**Mini Project Report**

**“Hierarchical Clustering”**

**Submitted by**

Mehul Patil (T184099)

Niranjan Patil (T184104)

Pranay Patil (T184115)

Prajwal Warad (T18423)

**Under Supervision of**

Prof. Jayvant H. Devare

Miss. Shobha Mourya



**School of Computer Engineering and Technology**

MIT ACADEMY OF ENGINEERING

December 2018

**Acknowledgments**

We would like to thank Miss. Shobha Mourya mam and Prof. Jayvant Devare sir for their support and advice through all the stages of project. Data Science is truly a wonderful subject and we have learnt the importance of this subject through this project. We enjoyed and learnt many new things through the course of this project.

December 2018

Prajwal Warad (T184123)

Niranjan Patil (Seat No)

Pranay Patil (Seat No)

Mehul Patil (Seat No)

**Contents**

**ACKNOWLEDGEMENTS** **1**

**CONTENTS** **2**

**1. INTRODUCTION** **3**

OBJECTIVE 4

METHODOLOGY 4

**2. DATA SET** **5**

DESCRIPTION 6

NUMBER OF ATTRIBUTES USED 6

**3. SYSTEM WORKING** **10**

SAMPLE CODE……………………………………………………………………………………………10

SCREENSHOT RESULT 6

**4. CONCLUSION** **12**

**5. FUTURE WORK**  **14**

**6. REFERENCES** **16**

**Introduction**

**Overview**

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense) to each other than to those in other groups (clusters). It is a main task of exploratory data mining, and a common technique for statistical data analysis, used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, bioinformatics, data compression, and computer graphics.

Cluster analysis itself is not one specific algorithm, but the general task to be solved. It can be achieved by various algorithms that differ significantly in their understanding of what constitutes a cluster and how to efficiently find them. Popular notions of clusters include groups with small distances between cluster members, dense areas of the data space, intervals or particular statistical distributions.

Cluster analysis is a powerful toolkit in the data science workbench. These similarities can inform all kinds of business decisions; for example, in marketing, it is used to identify distinct groups of customers for which advertisements can be tailored.

Cluster analysis seeks to find groups of observations that are similar to one another, but the identified groups are different from each other. This similarity/difference is captured by the metric called distance.

We will learn about the fundamental principles of hierarchical clustering - the linkage criteria and the dendrogram plot - and how both are used to build clusters.

**Objective**

* Learn how to calculate the distance between observations for both continuous and categorical features.
* Develop an intuition for how the scales of your features can affect distance.
* How do you find groups of similar observations (clusters) in your data using the distances that you have calculated?
* Learn about the fundamental principles of hierarchical clustering.

**Methodology**

* Scaling of two observations.
* Calculating distance between observations.
* Finding the closest observation to the pair.
* Comparing two or more observations.
* Produce dendrogram from the clusters.
* Put it with the main data and display clusters.

**Data set**

We have bundled the project based on two datasets.

1. Employee Payroll:

We have the list of names of employees, their occupation and their salary. So we have applied hierarchical clustering to create clusters based on similar occupation and similar payroll. We have used four attributes here.

No: of attribute used – 3,

Attributes: Field, year and salary

2. Iris flower dataset:

Here we have list of flower species and their characteristics like petal length, sepal length and width etc. So using hierarchical clustering we have clustered them to clusters based on which specie they belong to. We have used four attributes here as well.

No: of attribute used – 5,

Attributes: Sepal length, Sepal width, Petal length, Petal Width and species

**System Working**

* **First dataset code:**

**Imputation:**

# Load Packages

getwd()

install.packages("dplyr")

library(dplyr)

install.packages("ggplot2")

library(ggplot2)

install.packages("reader")

library(reader)

# Read Csv File

oes = read.csv("C:/Users/Mehul/Documents/oesnew2.csv")

glimpse(oes)

#load data

data <- oes

#Get summary

summary(oes)

install.packages("missForest")

library(missForest)

#Generate 10% missing values at Random

oes.mis <- prodNA(oes, noNA = 0.1)

iris.mis <- prodNA(iris,noNA=0.1)

#Check missing values introduced in the data

summary(oes.mis)

#iris.mis <- subset(oes.mis, select = -c(Species))

summary(oes.mis)

#install MICE

install.packages("mice")

library(mice)

md.pattern(oes.mis)

install.packages("VIM")

library(VIM)

mice\_plot <- aggr(oes.mis, col=c('navyblue','yellow'),

                  numbers=TRUE, sortVars=TRUE,

                  labels=names(oes.mis), cex.axis=.7,

                  gap=3, ylab=c("Missing data","Pattern"))

imputed\_Data <- mice(oes.mis, m=5, maxit = 50, method = 'pmm', seed = 500)

summary(imputed\_Data)

mice(data = oes.mis, m = 5, method = "pmm", maxit = 50, seed = 500)

#check imputed values

imputed\_Data$imp$Fields.Years

imputed\_Data$imp$X2001

imputed\_Data$imp$X2007

imputed\_Data$imp$X2003

imputed\_Data$imp$X2002

**Main clustering:**

# Load Packages

getwd()

install.packages("dplyr")

library(dplyr)

install.packages("ggplot2")

library(ggplot2)

install.packages("reader")

library(reader)

# Read Csv File

oes = read.csv("C:/Users/Admin/Desktop/Study/T.Y.B.Tech/Semester Long/R Programming/Project/Occupational Employment Statistics.csv")

glimpse(oes)

library(dendextend)

# Calculate euclidean distance between the occupations

dist\_oes <- dist(oes, method = "euclidean")

# Generate an average linkage analysis

hc\_oes <- hclust(dist\_oes, method = "average")

# Create a dendrogram object from the hclust variable

dend\_oes <- as.dendrogram(hc\_oes)

# Plot the dendrogram

plot(dend\_oes)

# Color brances by cluster formed from the cut at a height of 100000

dend\_colored <- color\_branches(dend\_oes, h = 100000)

# Plot the colored dendrogram

plot(dend\_colored)

library(tibble)

library(tidyr)

# Use rownames\_to\_column to move the rownames into a column of the data frame

df\_oes <- rownames\_to\_column(as.data.frame(oes), var = 'occupation')

# Create a cluster assignment vector at h = 100,000

cut\_oes <- cutree(hc\_oes, h = 100000)

# Generate the segmented the oes dataframe

clust\_oes <- mutate(df\_oes, cluster = cut\_oes)

# Create a tidy data frame by gathering the year and values into two columns

gathered\_oes <- gather(data = clust\_oes,

                       key = year,

                       value = mean\_salary,

                       -cluster, -occupation)

head(gathered\_oes)

**Screenshots:**

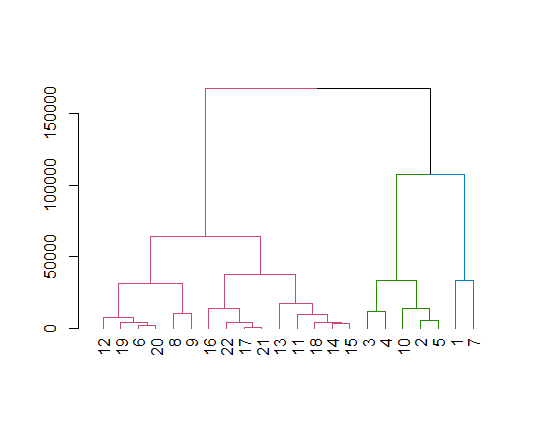


Figure 1 Dendogram

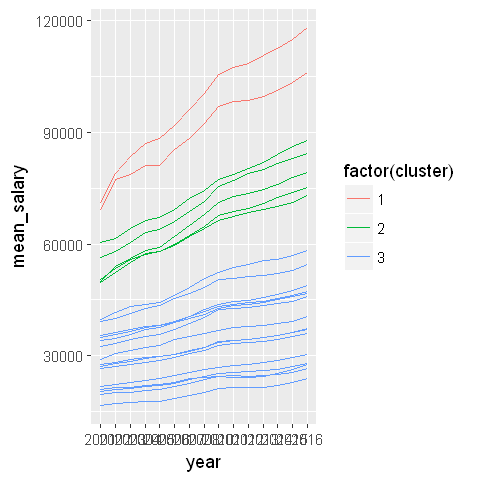


Figure 2 Graph

* **Second dataset code**

install.packages("dplyr")

install.packages("ggplot2")

install.packages("stringr")

install.packages('gdata')

install.packages("colorspace")

install.packages("dendextend")

library(dplyr)

library(ggplot2)

library(stringr)

library(gdata)

library(colorspace)

library(dendextend)

set.iris\_dataworld\_test()

library(gdata)

file\_name=read.csv("Iris flower data set.csv")

getwd()

summary(file\_name)

is.na(file\_name)

str(file\_name)

library(colorspace)

flower\_frame=as.data.frame(scale(file\_name))

summary(file\_name)

exceptional\_label=file\_name$Species

file\_name$Species=NULL

summary(file\_name)

summary(flower\_frame)

head(flower\_frame)

dist\_mat=dist(flower\_frame, method='euclidean')

hclust\_avg=hclust(dist\_mat, method='average')

hclust\_avg

plot(hclust\_avg)

cut\_avg=cutree(hclust\_avg, k=3)

rect.hclust(hclust\_avg , k = 3, border = 2:6)

abline(h = 3, col = 'red')

suppressPackageStartupMessages(library(dendextend))

avg\_dend\_obj <- as.dendrogram(hclust\_avg)

avg\_col\_dend <- color\_branches(avg\_dend\_obj, h = 3)

plot(avg\_col\_dend)

flower\_cl=mutate(file\_name, cluster=cut\_avg)

suppressPackageStartupMessages(library(dplyr))

count(flower\_cl, cluster)

suppressPackageStartupMessages(library(ggplot2))

ggplot(flower\_cl, aes(x=Dataset.Order, y=Sepal.length, color=factor(cluster)))+geom\_point()

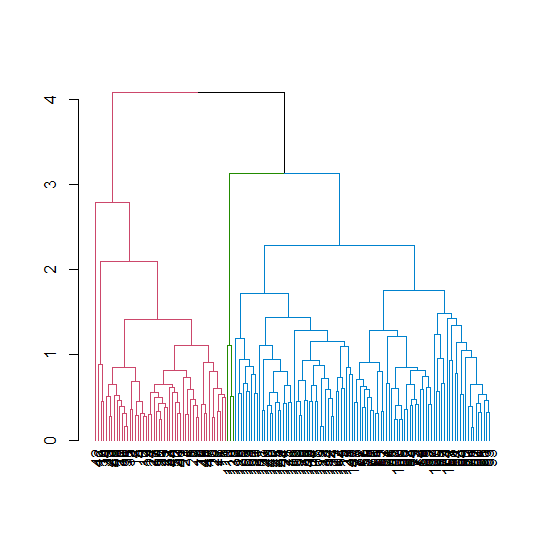
ggplot(flower\_cl, aes(x=Dataset.Order, y=Sepal.width, color=factor(cluster)))+geom\_point()

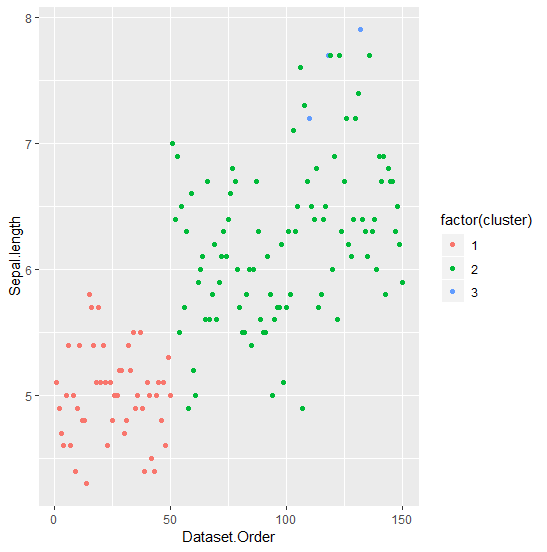
ggplot(flower\_cl, aes(x=Dataset.Order, y=Petal.length, color=factor(cluster)))+geom\_point()

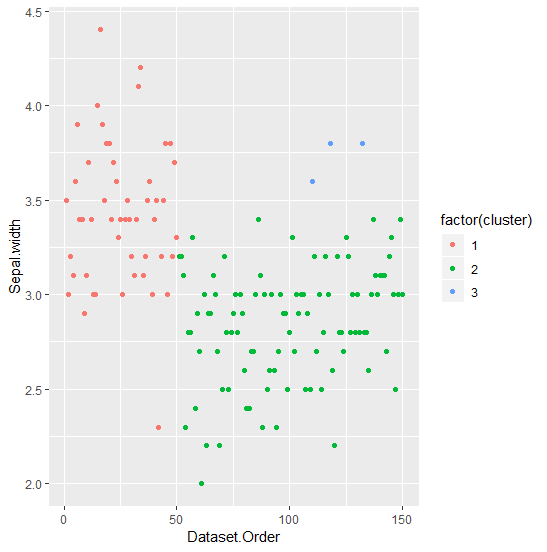
ggplot(flower\_cl, aes(x=Dataset.Order, y=Petal.width, color=factor(cluster)))+geom\_point()

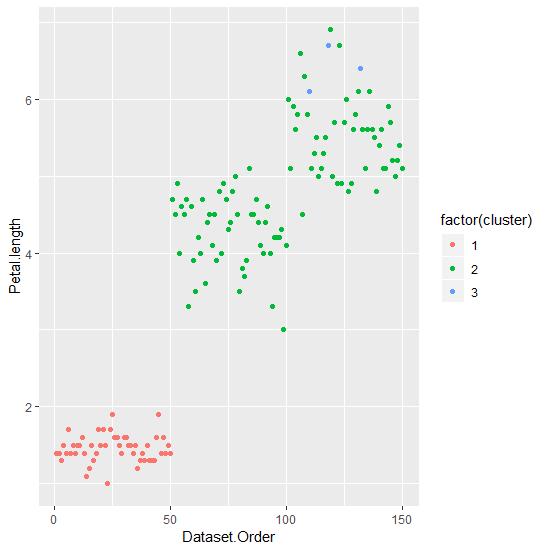
table(flower\_cl$cluster, exceptional\_label)

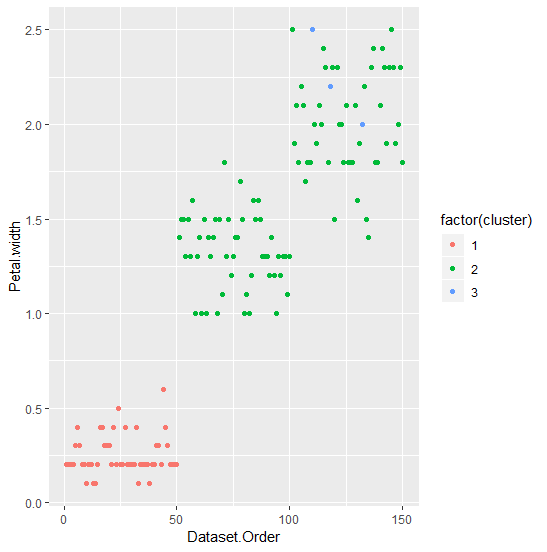
**Screenshots:**

****

****

****

****

****

**Conclusion & Future work**

Through the project we have acquired knowledge about clustering and its types and how to implement it into our day today topics.

We can use k-clustering method as well on these datasets. We are going to do that in the future. In this project we have used average linkage to plot distance. There are three other types of linkages that we can use to plot.

**Reference**

* <https://archive.ics.uci.edu/ml/datasets/iris>
* <https://cran.r-project.org/web/packages/dendextend/vignettes/Cluster_Analysis.html>
* <http://www.nosimpler.me/hierarchical-clustering/>