SCHOOL OF MEDICINE
UNIVERSITY OF THESSALY

POSTGRADUATE PROGRAMME (MSC)

«RESEARCH METHODOLOGY IN BIOMEDICINE, BIOSTATISTICS AND CLINICAL BIOINFORMATICS AT UNIVERSITY OF THESSALY»

Master's Thesis

"Comparative Analysis of the Different Clustering Methods using Python"

"Συγκριτική Μελέτη Διαφορετικών Μεθόδων Ομαδοποίησης με χρήση Python"

Supervisors
Batsidis Apostolos
Axel Kowald
Elias Zintzaras

Dadouli Katerina
AEM: 00070
E-mail: katerina1dad@gmail.com
Academic year:2015-2016

“Comparative Analysis of the Different Clustering Methods using Python”

Chapter 1 Abstract

The purpose of cluster analysis is to assign items in groups ("clusters"), so that items belonging to the same cluster are more similar than those items belonging to different clusters. Because of the different ways determining the clusters, there are many different techniques of cluster analysis. In this master's thesis three popular methods of cluster analysis are presented and were compared using Python. In this context, Chapter 2 (Introduction) introduces the concept of Cluster Analysis. In Chapter 3 (Methods), a brief introduction of the three well known clustering algorithms used in this thesis is presented. Moreover the advantages and disadvantages of these methods based on empirical studies are given. In Chapter 4 (Results), the experimental part of the three methods of clustering using a specific data set and the software Python is given. Chapter 5 (Conclusion) gives a summary.

Chapter 2 Introduction

Cluster analysis is a multivariate method and the objective of it, is to assign observations to groups (“clusters") so that observations within each group are similar to one another with respect to variables or attributes of interest, and the groups them-selves stand apart from one another(see for instance Peter Tryfos, 1997). In other words, the groups or clusters should be as homogeneous as possible and the differences among the various groups as large as possible. Cluster analysis is applied in many fields such as the natural sciences, the medical sciences, economics, marketing, etc. In marketing, for instance, it is useful to build and describe the different segments of a market from a survey on potential consumers. An insurance company, on the other hand, might be interested in the distinction among classes of potential customers so that it can derive optimal prices for its services. In medical sciences, specifically in the field of psychiatry, based on scores on psychological inventories, patients can be clustered into subgroups that have similar response patterns. This may be helpful in targeting appropriate treatment and studying typologies of diseases.

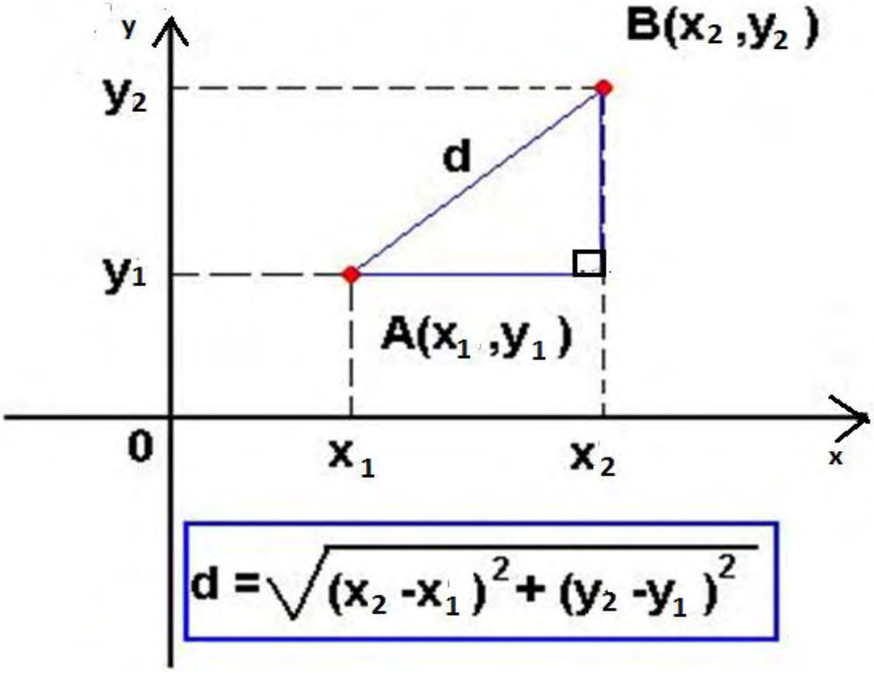
Cluster analysis can be divided into the following fundamental steps (see for instance Hardler and Simar, 2007)

1 .Formulate the problem

Most important is selecting the variables on which the clustering is based. Inclusion of even one or two irrelevant variables may distort a clustering solution. Cluster analysis has no mechanism for differentiating between relevant and irrelevant variables. Therefore the choice of variables included in a cluster analysis must be underpinned by conceptual considerations. This is very important because the clusters formed can be very dependent on the variables included. (See for instance Cornish 2007)

2. Choice of a proximity measure-Select a Distance Measure

The starting point of a cluster analysis is a data matrix X (n × p) with n measurements (objects) on p variables. The proximity (similarity) among objects is described by a matrix D (n × n), whose components d_{ij} give the similarity coefficient or the distance between two points x_i and x_j. The nature of the observations plays an important role in the choice of proximity measure. The data used in cluster analysis can be interval, ordinal or categorical. A variety of similarity (distance) measures exist for binary data (e.g., Jaccard, Tanimoto, Simple Matching coefficients) and for continuous data (e.g., Lr-norms). For interval data the most common distance measure used is the Euclidean distance. The Euclidean distance between the two points is the hypotenuse of the triangle as shown in the next graph:



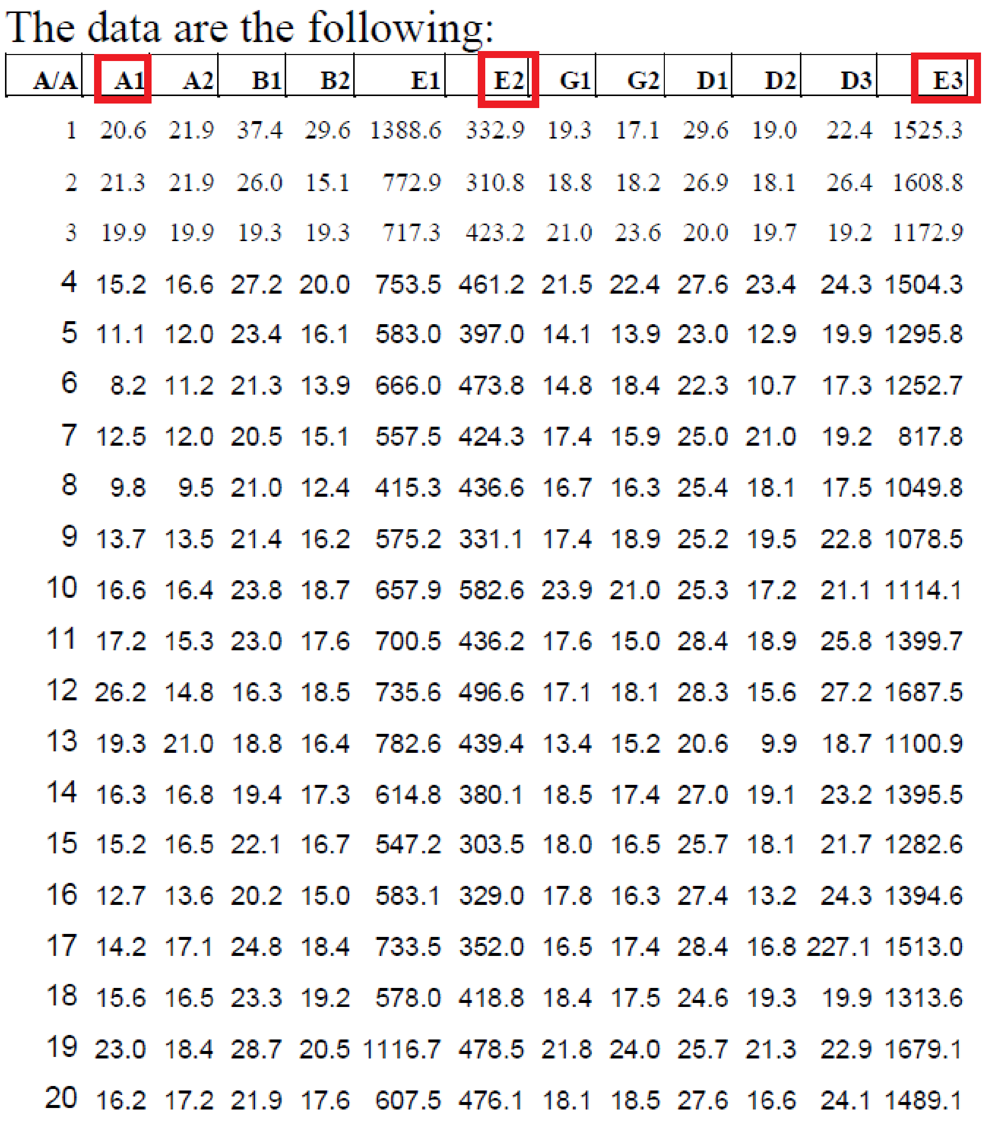
As far as "standardizing" is concerned, when variables are measured on different scales, (e.g. a variable that ranges between 0 and 100 and a variable that ranges between 0 and 1) then it is recommended to standardizing variables since variables with large values contribute more to the distance measure than variables with small values.

Especially our code uses "normalization" of data. To normalize data, traditionally means to fit all the data within unity (1), so all data values will take on a value of 0 to 1. The following equation is what should be used to implement a unity-based normalization:

$$x_{i,0\ to\ 1} = \frac{x_i - x_{min}}{x_{max} - x_{min}}$$

where:
x_i*=if the value of each data point i*
x_{min} *=the minima among all data points* x_i
x_{max} *=the maxima among all data points* x_i
$x_{i,0\ to\ 1}$*=the data point which is normalized between 0 and 1.*
(See for instance: Ben Etzkorn, 2011)

For example, we can see in the following table 12 morphometric measurements of sacrum and we must identify groups of individuals with similar morphological sacrum. As we can see, A1 variable ranges from 9.8mm to 26.2mm(length), E2 ranges from 310.8mm to 582.6mm (length), and E3 ranges from 817.8mm^2 to 1687.5mm^2 (area). So in this case standardization is recommended because without it the variable A1 has no chance of influencing in solution given that the other variables are much larger.



The data are the following:

A/A	A1	A2	B1	B2	E1	E2	G1	G2	D1	D2	D3	E3
1	20.6	21.9	37.4	29.6	1388.6	332.9	19.3	17.1	29.6	19.0	22.4	1525.3
2	21.3	21.9	26.0	15.1	772.9	310.8	18.8	18.2	26.9	18.1	26.4	1608.8
3	19.9	19.9	19.3	19.3	717.3	423.2	21.0	23.6	20.0	19.7	19.2	1172.9
4	15.2	16.6	27.2	20.0	753.5	461.2	21.5	22.4	27.6	23.4	24.3	1504.3
5	11.1	12.0	23.4	16.1	583.0	397.0	14.1	13.9	23.0	12.9	19.9	1295.8
6	8.2	11.2	21.3	13.9	666.0	473.8	14.8	18.4	22.3	10.7	17.3	1252.7
7	12.5	12.0	20.5	15.1	557.5	424.3	17.4	15.9	25.0	21.0	19.2	817.8
8	9.8	9.5	21.0	12.4	415.3	436.6	16.7	16.3	25.4	18.1	17.5	1049.8
9	13.7	13.5	21.4	16.2	575.2	331.1	17.4	18.9	25.2	19.5	22.8	1078.5
10	16.6	16.4	23.8	18.7	657.9	582.6	23.9	21.0	25.3	17.2	21.1	1114.1
11	17.2	15.3	23.0	17.6	700.5	436.2	17.6	15.0	28.4	18.9	25.8	1399.7
12	26.2	14.8	16.3	18.5	735.6	496.6	17.1	18.1	28.3	15.6	27.2	1687.5
13	19.3	21.0	18.8	16.4	782.6	439.4	13.4	15.2	20.6	9.9	18.7	1100.9
14	16.3	16.8	19.4	17.3	614.8	380.1	18.5	17.4	27.0	19.1	23.2	1395.5
15	15.2	16.5	22.1	16.7	547.2	303.5	18.0	16.5	25.7	18.1	21.7	1282.6
16	12.7	13.6	20.2	15.0	583.1	329.0	17.8	16.3	27.4	13.2	24.3	1394.6
17	14.2	17.1	24.8	18.4	733.5	352.0	16.5	17.4	28.4	16.8	227.1	1513.0
18	15.6	16.5	23.3	19.2	578.0	418.8	18.4	17.5	24.6	19.3	19.9	1313.6
19	23.0	18.4	28.7	20.5	1116.7	478.5	21.8	24.0	25.7	21.3	22.9	1679.1
20	16.2	17.2	21.9	17.6	607.5	476.1	18.1	18.5	27.6	16.6	24.1	1489.1

(Source: Zintzaras, 2016)

3. Choice of group-building algorithm- select a clustering procedure

On the basis of the proximity measures the objects assigned to groups so that differences between groups become large and observations in a group become as close as possible. There are essentially two types of clustering methods: Hierarchical and Non-hierarchical (or partioning).

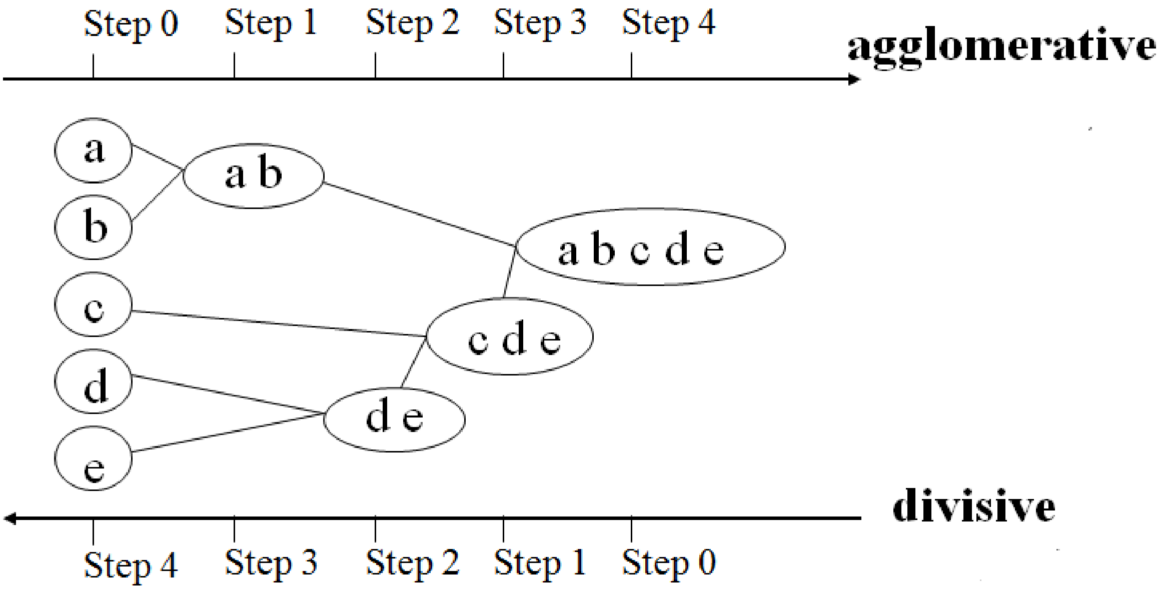
***Hierarchical methods**. Hierarchical clustering is one of the most straightforward methods. It can be either agglomerative or divisive.*

*– **Agglomerative methods**, in which subjects start in their own separate cluster. The two 'closest' (most similar) clusters are then combined and this is done repeatedly until all subjects are in one cluster. At the end, the optimum number of clusters is then chosen out of all cluster solutions.*

*– **Divisive methods**, in which all subjects start in the same cluster and the above strategy is applied in reverse until every subject is in a separate cluster.*

(See for instance Rosie Cornish 2007)

This is shown, in the following graph.



Agglomerative methods are used more often than divisive methods, so in this thesis agglomerative methods will be used.

The agglomerative algorithm consists of the following steps:

```
1.        Construct the finest partition
2.        Compute the distance matrix D.
DO
3.        Find the two clusters with the closest distance.
4.        Put those two clusters into one cluster.
5.        Compute the distance between the new groups and obtain a reduced distance matrix D.
UNTIL all clusters are agglomerated into X.
```

(See for instance Wolfgang Härdle and Léopold Simar 2007)

***Non-hierarchical methods** (often known as k-means clustering methods)*

The partioning algorithms start from a given group definition and proceed by exchanging elements between groups until a certain score is optimized.

The main difference between the two clustering techniques is that in hierarchical clustering once groups are found and elements are assigned to the groups, this assignment cannot be changed. In partitioning techniques, on the other hand, the assignment of objects into groups may change during the algorithm application.

In agglomerative clustering, once a cluster is formed, it cannot be split; it can only be combined with other clusters. Agglomerative hierarchical clustering doesn't let cases separate from clusters that *they've joined. Once in a cluster, always in that cluster (see for instance Norusis, 2016).*

4. Number of clusters

There is no right or wrong answer as to how many clusters someone needs. It depends on what you're going to do with them. To find a good cluster solution, you must look at the characteristics of the clusters at successive steps and decide when you have an interpretable solution or a solution that has a reasonable number of fairly homogeneous clusters (see for instance Norusis, 2016).

A good clustering solution should have high intra-cluster similarity and low inter-cluster similarity, i.e., data points within the same cluster should be similar but are dissimilar to data points in other clusters(see for instance Benjamin C. M. Fung, Ke Wang, and Martin Ester, Simon, 2009).

Based on the previous discussion it is obvious that there are a number of different methods used to determine which clusters should be joined at each stage. In this thesis three methods, namely the nearest neighbor method (single linkage method), the furthest neighbor method (complete linkage method) and the average linkage method will be considered. The presentation of those three methods is the subject of the next section.

Chapter 3 Methods

There are a number of different methods used to determine which clusters should be joined at each stage. In this thesis the following well known methods will be considered.

Method 1: Nearest neighbor method (single linkage method)

In this method the distance between two clusters is defined to be the minimum distance between the two closest members, or neighbors. This method is relatively simple but is often criticized because it doesn't take into account the cluster structure .This last characteristic can result in a problem called chaining whereby clusters end up being long and straggly. However, it is better than the other methods when the natural clusters are not spherical or elliptical in shape (see for instance Rosie Cornish 2007)

The distance between an individual (k) and a group formed by individuals i and j is calculated as follows:

$$d_{(ij)k} = \min\{d_{ik}, d_{jk}\}.$$

Therefore, $d_{(ij)k}$is the smallest element of the set of the distances of the pairs of the individuals ((i and k) and (j and k)). The connections between individuals and groups or among groups are made by the distance between the groups defined as that between the most similar individuals in these groups.

The distance between two groups is calculated as follows:

$$d_{(ij)(kl)} = \min\{d_{ik}, d_{il}, d_{jk}, d_{jl}\}.$$

The distance between the two groups formed by individuals, (i and j) and (k and l), is the smallest element of the set of which the elements are the distances among the pairs of individuals ((i and k), (i and l), (j and k) and (j and l).

(See for instance Luiza Barbosa da Matta, L&#237;via Gracielle Oliveira Tom&#233;, Caio C&#233;sio Salgado , Cosme Dami&#227;o Cruz, Let&#237; cia de Faria Silva (2015)).

Method 2: Furthest neighbor method (complete linkage method)

In this case the distance between two clusters is defined to be the maximum distance between members — i.e. the distance between the two subjects that are furthest apart. This method tends to produce compact clusters of similar size but, as for the nearest neighbor method, does not take into account the cluster structure. It is also quite sensitive to outliers (see for instance Rosie Cornish 2007).

In other words, complete linkage distance is the converse of single linkage since the least similar pair of documents in two clusters forms the basis for the measurement of inter-cluster similarity: thus, each document in a cluster is more similar to the most dissimilar in that cluster than to the most dissimilar document in any other cluster. This definition of cluster membership is very much stricter than that for single linkage, and thus the large straggly clusters in the latter case are here replaced by large numbers of small, tightly bound clusterings. (See for instance A.El-Hamdouchi and P.Willett 1989)

The distance between an individual (k) and a group consisting of individuals i and j is calculated as follows:

$$d_{(ij)k}=\max\{d_{ik},d_{jk}\}.$$

Therefore d(ij)k is the highest element in the set of individual pairs ((i and k) and (j and k)).

The distance between two sets is calculated as follows:

$$d_{(ij)(kl)}=\max\{d_{ik},d_{il},d_{jk},d_{jl}\}.$$

The distance between two groups formed by the individuals ((i and j) and (k and l)) is determined by the highest element in the set of which the elements are the distances among the pairs of the individuals in the clusters ((i and k), (i and l), (j and k) and (j and l))(see for instance Luiza Barbosa da Matta, L&#237;via Gracielle Oliveira Tom&#233;, Caio C&#233;sio Salgado , Cosme Dami&#227;o Cruz, Let&#237; cia de Faria Silva (2015).)

Method 3: Average (between groups) linkage method (sometimes referred to as UPGMA = unweighted pair-group method using arithmetic averages (UPGMA))

The distance between two clusters is calculated as the average distance between all pairs of subjects in the two clusters. This is considered to be a fairly robust method (see for instance Rosie Cornish 2007)

In other words, this method represents a mid-point between the two extreme types of linkage method, i.e. single linkage and complete linkage method.(see for instance A.El-Hamdouchi and P.Willett, 1989)

An overall expression for the unweighted mean among groups can be represented as follows:

$$d_{(ij)k}=\frac{n_i}{n_i+n_j}d_{ik}+\frac{n_j}{n_i+n_j}d_{jk}$$

Therefore $d_{(ij)(k)}$ is defined as the distance between the group (ij) with inner size n_i and n_j, respectively, and group k. In this equation, the i, j and k indices are characterized as individuals or *groups (see for instance Luiza Barbosa da Matta, L&#237;via Gracielle Oliveira Tom&#233;, Caio C&#233;sio Salgado , Cosme Dami&#227;o Cruz, Let&#237;cia de Faria Silva (2015)).*

The next table contains the most commonly used agglomerative algorithms and cluster criteria for cluster development. For each one of them some weak and strong points (advantages and disadvantages) are reported.

Agglomerative Algorithm	*Cluster Criterion*	*Weak Point*	*Strong Point*
Single linkage or nearest neighbor approach.	*Minimum distance or the closest single pair.*	*“Snaking” effect*	*1. “Correctedness” maximization of pair of clusters* *2. Fewer clusters than other methods.*
Complete linkage or the furthest neighbor or diameter method.	*Maximum distance or the distance between their two furthest members.*	*Refers to a single pair.*	*1. Intra-cluster distances minimization* *2. Compact cluster formation*
Average linkage	*Average distance from all individuals in one cluster to all individuals in another.*	*Cluster production with approximately the same variance.*	*Produced hierarchy is the same with the single or complete linkage algorithm.*

Source:From Dillon,W.R. and Goldstein, M., Multivariate Analysis, New York: 1984; Hair,J.F.,Anderson,R.E.,Tatham, R.L., and Black, W.C., Multivariate Data Analysis, Upper Saddle River, NJ: Prentice Hall, 1998; Johnson, D.E., Applied Multivariate Methods for Data Analysis, Pacific Grove, CA: Duxbury Press, 1998.

<u>Clarifications</u>

“Snaking” or “Chaining” effect is the tendency to form loosely bound clusters with little internal cohesion.(See for instance A.El. Hamdouchi and P.Willett, 1989)

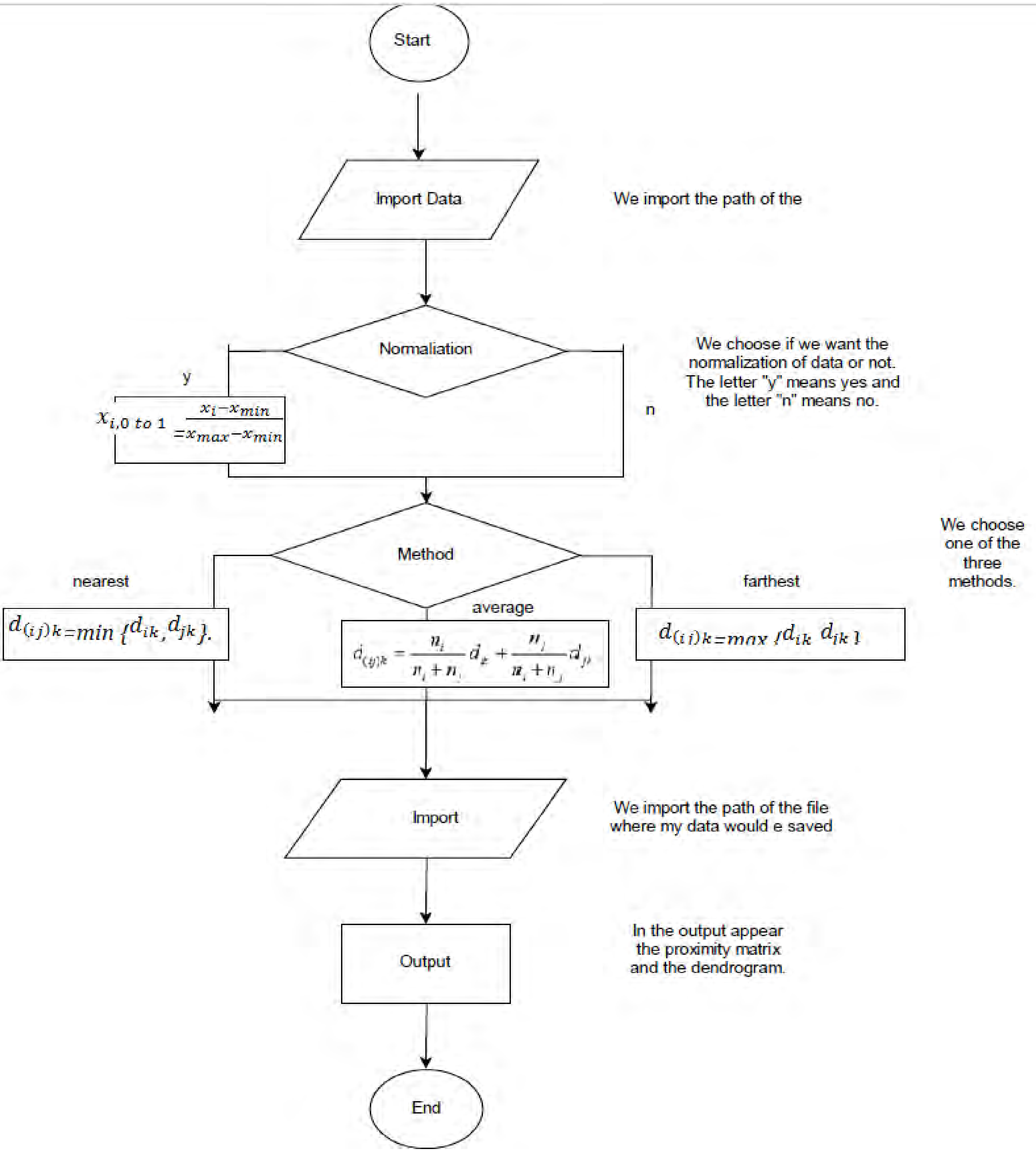
“Noise” and “outliers” are already mentioned many times in this thesis. So what they are? And which is the difference between them? Outliers are data objects with characteristics that are considerably different than most of the other data objects in the data set. Noise is random error or variance in a measured variable (Jiawei Han, Micheline Kamber, Jian Pei, 2012). In other words noise refers to modification of original values (Bhavesh Patankar, Dr. Vijay Chavda 2015). Noisy data is meaningless data. The term has often been used as a synonym for corrupt data (L. Sunitha, M.Bal Raju, B.Sunil Srinivas, 2013). Let’s assume that the values of SBP of six men are 150,147,180,145,280. If the nurse doesn’t enter right the data on database and she writes for example 150, 174, 180, 154, abc, 280 then she will create noise. Noise also could be created if the sphygmomanometer is not working properly. Finally it is easy to understand that the outlier in this example is the value 280.

Chapter 4 Description of the code.

The code uses the Euclidian distance to find the "distance" between patients. Moreover the code contains a choice about the "normalization" of variables or not, i.e. the user can choose if the variables should be "normalized" or not. The program also gives the opportunity to choose clustering method. There are three choices/clustering methods, nearest neighbor, farthest neighbor and average method. Finally, in the output appear the proximity matrix and the dendrogram.

Finally, there is the flowchart and an example which are useful in understanding code.

Flowchart

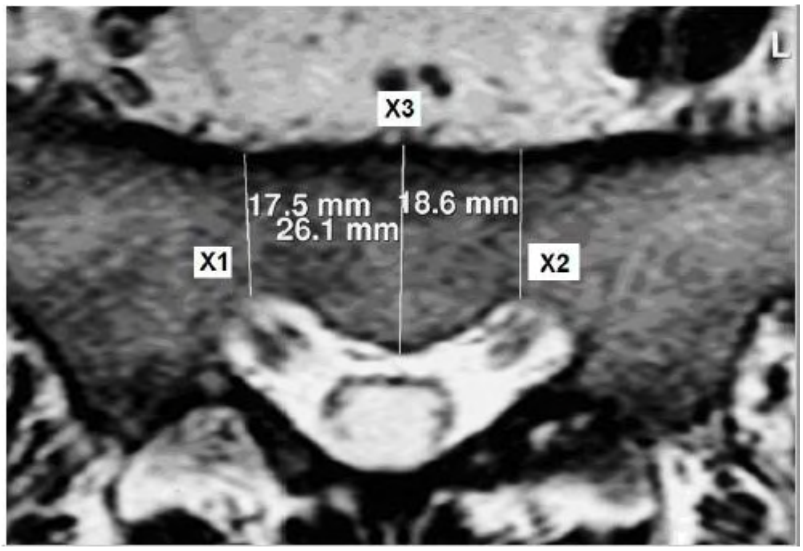

Start
Import Data
We import the path of the
Normaliation
We choose if we want the normalization of data or not. The letter "y" means yes and the letter "n" means no.
y
n
$x_{i,0\ to\ 1}=\frac{x_i-x_{min}}{x_{max}-x_{min}}$
Method
We choose one of the three methods.
nearest
average
farthest
$d_{(ij)k}=min\{d_{ik},d_{jk}\}.$
$d_{(ij)k}=\frac{n_i}{n_i+n_j}d_{ik}+\frac{n_j}{n_i+n_j}d_{jk}$
$d_{(ij)k}=max\{d_{ik}\ d_{jk}\}$
Import
We import the path of the file where my data would e saved
Output
In the output appear the proximity matrix and the dendrogram.
End

Example

We are interested in exploring groups of patients with similar morphometric profile in the sacrum.

The identification of these groups may direct us in choosing the appropriate surgical technique in patients with pelvic ring injuries.

In five patients with pelvic ring injuries, the following morphological variables of the transverse plane of the sacrum were recorded:



The dataset is the following:

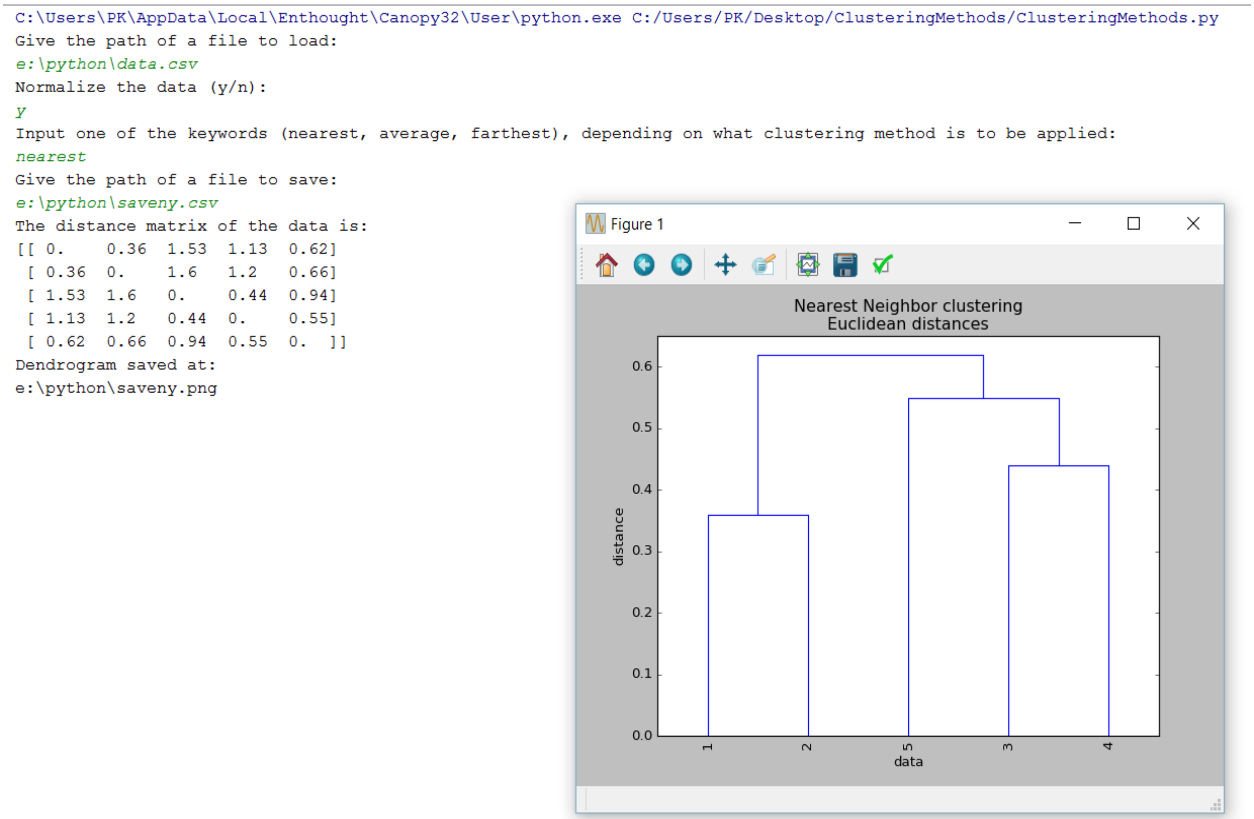
Patient	*X1*	*X2*	*X3*
1	*x11=22*	*x12=21*	*x13=28*
2	*x21=20*	*x22=22*	*x23=30*
3	*x31=14*	*x23=15*	*x33=21*
4	*x41=16*	*x24=16*	*x43=24*
5	*x51=18*	*x25=19*	*x53=26*

(Source: Zintzaras , 2016)

As we have already mentioned, firstly we import the path of the file. The data points are read from a data file (csv file) and they don't have to be typed in. At this point the file contains the example's data. It is necessary to mention that we can have a flexible number of dimensions i.e. the program can also analyze data with more than three variables. Afterwards we choose if we want the normalization of data or not. The letter "y" means "yes" and the "n" means "no". The third step is to choose one of the three clustering methods and finally we import the path of the file where the data will be saved.

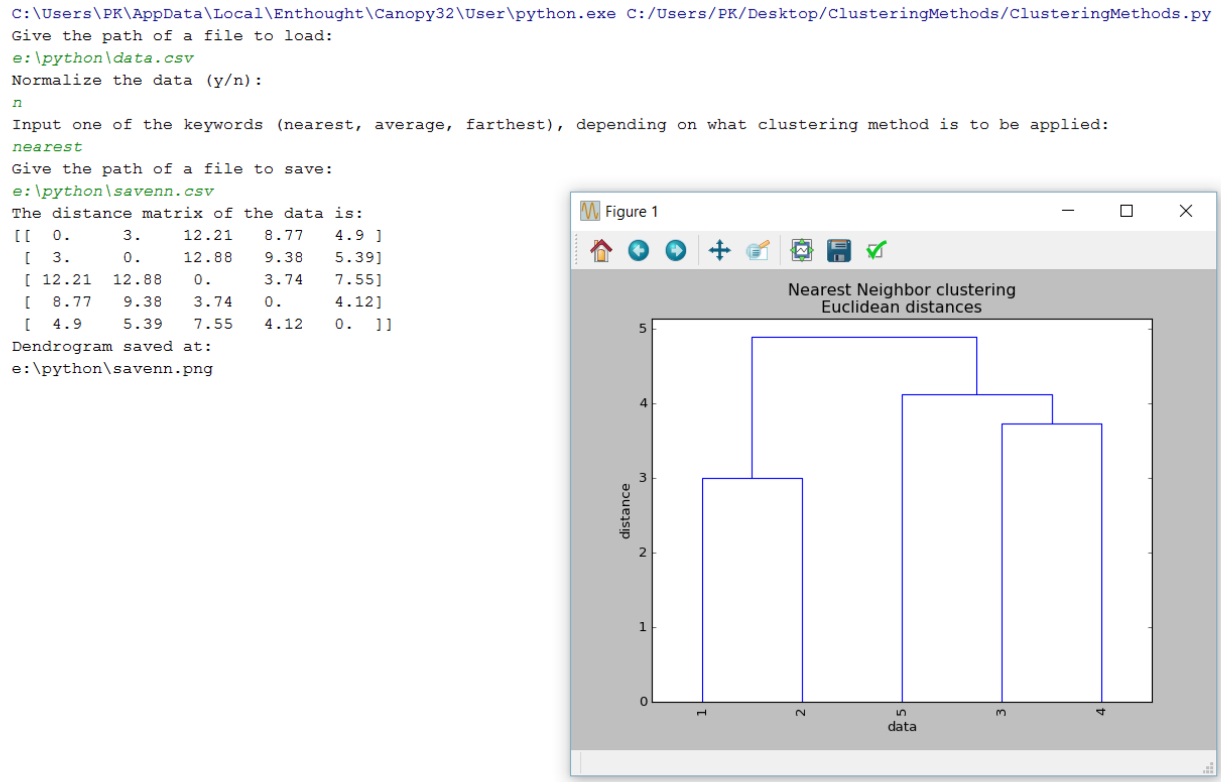
All possible results are:

Method: Nearest neighbor analysis
Normalization: yes



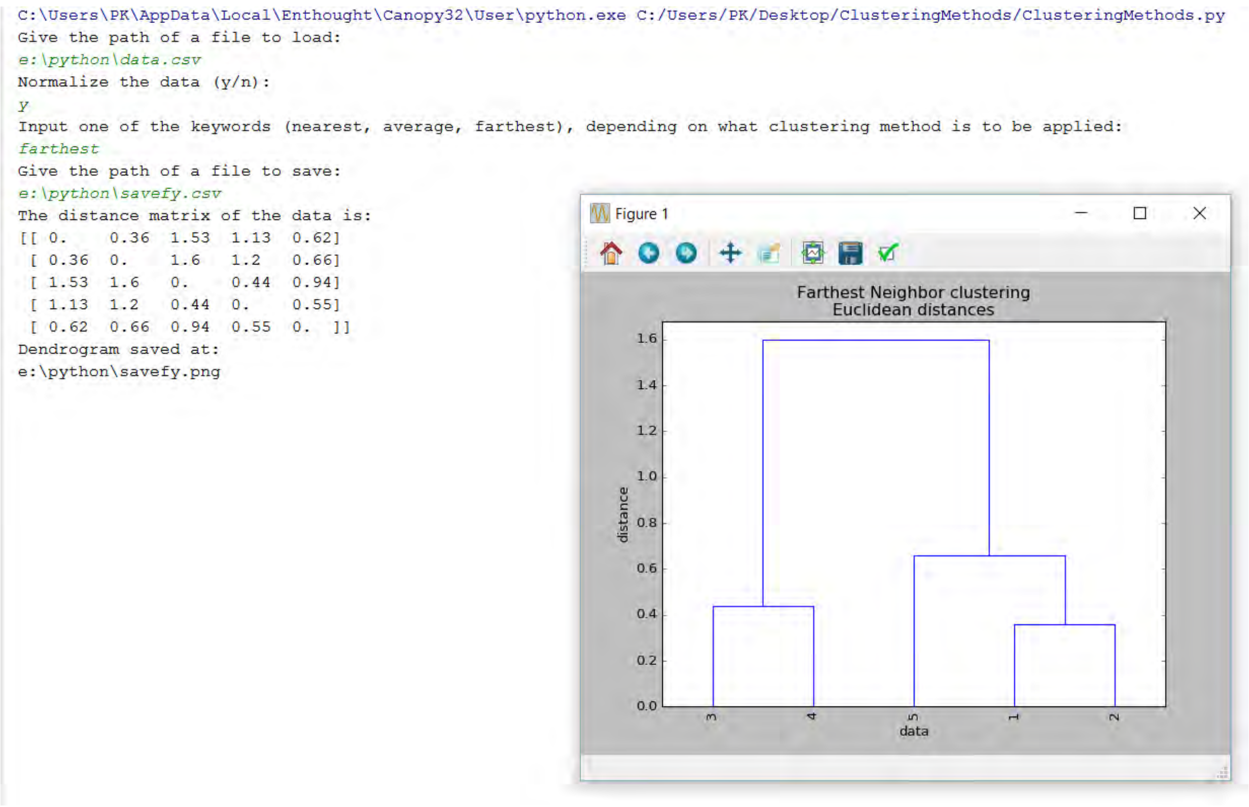
```
C:\Users\PK\AppData\Local\Enthought\Canopy32\User\python.exe C:/Users/PK/Desktop/ClusteringMethods/ClusteringMethods.py
Give the path of a file to load:
e:\python\data.csv
Normalize the data (y/n):
y
Input one of the keywords (nearest, average, farthest), depending on what clustering method is to be applied:
nearest
Give the path of a file to save:
e:\python\saveny.csv
The distance matrix of the data is:
[[ 0.    0.36  1.53  1.13  0.62]
 [ 0.36  0.    1.6   1.2   0.66]
 [ 1.53  1.6   0.    0.44  0.94]
 [ 1.13  1.2   0.44  0.    0.55]
 [ 0.62  0.66  0.94  0.55  0.  ]]
Dendrogram saved at:
e:\python\saveny.png
```

Method: Nearest neighbor analysis
Normalization: no



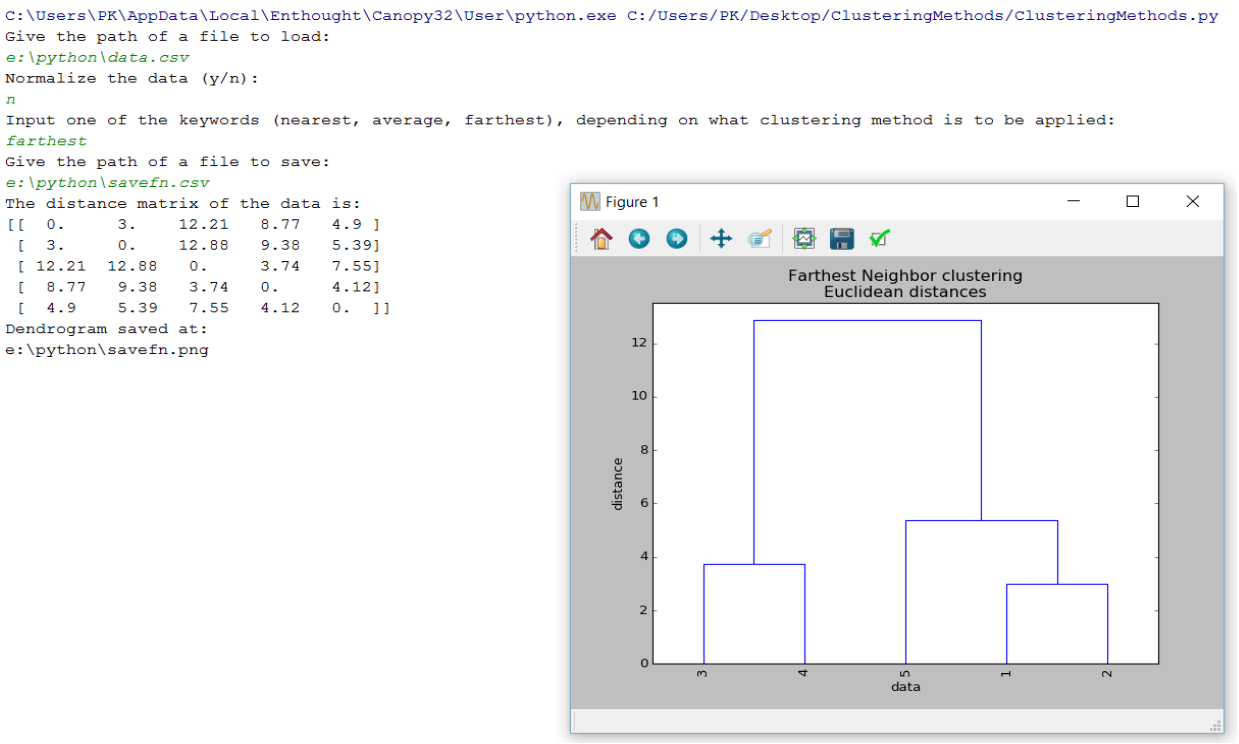
```
C:\Users\PK\AppData\Local\Enthought\Canopy32\User\python.exe C:/Users/PK/Desktop/ClusteringMethods/ClusteringMethods.py
Give the path of a file to load:
e:\python\data.csv
Normalize the data (y/n):
n
Input one of the keywords (nearest, average, farthest), depending on what clustering method is to be applied:
nearest
Give the path of a file to save:
e:\python\savenn.csv
The distance matrix of the data is:
[[  0.     3.    12.21   8.77   4.9 ]
 [  3.     0.    12.88   9.38   5.39]
 [ 12.21  12.88   0.     3.74   7.55]
 [  8.77   9.38   3.74   0.     4.12]
 [  4.9    5.39   7.55   4.12   0.  ]]
Dendrogram saved at:
e:\python\savenn.png
```

Method: Farthest neighbor analysis
Normalization: yes



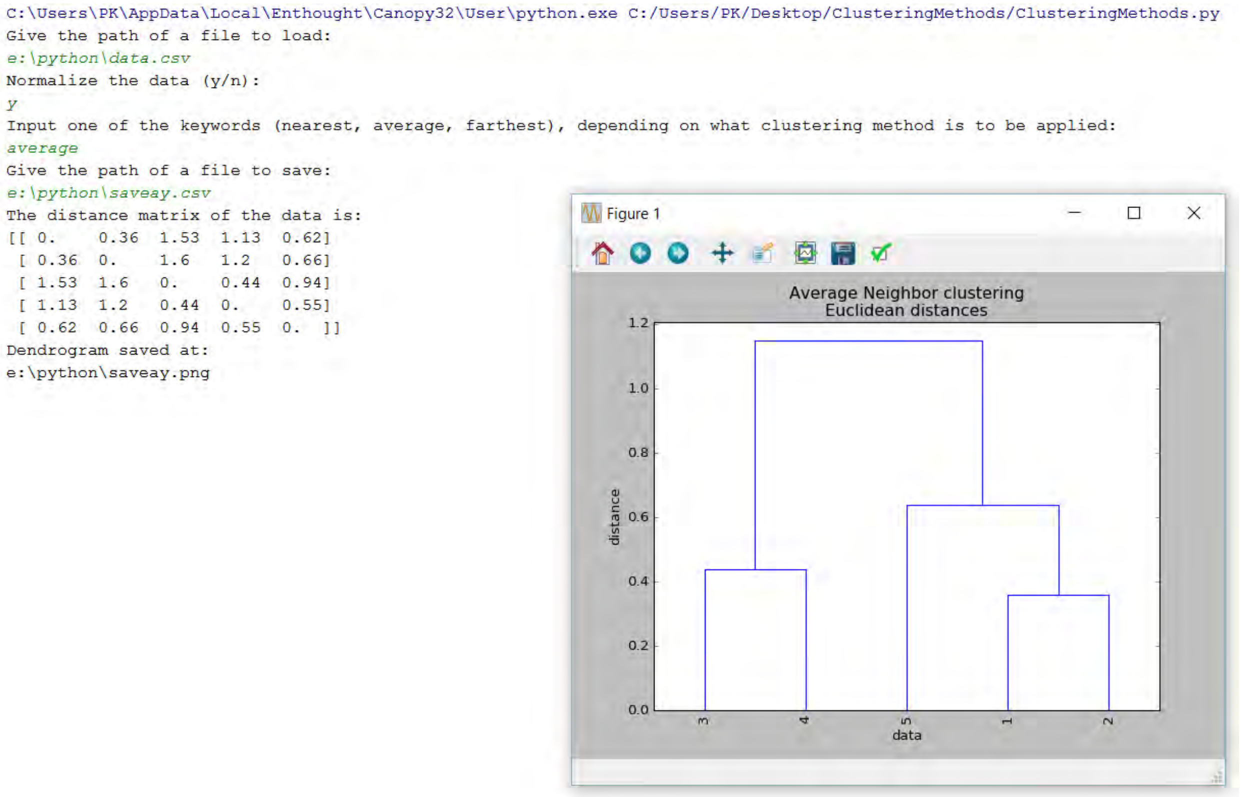
```
C:\Users\PK\AppData\Local\Enthought\Canopy32\User\python.exe C:/Users/PK/Desktop/ClusteringMethods/ClusteringMethods.py
Give the path of a file to load:
e:\python\data.csv
Normalize the data (y/n):
y
Input one of the keywords (nearest, average, farthest), depending on what clustering method is to be applied:
farthest
Give the path of a file to save:
e:\python\savefy.csv
The distance matrix of the data is:
[[ 0.    0.36  1.53  1.13  0.62]
 [ 0.36  0.    1.6   1.2   0.66]
 [ 1.53  1.6   0.    0.44  0.94]
 [ 1.13  1.2   0.44  0.    0.55]
 [ 0.62  0.66  0.94  0.55  0.  ]]
Dendrogram saved at:
e:\python\savefy.png
```

Method: Farthest neighbor analysis
Normalization: no



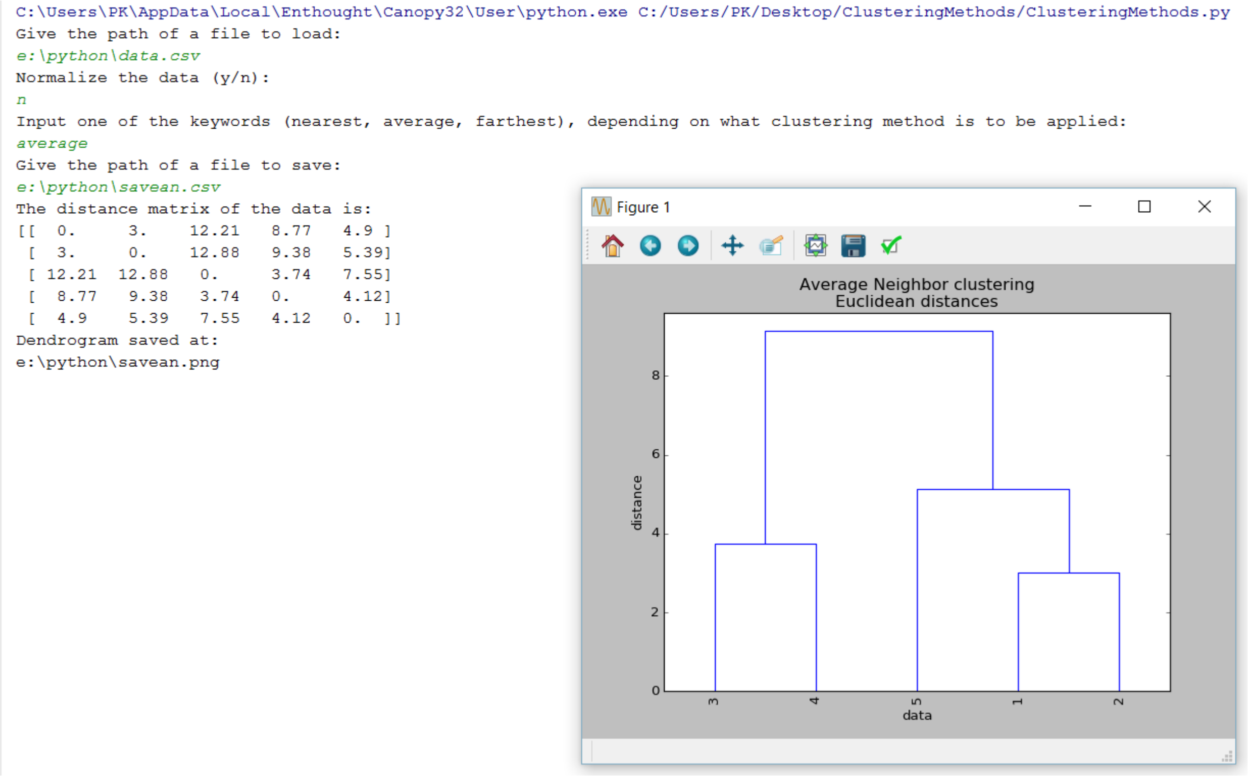
```
C:\Users\PK\AppData\Local\Enthought\Canopy32\User\python.exe C:/Users/PK/Desktop/ClusteringMethods/ClusteringMethods.py
Give the path of a file to load:
e:\python\data.csv
Normalize the data (y/n):
n
Input one of the keywords (nearest, average, farthest), depending on what clustering method is to be applied:
farthest
Give the path of a file to save:
e:\python\savefn.csv
The distance matrix of the data is:
[[  0.      3.     12.21    8.77    4.9 ]
 [  3.      0.     12.88    9.38    5.39]
 [ 12.21   12.88    0.      3.74    7.55]
 [  8.77    9.38    3.74    0.      4.12]
 [  4.9     5.39    7.55    4.12    0.  ]]
Dendrogram saved at:
e:\python\savefn.png
```

Method: Average method
Normalization: yes



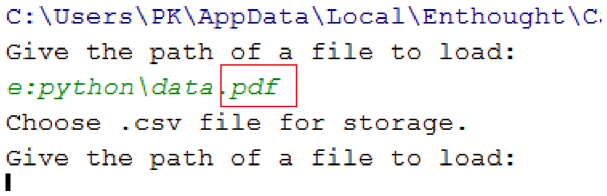
```
C:\Users\PK\AppData\Local\Enthought\Canopy32\User\python.exe C:/Users/PK/Desktop/ClusteringMethods/ClusteringMethods.py
Give the path of a file to load:
e:\python\data.csv
Normalize the data (y/n):
y
Input one of the keywords (nearest, average, farthest), depending on what clustering method is to be applied:
average
Give the path of a file to save:
e:\python\saveay.csv
The distance matrix of the data is:
[[ 0.    0.36  1.53  1.13  0.62]
 [ 0.36  0.    1.6   1.2   0.66]
 [ 1.53  1.6   0.    0.44  0.94]
 [ 1.13  1.2   0.44  0.    0.55]
 [ 0.62  0.66  0.94  0.55  0.  ]]
Dendrogram saved at:
e:\python\saveay.png
```

Method: Average method
Normalization: no

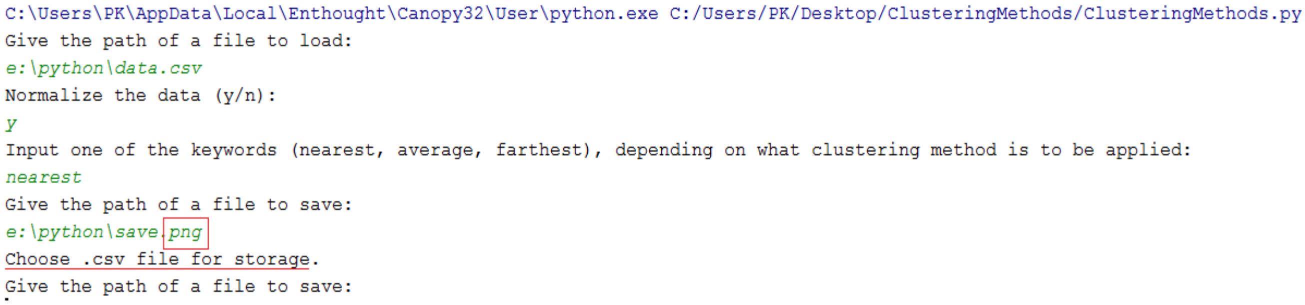


```
C:\Users\PK\AppData\Local\Enthought\Canopy32\User\python.exe C:/Users/PK/Desktop/ClusteringMethods/ClusteringMethods.py
Give the path of a file to load:
e:\python\data.csv
Normalize the data (y/n):
n
Input one of the keywords (nearest, average, farthest), depending on what clustering method is to be applied:
average
Give the path of a file to save:
e:\python\savean.csv
The distance matrix of the data is:
[[  0.     3.    12.21   8.77   4.9 ]
 [  3.     0.    12.88   9.38   5.39]
 [ 12.21  12.88   0.     3.74   7.55]
 [  8.77   9.38   3.74   0.     4.12]
 [  4.9    5.39   7.55   4.12   0.  ]]
Dendrogram saved at:
e:\python\savean.png
```

At the first and last step, it is necessary the file is a CSV file.

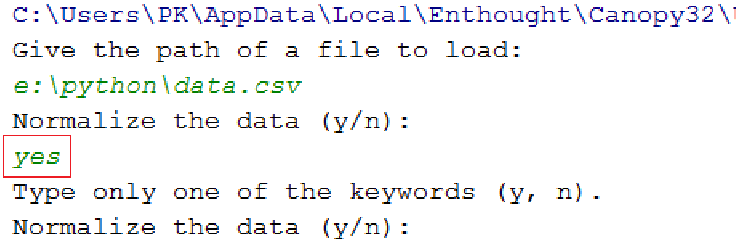


```
C:\Users\PK\AppData\Local\Enthought\C
Give the path of a file to load:
e:python\data.pdf
Choose .csv file for storage.
Give the path of a file to load:
```



```
C:\Users\PK\AppData\Local\Enthought\Canopy32\User\python.exe C:/Users/PK/Desktop/ClusteringMethods/ClusteringMethods.py
Give the path of a file to load:
e:\python\data.csv
Normalize the data (y/n):
y
Input one of the keywords (nearest, average, farthest), depending on what clustering method is to be applied:
nearest
Give the path of a file to save:
e:\python\save.png
Choose .csv file for storage.
Give the path of a file to save:
```

At the second step, it is necessary to import one of the two choices, "y" or "n".



```
C:\Users\PK\AppData\Local\Enthought\Canopy32\
Give the path of a file to load:
e:\python\data.csv
Normalize the data (y/n):
yes
Type only one of the keywords (y, n).
Normalize the data (y/n):
```

Finally, we should mention the tools used to implement the code. The python IDE used is PyDev utilized through the LiClipse editor, a special, lightweight version of Eclipse (http://www.liclipse.com/download.html). We also used the libraries math, numpy, scipy, matplotlib, os and csv for the necessary functions described in the code (sqrt, pow, inf, dendrogram, pyplot, csv, isfile, exists). The python version interpreting the code is python 2.7. Both the libraries included and the interpreter used, were found in the Enthought Canopy package distribution (https://store.enthought.com/downloads/#default).

Chapter 5 Conclusion

In this thesis three hierarchic clustering methods, single linkage (nearest neighbor method), complete linkage (furthest neighbor method) and average method were analyzed. Also their weak and strong points were discussed. Finally each algorithm was presented using Python and for the presentation was used the same example/data so that the reader can see the differences. "The choice of a clustering method depends on the material and the objectives at matter because different clustering methods can lead to different results. No method is considered superior, but some methods are more suitable for certain situations than others (see KAUFMAM; ROSSEEUW, 2005).

References

1. *El-Hamdouchi and P.Willett 1989, "Comparison of Hierarchic Agglomerative Clustering Methods for Document Retrieval", The computer journal, volume 32 pages 220-227*
2. *Ben Etzkorn 2011, http://www.benetzkorn.com/2011/11/data-normalization-and-standardization/*
3. *Bhavesh Patankar, Dr. Vijay Chavda ,2015, A survey on Improving Classification Accuracy in Data Mining,IJSRSET,volume 1,Issue 2*
4. *Dillon, W.R. and Goldstein, M., Multivariate Analysis, New York: 1984;*
5. *Fung B.C.M., Wang K., Ester M.(2009). Hierarchical Document Clustering&#8221;, in Wang J. (Ed.): Encyclopedia of Data Warehousing and Mining, IGI Global*
6. *Hair, J.F., Anderson, R.E., Tatham, R.L., and Black, W.C., Multivariate Data Analysis, Upper Saddle River, NJ: Prentice Hall, 1998;*
7. *Jiawei Han, Micheline Kamber, and Jian Pei, 2012, Chapter 12, Data Mining: Consepts and Techniques*
8. *Johnson, D.E., Applied Multivariate Methods for Data Analysis, Pacific Grove, CA: Duxbury Press, 1998.*
9. *Kaufman, Rosseeuw, Finding groups in data, An introduction to cluster analysis , 2005*
10. *L.Sunita, M. Bal Raju, B. Sunil Srivinas, 2013, A Comperative Study between Noisy and Outlier Data in Data Mining, INPRESSCO,Vol.3,No.2*
11. *Luiza Barbosa da Matta, L&#237;via Gracielle Oliveira Tom&#233;, Caio C&#233;sio Salgado , Cosme Dami&#227;o Cruz, Let&#237;cia de Faria Silva (2015). Hierarchical genetic clusters for phenotypic analysis, Acta Scientiarum Agronomy, Maring&#225;, v. 37, n. 4, p. 447-456.*
12. *Norusis, 2016 IBM SPSS STATISTICS 19 Guide to Data analysis http://www.norusis.com/pdf/SPC_v13.pdf*
13. *Peter Tryfos, 1998 METHODS FOR BUSSNESS ANALYSIS AND FORCASTING*
14. *Rosie Cornish, 2007 Chapter 3 Cluster Analysis Mathematics learning support Center http://www.lboro.ac.uk/departments/mlsc/*
15. *Wolfgang Härdle and Léopold Simar (2007, Applied Multivariate Statistical Analysis)*
16. *Zinzaras, 2016, Course Advanced_Statistics,*