)
> fitd
DBSCAN clustering for 150 objects. Parameters: eps = 0.7, minPts = 5. The clustering cont
ains 2 cluster(s) and 6 noise points.   0  1  2
                                         6 48 96
```

Available fields: cluster, eps, minPts. This dbscan nicely groups 2 different cluster of 48 and 96 cases were completely similarities but there were 6 noise cases with dissimilarities out of 150 records of numerical data sets. > plot(iris,col=fitd$cluster)

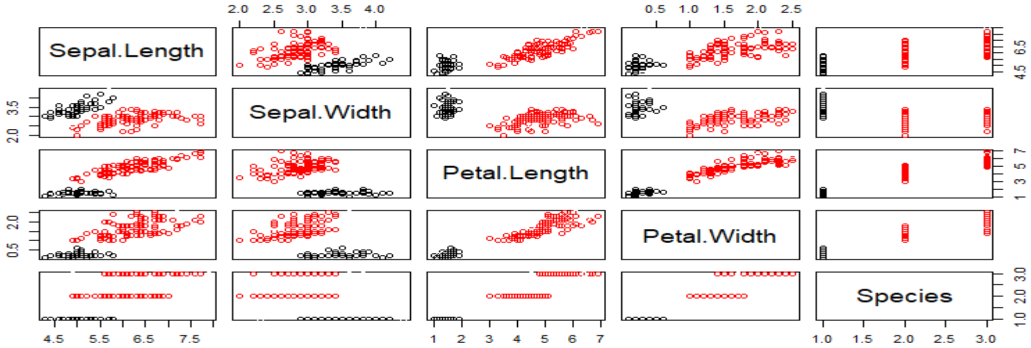

Fig: 16 Hierarchal Clustering with 2Ccluster of Iris

This figure also further same results of kmean and hierarchal clustering when we set out 2 cluster design of iris data sets. There were two groups of data having best clustering. Thus, when we add more variables makes its values with similar with group after elbow is necessary for data analysist. Therefore, principle component reduction techniques support clustering of data sets before data analysing.

Using R Programming

Another data sets of 27 records with 8 variables could analysed other types clustering.

```
>library(readxl)
> Utilities <- read_excel("C:/Users/Rimal Sir/Desktop/Utilities.xls")
> View(Utilities)
> zz=Utilities[,-c(1,1)] # exclude name field
> zz
# A tibble: 27 x 8
   Charge Salary  load demand  fuel gasbill elecbill totalbill
 1     26    680   194     5.5  210   113.       68.2       4.82
 2     18    600   164     5.6  139.   95.9      43.3       4.64
# ... with 25 more rows
> plot(Salary~fuel,data=Utilities) with(Utilities,text(Salary~fuel,labels
=company,pos=1, cex=.5))
```

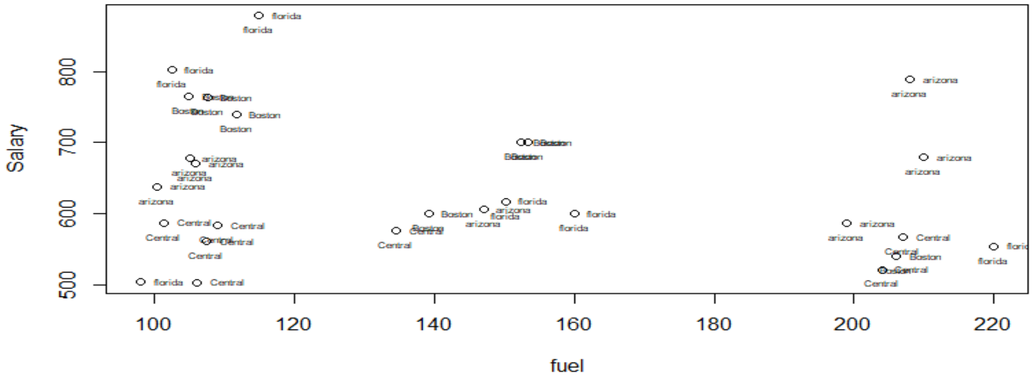

Fig:17 Sales Vs Fuel Consumptions of Utilities

This plot demonstrate sales vs fuel consumpton data of utilities database with name of country of each cluster.pos means position of points and cex is size of text in plot. T here were some company which has high fuel cost but low salary and the middle has averge in both fuel and salary where as one cluster has low to high salary will be one cluster. The normalization is sometime very required if data has high ranges of values this could be achieve using substracting mean and dividing standard deviation of all data sets where company name field should be removed first.

```
> m=apply(zz,2,mean)# 2 for data are in columns
> s=apply(zz,2,sd)# 2 for column
> z=scale(zz,m,s)#normalized dataset of utilities table
> print(z,digit=2)# 2 dugits data only
> print(z,digit=3)
       Charge  Salary    load demand    fuel gasbill elecbill totalbill
 [1,] -1.1773  0.3913  1.4084  1.069  1.5477   1.050   0.3123     0.651
 [2,] -1.5643 -0.4144  1.0187  1.107 -0.0920   0.722  -0.8581     0.505
[27,]  1.0482 -0.5453 -0.9171 -0.832 -0.9669  -0.802   1.0431    -0.401
attr(,"scaled:center")
```

```
> hc.c=hclust(d)
> hc.c
Call: hclust(d = d)
Cluster method   : complete  Distance        : euclidean  Number of objects: 27
> plot(hc.c,hang=1,labels=Utilities$company)
```

```
Charge    Salary     load    demand     fuel   gasbill  elecbill totalbill
50.33    641.15     85.59     2.69    143.15    58.77     61.59      4.00
attr(,"scaled:scale")
Charge    Salary     load    demand     fuel   gasbill  elecbill totalbill
20.67     99.30     76.97     2.63     43.19    51.41     21.32      1.26
> d=dist(z)# eucalidan distance calculation
> print(d,digits=2)# making more compact
1
2  2.27
3  2.52 2.54
4  4.48 4.30 5.62
5  3.98 2.93 5.15 3.20
6  2.39 2.59 3.78 4.04 2.55
7  4.74 3.98 5.79 3.09 1.57 2.95
8  5.41 5.44 7.25 4.67 3.16 3.65 2.81
………………………………………………………………

27 4.64 4.96 6.01 2.46 3.42 3.51 2.55 3.47 2.27 3.17 2.75 3.93 7.66 2.95
```

The Euclidean distance will be calculated with distance function while print function shows only digits specified so that data become normal looks of all records. The hierarchal cluster dendrogram can be easily calculated with hclust function the default is complete linkage with eculidan distance whose plot can be easily calculated with hang.

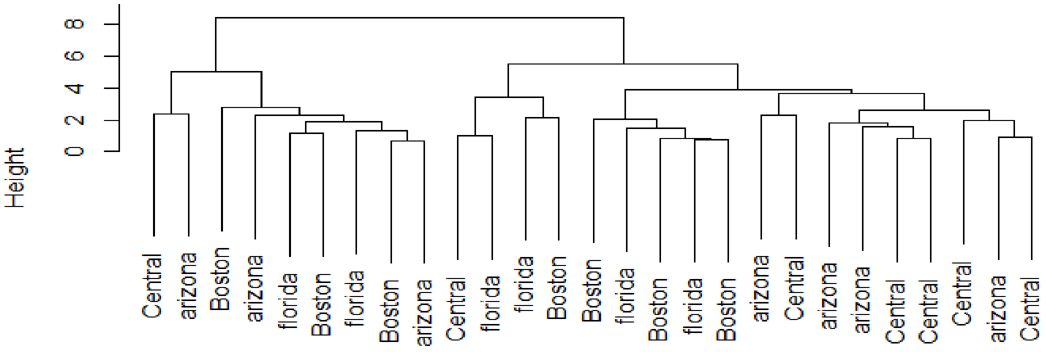

Fig: 18 Cluster Dendrogram of Utilities

The hierarchal dendogram display each company of its variable wastage with its name by its when certs and amazon are in similar data value formed one cluster which similarly goes all data values formed hierarchal clustering.

```
> hc.a=hclust(d,method="average")
> hc.a
Call: hclust(d = d, method = "average")
Cluster method   : average
Distance         : euclidean
Number of objects: 27
> plot(hc.a,hang=1)
```

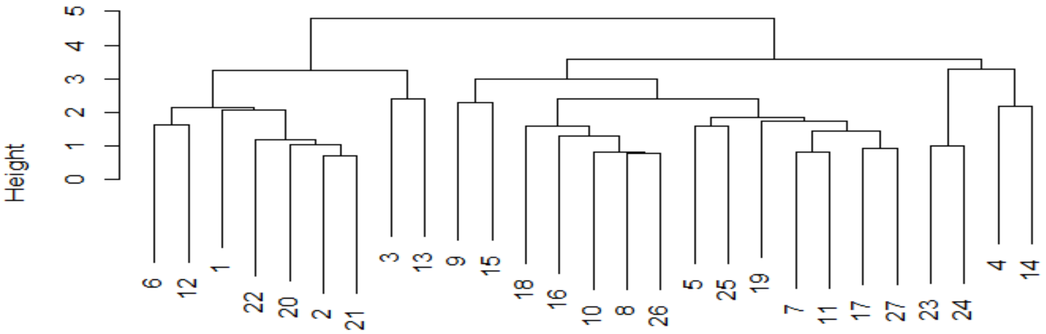

Fig: 19 Cluster Dendrogram of Hang1

From the above figure the company 22 and 23 are formed one cluster then goes subsequently union single cluster tree formed.

```
> m.c=cutree(hc.c,3)
> m.a=cutree(hc.a,3)
> table(m.c,m.a)
       m.a
m.c     1  2  3
     1  9  0  0
     2  0  4  0

     3  0  0 14
```

The cutree function cuts its height at specified location and the table command shows two-dimensional table of data sets presents below clusters cuts. The table show that the average linkage there were 9, 4 and 14 cluster whereas complete cluster there were 9, 4 and 14 clusters in each label, however there were not any company presents both companies.   The average of two cluster can calculated with aggregate function.

```
> aggregate(z,list(m.c),mean)
Group.1      Charge     Salary      load     demand       fuel    gasbill   elecbill  totalbill
1     0.1019961  -0.179411  1.271287  1.263528  0.582989  1.214592 -0.587976  1.133909
2    0.0685389 -1.167201 -0.374722 -0.452166  0.332376 -0.472027 -1.275203 -0.879289
3     0.6888436  0.448822 -0.710193 -0.683079 -0.469743 -0.645945  0.742328 -0.477716
```

The above data shows that the fuel has significant among three company varying .33 to -0.46 is lower than average consumption and total bill also has high variation mean with in cluster.

```
> aggregate(Utilities[,-c(1,1)],list(m.c),mean) # in o riginal units
  Group1   Charge   Salary load   demand  fuel   gasbill elecbill totalbill
1. 127.55556 623.3333 183.44444 6.0111111 168.3344 121.20556 49.05778  5.43325
2. 2 51.75000 525.2500  56.75000 1.5000000 157.5100  34.50000 34.40750  2.893158
3. 3 64.57143 685.7143  30.92857 0.8928571 122.8650  25.55929 77.41714  3.399837
```

Similarly, the load 183 is the highest mean of company 1 to 30 lowest mean of company 3. Silhoueplot demonstrate the bar diagram of all cluster members.

```
> library(cluster)
> plot(silhouette(cutree(hc.c,4),d))
```

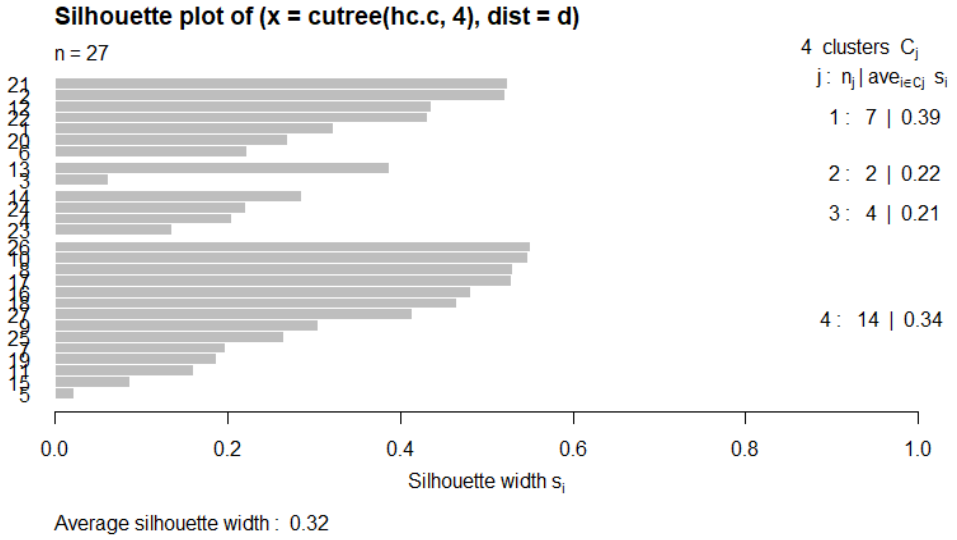

Fig: 20 Sllhouette Plot with 4 Cutee

The scree plot calculates sum of square within group.

```
> wss=(nrow(z)-1)*sum(apply(z,2,var))
> for(i in 1:20)wss[i]=sum(kmeans(z,centers=i)$withinss)
> plot(1:20,wss,type="b",xlab="Number Cluster",ylab="Within Groups")
```

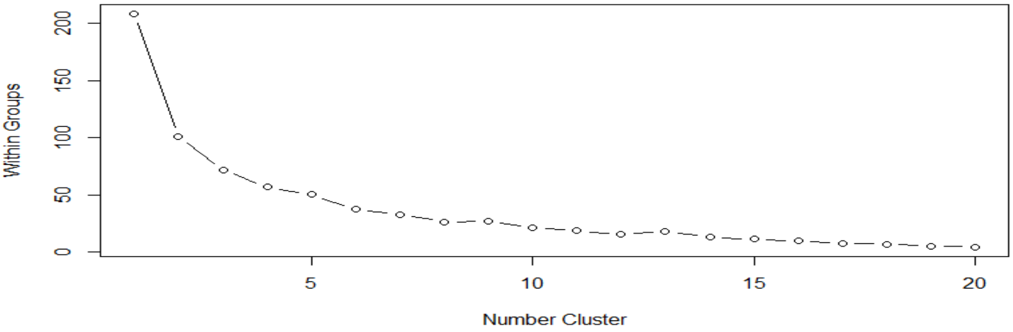

Fig: 21 Line Plot of Number of Cluster within groups

The scree plot shows that the three is high falls of data in 1 to 2 cluster when it increases it will not such significant when it reaches 5 cluster there is no such significant within group.

5. Conclusion

Cluster analysis is a technique used to classify objects or cases into related groups, is sometime called conglomerates. Cluster analysis is also classification analysis or numerical. if there is no group information or cluster membership for any of the objects is generally known as unsupervised clustering. Clustering procedures in cluster analysis can be hierarchical, non-hierarchical, or twostep procedures. A hierarchical procedure in the analysis of conglomerates is characterized by the development of a structure similar to a tree. A hierarchical procedure can be agglomerating. Agglomeration methods in cluster analysis consist of linking methods, variance methods, and centroid methods. Non-hierarchical methods in cluster analysis are often referred to as clustering. The two-step procedure can automatically determine the optimal number of groupings by comparing the model values. The choice of grouping procedures and the choice of distance measurement are correlated. The correlated dimensions of the groupings in the cluster analysis must be significant. Groups should be interpreted in terms of group centroids describes which variables or samples belong to which groups. Cluster analysis is popular in many fields like cancer research for the group expression profile to identify the molecular profile of patients with good or bad. The business segmentation marketing by identifying sub-groups of customers with similar profiles that could be included in a particular form of business promotion. agglomerated hierarchical grouping is deterministic only for the bounded distances when a single link is not used. DBSCAN is deterministic, except for the data set permutation in rare cases. Kmeans is deterministic except for initialization. It is possible to initialize with the first k objects, so it is also deterministic. The variables in which the cluster analysis will be performed. The hypothesis to be tested and the judgment of the experimenter should also be selected by the theory. It is necessary to select an appropriate measure of distance or similarity; The most used measure is the Euclidean distance or its square.

References

[1] A. Ruiz, “The cognitive coder infoworld.” 2017.

[2] S. Kakati, “What is data science. attend coursera, udemy and stanford online lagunita.” 2018.

[3] G. Fiona, “Overfitting regression analysis,” www.bmj.com, 2018.

[4] A. Nasridinov, “The third international conference on ieee, visual analytics for big data using R.” In Cloud and Green Computing (CGC), 2013.

[5] R. Ihaka, “Mining big data: Current status andforecast to the future,” 1996.

[6] M. RIJMENAM, “How does big data analytics help in decision making. p. https://datafloq.com/about/.” 2018.

[7] M. Hahsler, “dbscan: Fast density-based clustering with r. volume wright state university.” Version Febu, 2013.

[8] J. Leskovec and Anand Rajaraman Data Mining, Stanford University 2018

[9] Joasang Lim1, Joongjin Kook2 and Jinman Kim DBSCAN-D: A Density-Based Clustering Method of Directionality, 2017

[10] N. Golubovic Centaurus: *A Cloud Service for K-means Clustering*, 2015

Author’s Biography

Yagyanath Rimal is lecturer of Computer Science Information Technology at Pokhara University, Nepal. He received his Master of Science in Information Technology from SMU, India with Information Technology specialization and programming, he had wide area of interest like teaching student, writing books, doing

How to Cite

Rimal, Yagyanath, “Deterministic Machine Learning Cluster Analysis of Research Data: using R Programming”, *International Journal of Machine Learning and Networked Collaborative Engineering,* Vol. 03, No. 1, 2019, pp. 46-65.

doi : https://doi.org/10.30991/IJMLNCE.2019v03i01.004.