Performance Comparison with Hierarchical and Partitional Clustering Methods

OZER OZDEMIR
Department of Statistics
Eskisehir Technical University
Eskisehir
TURKEY

SIMGENUR CERMAN
Department of Statistics
Eskisehir Technical University
Eskisehir
TURKEY

Abstract: - In data mining, one of the commonly-used techniques is the clustering. Clustering can be done by the different algorithms such as hierarchical, partitioning, grid, density and graph based algorithms. In this study first of all the concept of data mining explained, then giving information the aims of using data mining and the areas of using and then clustering and clustering algorithms that used in data mining are explained theoretically. Ultimately within the scope of this study, "Mall Customers" data set that taken from Kaggle database, based partitioned clustering and hierarchical clustering algorithms aimed at the separation of clusters according to their costumers features. In the clusters obtained by the partitional clustering algorithms, the similarity within the cluster is maximum and the similarity between the clusters is minimum. The hierarchical clustering algorithms is based on the gathering of similar features or vice versa. The partitional clustering algorithms used; k-means and PAM, hierarchical clustering algorithms used; AGNES and DIANA are algorithms. In this study, R statistical programming language was used in the application of algorithms. At the end of the study, the data set was run with clustering algorithms and the obtained analysis results were interpreted.

Key-Words: - Data mining, Clustering, Clustering algorithms, R programming

Received: May 27, 2021. Revised: November 17, 2021. Accepted: November 28, 2021. Published: December 27, 2021.

1 Introduction

Data mining is a set of techniques and concepts that generate new information for decision-making systems. It means extracting information from large data sets. It is an interactive and obsessive process that includes many steps that cover the decisions made by the user. Many definitions have been made about data mining. [1] Data mining or information discovery in databases, as is known, is the non-trivial extraction of implicit, previously unknown, and potentially useful information from data.

[2] on the other hand, defines Data Mining as a step of the Information Discovery Process formed by the application of data analysis and discovery algorithms that reveal certain patterns within the data within the framework of acceptable numerical efficiency.

[3] categorize data mining functions into classification, segmentation, prediction, and description.

[4] defined the data as an analysis that reveals unexpected relationships by summarizing the data like a novel, in a useful and understandable way.

According to [5], it is to extract previously unknown and useful information from the data.

In [6], data mining is a rapidly developing field that can analyze large amounts of data collected from applications such as manufacturing, bioinformatics and geological information systems.

According to [7], data mining is the basis of the KDD process, which involves the inference of

algorithms that research data, develop the model, and discover previously unknown patterns.
Data mining according to [8]; It is the process of discovering interesting information in large amounts of data in databases, data warehouses or other data stores.
It is used in many disciplines with different names in data mining, statistics and other fields. It has also been mentioned by different names such as data archeology, data browsing, data fishing, self-knowledge extraction or self-knowledge discovery in databases.

2 Methods

Many clustering methods have been developed to perform cluster analysis, which is one of the most widely used data mining techniques.
Clustering methods create a hierarchy depending on the clustering methods they use.
At the top level of the hierarchical structure, clustering methods are divided into hierarchical and partitional methods. Hierarchical methods create nested set series, while split methods create single-level sets.

2.1 Partitioning Clustering Techniques

2.1.1 K-Means Clustering Algorithm

The algorithm was introduced by [9] and named [10], [13].
The k-Means Clustering Algorithm is one of the Most Used Algorithms in Data Mining World [14].
The letter "k" in the name of the algorithm actually indicates the number of clusters. The algorithm minimizes the Quadratic Error Function, which is widely used in error calculation, and looks for the number of clusters "k". Given "n" number of data sets should be placed in "k" sets in a way that minimizes the error function. For this reason, cluster similarity is measured by the proximity of the values in the cluster to the mean. This becomes the center of gravity of the cluster. The value in the center of the cluster is the representative value of the cluster and is called the medoids.
In order to fulfill these requests, the following steps should be performed on the algorithm side, respectively:

Class centers should be determined.

Samples should be classified according to distances.

After the classification, new centers should be determined.

Steps 2 and 3 should be repeated according to the flow chart until it becomes desired.

Each object in the data set is assigned to the cluster most similar to it according to the distance between itself and the cluster average, and the cluster averages are calculated again for each cluster. This iteration continues until the criterion function converges at a common point. Usually the criterion function is defined as follows.

$$E = \sum_{i=1}^{k} \sum_{p \epsilon C_i} |p - m_i|^2 \quad (1)$$

Where p and m_i are the error sum of squares for multidimensional objects
E: Displays the total error squares for all objects in the database
p: It shows the point that represents a certain object.
m_i : Represents the mean of C_i clusters.
The K-means algorithm can be summarized as follows:
Input: Displays the number of databases and clusters of n objects.
Output: The error indicates the set of clusters that makes the sum of squares minimum.
Step 1: The first cluster centers should be determined.
Step 2: The distance of each object to the selected center points is calculated and according to the results, all objects are placed in the cluster closest to them from the k clusters.
Step 3: The new center points of the formed clusters are replaced with the average value of all objects in that cluster.
Step 4: Step 2 and Step 3 are repeated until the center points do not change.
When the k-means algorithm is used in categorical data, two problems are encountered: formulation of cluster centers and secondly, calculation of distances between objects and cluster centers.
The first problem can be solved by using the simple fit coefficient for categorical data instead of the Euclidean distance measure. The second problem is solved by using the mod instead of the cluster average.
Thus, another form of the k-means technique is the k-modes technique. When the k-means technique and the k-modes technique are combined to cluster mixed numeric data and categorically valued data, it is called the technical prototype (principal model) technique [9].

2.1.2 K-Medoids Clustering Algorithm

The algorithm was introduced by [11]. In the k-medoids algorithm, which is different from k-means and EM algorithms, the cluster centers are determined not by the average values of the objects in the cluster, but by finding the objects closest to the cluster center (medoids). In this way, it is less sensitive to noisy and outliers compared to the k-means technique [15].

Two of the oldest k-medoids algorithms put forward are PAM (Partitioning Around Medoids) and CLARA (Clustering LARge Applications) algorithms [16].

PAM algorithm: In the PAM algorithm, the number of clusters is determined by the user. It works well in small databases. The calculation complexity of each iteration is solved by the following formula;

$$O[k(n-k)^2] \quad (2)$$

n: the number of objects and k: the number of clusters.

The K-medoids algorithm is summarized as follows.

Step 1: The number of clusters must be determined.

Step 2: The k objects should be selected as starting medoids.

Step 3: Assign the remaining objects to the set with the closest medoid x.

Step 4: The objective should calculate the function. (Error squares criterion: sum of distances of all objects for nearest medoids)

Step 5: The y point must be chosen randomly.

Step 6: If the replacement of x and y will minimize the objective function, the places of these two points (x and y) should be changed.

Step 7: Step 3 and Step 6 should be repeated sequentially until there is no change [17], [19].

The PAM algorithm can only be applied to small data sets. In large data sets, it cannot show the desired sensitivity.

In order to complete this shortcoming of the PAM algorithm, Kaufman and Rousseeuw developed the CLARA algorithm in 1990 [12].

2.2 Hierarchical Clustering Techniques

2.2.1 Agglomerative Hierarchical Clustering Algorithms

AGNES (Agglomerative Nesting), which is the associative clustering algorithm, was presented by [12]. It follows a building structure running from bottom to top. To begin with, each object is considered as a separate set. For each subsequent step of the algorithm, those atomic clusters with similar properties are combined, and the total number of clusters decreases by one after each merging process. The merging process ends when the desired number of clusters are obtained or when the distance between the two closest clusters reaches the given threshold value. If no termination condition is given, when the clustering process is complete, all objects are gathered into a single cluster. Some techniques are used to speed up the process. These techniques are BIRCH, CURE, ROCK and CHAMELEON algorithms.

2.2.2 Divisive Clustering Algorithms

The DIANA (Divisive Analysis) algorithm was introduced by [12]. It follows a top-down building structure. To begin with, all of the data objects are considered as a single set, and for each subsequent step of the algorithm, the objects with the highest similarity among themselves are brought together and a new set is created and the large set is divided into two. This process continues until each object forms a cluster on its own or a certain termination condition is met. Termination condition is achieved to obtain the desired number of clusters or to ensure that the distance between the two closest clusters is above the given threshold value [11].

When the parser clustering technique is compared with the combiner clustering technique; Discriminatory clustering techniques require more processing and are more likely to produce erroneous results. Therefore, adder clustering techniques are preferred for applications [18].

3 Problem Solution

The "Mall Customers" data set [19] has been transferred to the R Program.

Variables are;

Customer ID: unique customer identification number.

Gender: Male and Female.

Age: Age of customers.

Annual Revenue (k $): Annual revenue of customers in thousands of dollars.

Spending Score (1-100): It is the spending score assigned by the shopping center according to the purchasing behavior of the customer.

```
mallcustomers<-read.csv("Mall_Customers.csv",header=TRUE)
str(mallcustomers)

## 'data.frame':    200 obs. of  5 variables:
##  $ CustomerID            : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Gender                : chr  "Male" "Male" "Female" "Female" ...
##  $ Age                   : int  19 21 20 23 31 22 35 23 64 30 ...
##  $ Annual.Income..k..    : int  15 15 16 16 17 17 18 18 19 19 ...
##  $ Spending.Score..1.100.: int  39 81 6 77 40 76 6 94 3 72 ...
```

Figure 1: Mall Customer Data Set

The variable names have been changed to make it more understandable, they have been made more specific. Statistical Summary is produced. According to the R results;
The data set contains information about 200 clients, 88 men and 112 women. The annual income range is between 15 thousand and 137 thousand in dollars, and their age is between 18 and 70. Their spending points include values between 1 and 99.

```
#The names of the columns have been changed to avoid confusion.
col_names <- c('customerid',
               'Gender',
               'Age',
               'AnnualIncome',
               'SpendingScore')
names(mallcustomers) <- col_names
##Since the gender variable is not in numerical form, we convert the gender
variable into a two-level categorical variable.

mallcustomers$Gender <- as.factor(mallcustomers$Gender)

summary(mallcustomers)

##    customerid        Gender         Age         AnnualIncome    SpendingScore
##  Min.   :  1.00   Female:112   Min.   :18.00   Min.   : 15.00   Min.   : 1.00
##  1st Qu.: 50.75   Male  : 88   1st Qu.:28.75   1st Qu.: 41.50   1st Qu.:34.75
##  Median :100.50                Median :36.00   Median : 61.50   Median :50.00
##  Mean   :100.50                Mean   :38.85   Mean   : 60.56   Mean   :50.20
##  3rd Qu.:150.25                3rd Qu.:49.00   3rd Qu.: 78.00   3rd Qu.:73.00
##  Max.   :200.00                Max.   :70.00   Max.   :137.00   Max.   :99.00
```

Figure 2: Numericalization of Data and Statistical Summary

3.1 Correlation Analysis

Scatter Plots were used to examine the correlation and relationship between variables. A function has been defined to create the scatter plots.

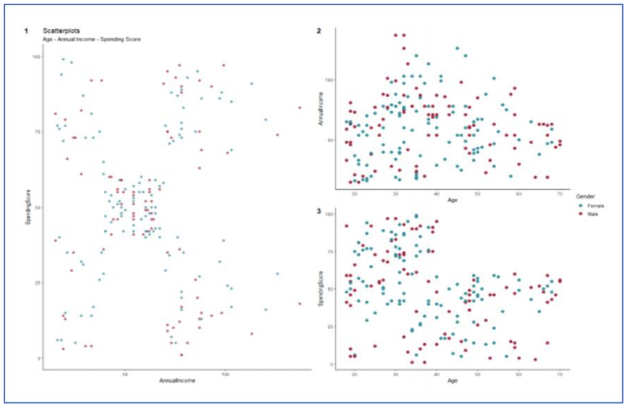

Figure 3: Correlation Analysis in Scatter Plot

There is no relationship between the variables of Score, Income and Age. Chart A in the figure below shows different groups with different combinations of Income and Score variables. What we can highlight in Chart B is that customers between the ages of 30 and 60 are the only customers who can exceed $ 80,000 a year. We can also see that customers who score over 60 points on the C chart belong to the group under 42.

```
# Preprocessing to data

# Since the clustering model to be implemented works by measuring the distanc
es between the different properties, it makes sense that they are numerical,
so a new transformation should be done. In the gender variable, let the value
be 0 for female customers and 1 for male customers.

mallcustomersprep <- mallcustomers[2:5]

mallcustomersprep$Gender <- if_else(mallcustomers$Gender == 'Female', 0, 1)
mallcustomersprep[2:4] <- scale(mallcustomersprep[2:4])
head(mallcustomersprep)

##   Gender        Age AnnualIncome SpendingScore
## 1      1 -1.4210029    -1.734646    -0.4337131
## 2      1 -1.2778288    -1.734646     1.1927111
## 3      0 -1.3494159    -1.696572    -1.7116178
## 4      0 -1.1346547    -1.696572     1.0378135
## 5      0 -0.5619583    -1.658498    -0.3949887
## 6      0 -1.2062418    -1.658498     0.9990891
```

Figure 4: Preprocessing of Data

Data preprocessing has been done. Algorithms give better results when all variables are in the same structure. Gender variable was categorized with 2 variables.

3.2 Clustering Analysis with K-Means Algorithm

A function has been defined to implement the K-Means algorithm. This function contains the K-average steps sequentially. First of all, the number of clusters is determined by the elbow method. It should be noted that the number of clusters in the K-means algorithm must be specified in advance. It has an important place in the distance criteria to be selected while applying the K-means algorithm. Euclidean Distance is used in this application. Additionally, the Accent and Pastel1 palettes are used to color the charts.
While clustering was made among age-income, score-age, score-income and finally all variables, the best cluster number was determined in the elbow method. Then a bar graph was drawn showing the sizes of the clusters to show the dimensions of the clusters. As a result, it was desired to observe how and to what extent cluster centers have outliers with data visualization. While doing this application, "ggplot2" library was used.

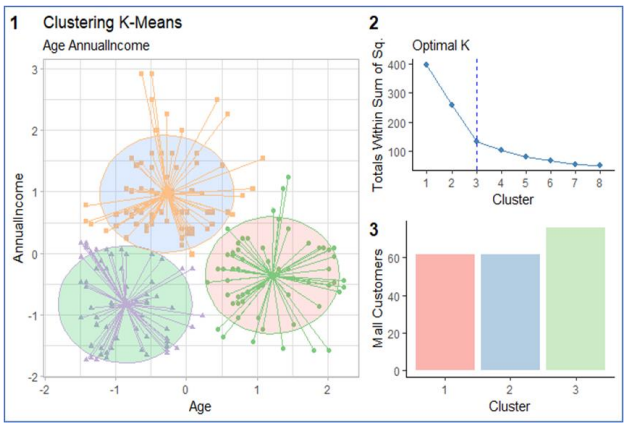

Figure 5: Age and Income Variables Graphical Display

When the Elbow Graph is examined, it is clear that we will get the best result when k = 3. The outlier is high, which is a situation we do not want. The cluster shown in green in general has more elements than the other 2 clusters. Clustering charts created between age and annual income values very close to each other.

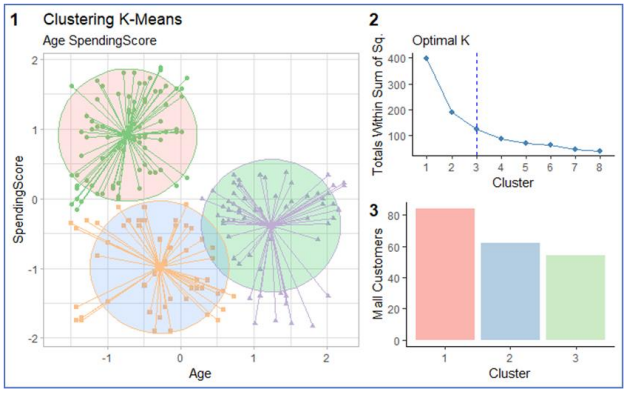

Figure 6: Age and Spending Score Variables Graphical Display

It is seen that we will get the best result when k = 3 in the elbow graph showing how many clusters we will get the best result of the connection between the spending score and age variables.

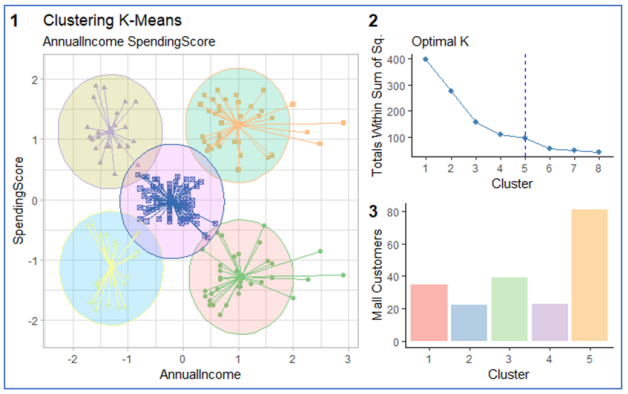

Figure 7: Annual Income and Spending Variables Graphical Display

Among the variables of expenditure and annual income, the best result according to the elbow chart is when k = 5. When we examine it, it is seen that the outliers have decreased.

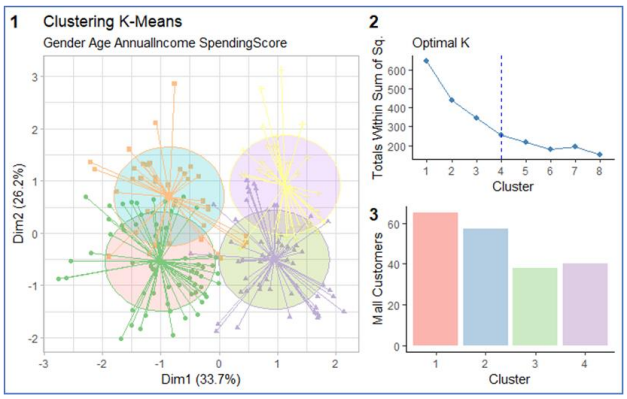

Figure 8: Age, Income, Spending Score and Gender Variables Graphical Display

When all variables are included, the best clustering is taken when the number of clusters = 4 The elbow chart shows us that clustering is best between 3 and 5.

Intra-cluster sum of squares is a measure of the variability of observations within each cluster. Clusters with higher values show more variability in observations within the cluster. The higher the number of observations, the greater the sum of squares. Therefore, the sum of in-cluster squares cannot usually be directly compared between clusters with different observation numbers. The ratio of the totals we have obtained to each other is 40.4%.

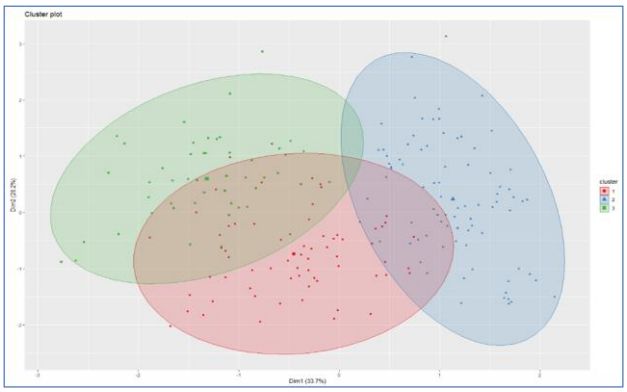

Figure 9: Dataset Representation Divided Into 3 Clusters

It looks like clusters are internal. There are several outliers. It is the desired state.

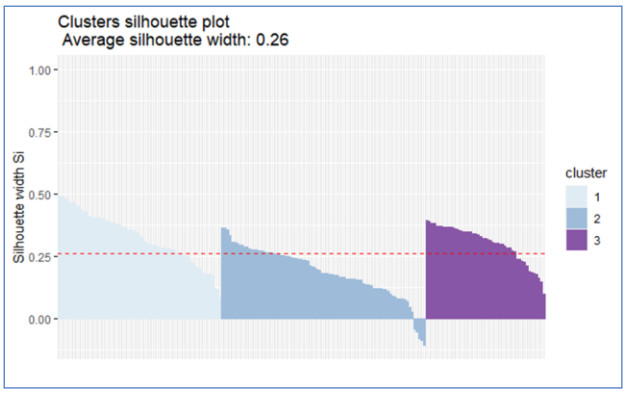

Figure 10: Silhouette Plot Representation Divided into 3 Clusters

Silhouette analysis can be used to examine the separation distance between emerging clusters. When the silhouette analysis is done for k = 3, it is seen that the sizes of the clusters are almost close to each other. Therefore, it is similar in size that can be considered best as "k".

3.3 Clustering Analysis with PAM Algorithm

PAM firstly takes k numbers randomly selected as the cluster center as in the k-means algorithm. Each time a new element is added to the set, when it determines the point that can contribute to the development of the cluster by trying the elements of the cluster, the point it finds is the new center and the old center is the normal cluster element.
"Factoextra","cluster" and "ggthemes" libraries were used in the application. The graphs plotted for the PAM algorithm are listed below. In the first graph, clusters divided into two variables are shown. In the cluster plot, the clusters are divided into two clusters of equal size numbered. Dividing it into 2 sets to data is a good choice.

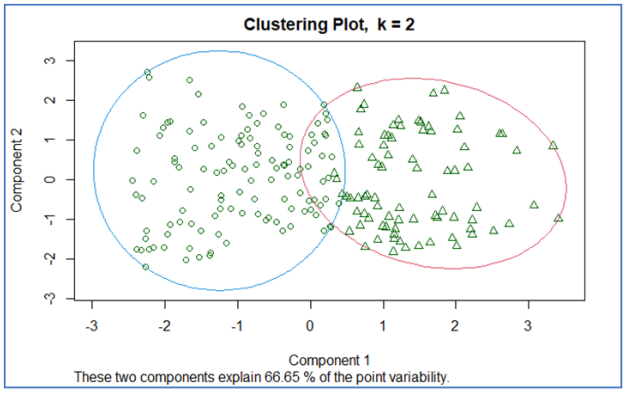

Figure 11: Clustering Plot with Pam Algorithm(k=2)

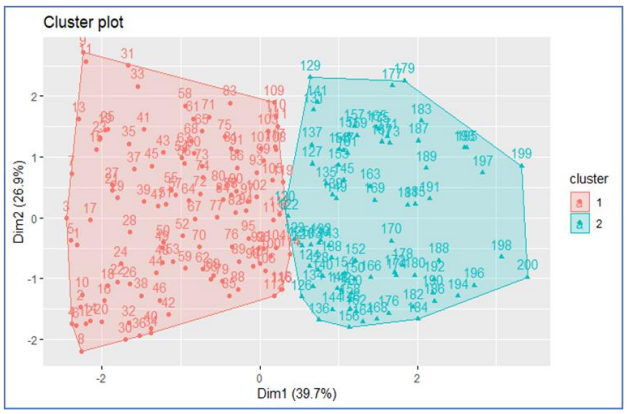

Figure 12: Clustering Plot with Pam Algorithm

In the Cluster Plot given above, the number of clusters is taken as 2 for Mall Customer data set. They are numbered and colored.

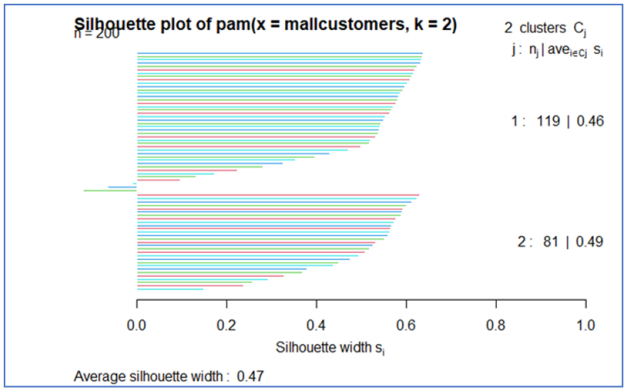

Figure 13: Silhouette Plot with Pam Algorithm(k=2)

In the Silhouette Plot given above, the number of clusters for the Mall Customer data set is taken as 2.

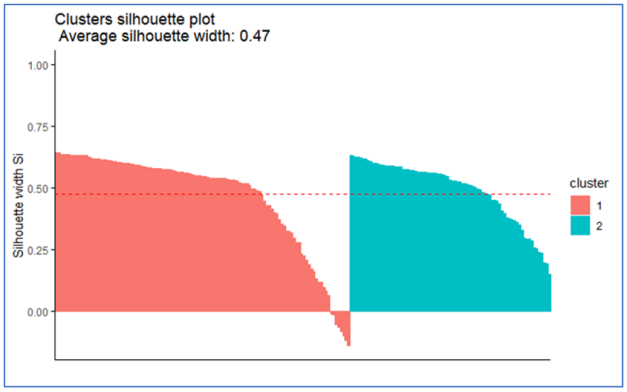

Figure 14: Silhouette Plot with Pam Algorithm

When the various silhouette graphics obtained are examined, it is seen that there are negative values. Elements in the negative state are not in the correct set. Elements must be identified and placed in the correct cluster. Since it is seen that the sizes of the clusters are close to each other and the number of negative value elements is low, taking the number

of clusters as 2 gave us positive results. Therefore, it is similar in size that can be considered best as "k".

3.4 Clustering Analysis with AGNES (Agglomerative Nesting)

It consists of 6 methods. The “ward” method is used in this application, minimizing the total intra-cluster variance. In this algorithm, clusters with a minimum inter-cluster distance are combined in the step.

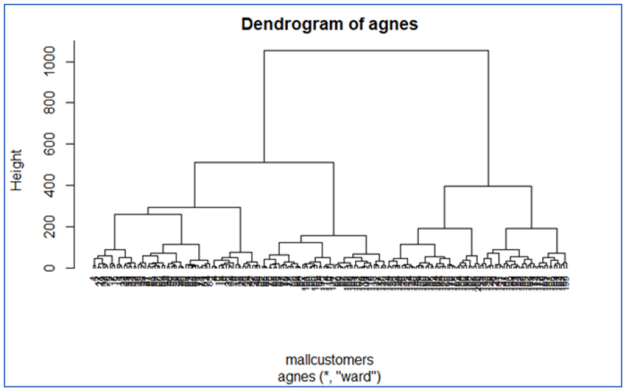

Figure 15: Dendogram of Agnes

It clusters based on Euclidean distance and correlation distance using single link aggregation. The algorithm starts by treating each object as a single set. The cluster pairs are then concatenated until all the clusters are combined into a single large cluster containing all objects. Dendrogram is a tree-based representation of objects.

At the bottom is the dataset. As you climb up the tree, the branches that are similar to each other begin to merge with the branches. These also converge at higher levels. There is no exact way to decide where to cut the tree. Dendogram can be drawn horizontally or vertically.

3.5 Clustering Analysis with DIANA (Divisive Analysis)

The DIANA algorithm is considered the opposite of the AGNES algorithm.

We obtained the merge values, order of objects values, height values, and divisive coefficient values in R programming with the Diana function. Diana works in a "top down" process. It begins with the root in which all objects are included in a single cluster. The process is iterated with all objects are in their own cluster. Dendrogram was drawn with the obtained values. The most heterogeneous cluster a split into two at each step.

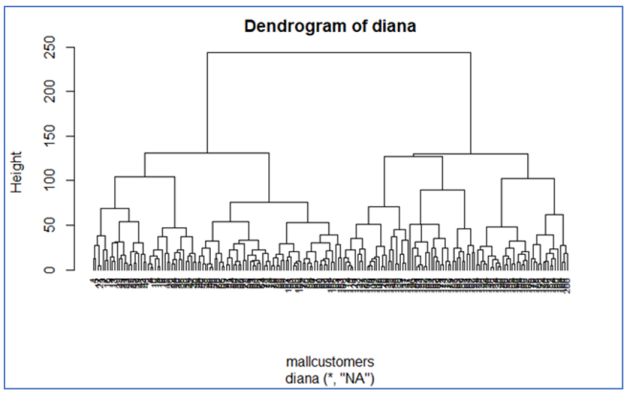

Figure 16: Dendogram of DIANA

There is no need to determine the number of clusters, it can be divided as many times as you want. It gives more real results, is faster, but did not give explanatory values for this data set.

4 Conclusion

In this study, it is compared which of the two clustering methods will give more explanatory results when R programming and customer segmentation analysis are applied. R Programming is used successfully for statistical operations. The results obtained are very rich in terms of visuality. Firstly, the K-Means clustering algorithm is discussed. The K-Means algorithm applied with customer data has been examined with trials. Euclidean application is used for data transitions. It took 0 seconds to complete the operation. K-Means algorithm has been very useful for this study in terms of ease of use, speed of operation and explainability. PAM clustering approach is less sensitive to outliers and provides an alternative to k-means to deal with these situations. It is also beneficial for Mall Clients to validate PAM and it has been explained. AGNES and DIANA algorithms work in reverse. It was tried to be explained through dendograms. The AGNES and DIANA coefficients were quite close to each other. The difficulty in clustering analysis is that basically different clustering techniques give significantly different results on the same data. In addition, there is no algorithm that gives all the desired outputs. In this thesis, performance comparison between partitional clustering and hierarchical clustering algorithms is made. In the study, for the Mall Customers data set; It was seen that partitional clustering algorithms obtained more explainable results than hierarchical clustering algorithms. It was concluded that the hierarchical clustering algorithms discussed obtained less explainable results for this data set.

References:

[1] Piatetsky-Shapiro G. (1996). Data mining and knowledge discovery in business databases. In: Raś Z.W., Michalewicz M. (eds) Foundations of Intelligent Systems. ISMIS 1996. Lecture Notes in Computer Science (Lecture Notes in Artificial Intelligence), vol 1079. Springer, Berlin, Heidelberg. https://doi.org/10.1007/3-540-61286-6_131.

[2] Fayyad, U. , Gregory P.S, Padhraic S. (1996), From Data Mining to Knowledge Discovery in Databases, vol 17, no. 3.

[3] Westphal, C. ve Blaxton, T. (1998). Data Mining Solutions: Methods and Tools for Solving Real-World Problems. U.S.A.: Wiley & Sons.

[4] Hand, D., Mannila H., Symth P. (2001), Principles of Data Mining.

[5] Witten, I. H., Frank E., Hall, M. A, Pal, C. J, (2005), Data Mining Practical Machine Learning Tools and Techniques, 4th Edition.

[6] Ye, N, (2003), "The Handbook of Data Mining", Lawrence Erlbaum Associates.

[7] Maimon, O., Rokach L., (2006), Data Mining and Knowledge Discovery Handbook.

[8] Han, J. and Kamber, M., (2012) Data Mining-Concepts and Techniques, 3rd Edition, Morgan Kauffman Publishers.

[9] Cox, D.R., (1957), Note on Grouping. Journal of Amer. Statist. Assoc., 52, 543-547.

[10] Ball, G. H., & Hall, D. J. (1967). A clustering technique for summarizing multivariate data. *Behavioral science*, *12*(2), 153-155.

[11] Rdusseeun, L. K. P. J., & Kaufman, P. (1987). Clustering by means of medoids. In *Proceedings of the Statistical Data Analysis Based on the L1 Norm Conference, Neuchatel, Switzerland* (pp. 405-416).

[12] Kaufman, L., and Rousseeuw, P. J., (1990), *Finding groups in data: An introduction to cluster analysis*. Wiley, Hoboken, NJ.

[13] MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability* (Vol. 1, No. 14, pp. 281-297).

[14] Ohn Mar San, Van-Nam Huynh and Yoshiteru Nakaromi, (2004), An Alternative Extension of the K-Means Algorithm for Clustering Categorical Data.

[15] R. Treur, (2003), Spatial Clustering Methods in Data Mining.

[16] Pavel Berkhin, (2002), Survey of Clustering Data Mining Techniques.

[17] Musa Peker, (2016), A decision support system to improve medical diagnosis using a combination of k-medoids clustering based attribute weighting and SVM, *Journal of Medical Systems*, 40, 116.

[18] Grabmeier, J., & Rudolph, A. (2002). Techniques of cluster algorithms in data mining. *Data Mining and knowledge discovery*, *6*(4), 303-360.

[19] https://www.kaggle.com/vjchoudhary7/customer-segmentation-tutorial-in-python.