JSDS: JOURNAL OF STATISTICS AND DATA SCIENCE
VOLUME 1, No 1, March 2022

https://ejournal.unib.ac.id/index.php/jsds/index

Agglomerative Nesting (AGNES) Method and Divisive Analysis (DIANA) Method For Hierarchical Clustering On Some Distance Measurement Concepts

Susi Wijuniamurti*, Sigit Nugroho and Ramya Rachmawati
Mathematics Department, Mathematics and Natural Sciences Faculty, University of Bengkulu, Bengkulu, Indonesia

* Corresponding Author. Email: swijuniamurti@gmail.com

Article Info

Article History:
Received: December 17, 2021
Revised: January 4, 2022
Accepted: January 18, 2022
Available Online: March 1, 2022

Key Words:
Hierarchical Cluster Analysis,
AGNES
DIANA
Euclid Distance
Manhattan Distance

Abstract

Clustering data through hierarchical approach could be performed by Agglomerative Nesting (AGNES) Method and Divisive Analysis (DIANA) Method. The objective of this research is to compare both the methods based on Euclid and Manhattan distance measurements. Of this research the clustering procedures of agglomerative method are conducted by exploring all techniques including single linkage, complete linkage, average linkage, and Ward. The data used are the National Socio-Economic Survey (SUSENAS) data which are selected specifically for the percentage of over 5 year old residents in each province, for both living in urban or rural, who access the internet in the last 3 months in 2017 but classified according purpose of accessing. By applying Mean Square Error (MSE) for 2 and 3 clusters, it can be concluded that the single linkage technique is the best performance of clustering procedure for both Euclidean and Manhattan distances.

1. INTRODUCTION

Cluster analysis using hierarchical approach constitutes a multivariate analysis technique for clustering observed data in such a way that each obtained cluster is an homogeneous cluster with respect to the techniques used for clustering such as single linkage, complete linkage, average linkage, and Ward. Substantially cluster analysis works by identifying a group of objects with similar characteristic which is different from the one of other objects so that objects located in a same group will be relatively more homogeneous than the ones of different group. The number of groups which could be identified depends on the size of data and the variety of object data. Moreover the data used could be the types of interval, frequency, and binary. Likewise each of the object data groups must represent a variable of the same type, not mixed with other groups of different variable types.

This research discusses the performance of how Agglomerative Nesting (AGNES) method and Divisive Analysis (DIANA) method work as well as the comparison of how effective they are. The former clusters the observed data by taking into account the techniques of clustering such as single linkage, complete linkage, average linkage, and Ward while the latter clusters the observed data through a top-down clustering approach beginning with one cluster for all observed data and then followed by splitting it recursively as one moves down the hierarchy. The implementation of both methods to the observed data is based on Euclid and Manhattan distance measurements. Finally the effectiveness of both methods on clustering is measured by determining Means Squared Error (MSE) of both distances obtained from applying both methods to the observed data. The effectiveness of both methods is determined on the basis of the accuracy of distance calculation with the help of R Program.

2. METHOD

This research is concerned with clustering data through hierarchical approach, especially by using two methods, those are Agglomerative Method and Divisive Method.

2.1 Agglomerative Method

Implementing Agglomerative Method for clustering data could be performed through its algorithm whose steps are as follows [1]:

1. Start from n clusters, each cluster contains only one object as its member.
2. Suppose $D_{n\times n} = [d_{rs}]$ is a close proximity matrix. Find an inequality matrix D for the most similar pair. For example, the chosen pair is united with the element d_{rs} so that the object r and s are chosen.
3. Link together the object r and s into one new cluster (rs) using some criteria and decrease the number of clusters by 1 through deleting row and column of object r and s. Calculate the inequality between cluster (rs) and all remain using criteria and add row and column to the new inequality matrix.
4. Repeat step 2 and 3 up to (n-1) times in order to form all objects in a single cluster. In every step, identify the cluster union and the value of inequality where the clusters are united.

2.2 Divisive Method

Divisive Hierarchical method could be positioned as the opposite of Agglomerative Hierarchical method. At the beginning, all data points are the members of single cluster. The next process is splitting the single cluster recursively as one moves down the hierarchy in order to get smaller cluster. Hierarchical Clustering Method has two types, those are monothetic and polythetic [2]. The algorithm of DIANA method can be explained through the following steps [3].

1. Suppose C is a cluster. Define diameter C as:

$$Diam(C) = \max_{x,y\in C} d(x,y).$$

2. Suppose C with $|C| \geq 2$ is a devided cluster resulting in clusters A and B. It means that $A \cap B = \emptyset$ and $A \cup B = C$. Initially $A = C$ and $B = \emptyset$ and the algorithm is to find A and B by displacing a point of A to B repeatedly. At the beginning a point y_1 is displaced from A to B provided that

$$D(x, A\backslash\{x\}) = \frac{1}{|A|-1} \sum_{y\in A, y\neq x} d(x,y),$$

where $d(.,.)$ is defined as a distance measurement between points of data. As a result, A_{new} and B_{new} can be obtained where

$$A_{new} = A_{old}\backslash\{y_1\} \quad \text{and} \quad B_{new} = B_{old} \cup \{y_1\}.$$

3. Investigate another point in A which must be displaced to B. Suppose $x \in A$ and the investigation function is defined by

$$D(x, A\backslash\{x\}) - D(x,B) = \frac{1}{|A|-1} \sum_{y\in A, y\neq x} d(x,y) - \frac{1}{|B|} \sum_{z\in B} d(x,z).$$

Suppose that the point y_2 maximizes the function above and the maximum result is positive, then y_2 must be displaced from A to B. If the maximum result is negative or zero, stop the process and the split of C to A and B end.

2.3 The Comparator Measure for Clustering Methods

A good cluster is the one which have high homogeneity among points within cluster and high heterogenity among clusters. In this research the measure used to compare methods in terms of the goodness is Mean Square Error (MSE). The formula used is as follows:

$$MSE = \frac{\sum_{k=1}^{l} \sum_{j=1}^{n_k} d_{kj}}{\sum_{k=1}^{l} n_k}$$

where d_{kj} is the distance between the j-th observed point to the k-th cluster.

3. RESULTS AND DISCUSSION

3.1 Description of Data

The data used in this research are secondary data as a result of National Socio-Economic Survey (SUSENAS) conducted by Central Bureau of Statistics (BPS). The data have been published in the form of a book entitled "Statistik Kesejahteraan Rakyat 2017" page 248. The data constitute the one presenting percentage of over 5 year old residents living in urban and rural area in each province but those specifically access internet within the last three months accompanied by the objective of accessing the internet in 2017.

3.2 Calculation of Distances

In this research distances used as the basis of cluster hierarchical analysis are the distances of Euclid and Manhattan. The calculation of distances is conducted for determining the similarity of inter-objects and the inequality of each objects. The formula for determining the Euclid distances are as follows [3]:

$$d_{euc}(x,y) = \left[\sum_{j=1}^{d}\left(x_j - y_j\right)^2\right]^{1/2},$$

where $d_{euc}(x,y)$ is distance between province x and province y. Since the number of province is 34 provinces and the number of observation variables is 10 variables, then it will generate a matrix of distance of 34×34 dimension.

Based on measurement of Euclidean distance it can be obtained the distance between Maluku Province and Southeast Sulawesi whose shortest distance is 14.2. It means that Maluku and Southeast Sulawesi have almost the same characteristic in terms of their objective of accessing internet and their spread of internet in Indonesia. However the farthest distance is 78.7 representing distance between Southeast Sulawesi and East Kalimantan.

The formula for calculating the Manhattan distance uses the following equation [3]:

$$d_{man}(x,y) = \sum_{j=1}^{d}\left|x_j - y_j\right|.$$

However when the points x and y have values n several variables, the Manhattan distance can be defined as [4]:

$$d_{manw}(x,y) = \sum_{k=1}^{d}\frac{w_j\left|x_j - y_j\right|}{\sum_{k=1}^{d} w_j}$$

As at Euclid's distance, the Manhattan distance $d_{man}(x,y)$ is the distance between the provinces x and y; the number of provinces is 34, and the observation variable is 10. The generated matrix is a matrix of size 34×34. Based on the calculation of the Manhattan distance with the help of the R application program, the closest distance is 43, that is the distance between Banten Province and West Java Province. While the farthest distance is 240, which is the distance between East Kalimantan and Southeast Sulawesi.

It can be seen from the range of distance values between Euclidean distance [14.2 , 78.7] and Manhattan distance [43, 240], Euclid distance is smaller and better than Manhattan distance because the smaller the distance, the more similar the measured variables.

3.3 Clustering Process

In this study, the clustering method used is the single linkage method, the complete method, the average method, the Ward method, and the divisive method. Each method is calculated based on a measure of the distance between Euclid and Manhattan. The calculation process is carried out using the R program. The results of the clustering using the R program are described in the form of a dendogram.

The dendogram results for each distance and clustering method were separated into 2 clusters and 3 clusters. Based on the observed dendogram results, for each single linkage method, complete method, average method, Ward method, and divisive method with different distance measures, there is no significant difference. However, when the methods are compared, the shape of the dendogram and the formed cluster members are very different. This is due to the different calculation processes.

3.4 Clustering Method Comparison

In this study, the determination of measure for comparing methods, namely MSE, is carried out with the help of Microsoft Excel. Each MSE is calculated based on the number of the formed clusters and the distance proximity used. The calculation results are concluded as follows:

a. MSE for 2 Clustering

The MSE calculation for each method using Euclidean Distance has yielded the following results as shown in Table 1

Tabel 1. The Two-Clustering Based MSE Values for Euclidean Distance

Method	MSE
Single	120.2935
Complete	163.2250
Average	164.5721
Ward	163.2250
Diana	159.4938

Based on Table 3.1, it can be seen that the MSE value for Single method has shown the smallest MSE value, that is 120.2935. In this case, the method has shown the best performance. On the other hand, the MSE calculation for each method using Manhattan Distance has yielded the following results as shown in Table 2.

Tabel 2. The Two-Clustering Based MSE Values for Manhattan Distance

Method	MSE
Single	123.1836
Complete	164.5721
Average	157.4010
Ward	146.2550
Diana	150.7328

Based on Table 3.2, it can be seen that the MSE value for Single Linkage method indicates the smallest MSE value, that is 123.1836. It means that Single Linkage method has shown the best performance.

Considering Table 1 and Table 2 it can be concluded that the province-based clustering according to the purpose of accessing internet in 2017 through the separation of 2 clusters can be done with a better result by Single Linkage method.

b. The Three Clustering-Based MSE

MSE calculations for three clustering using all considered method have yielded the results as shown in Table 3. as follows:

Tabel 3. The Three Clustering Based MSE Values for Eucliden Distance

Method	MSE
Single	116.4110
Complete	188.2378
Average	160.9910
Ward	190.3621
Diana	182.0976

Based on Table 3, it can be seen that the MSE value for Single method indicates the smallest MSE value, that is 116.4110. Therefore, Single method can be considered as the best performance method. However, the MSE calculation using Manhattan distance has given another insight as shown in Table 3.4.

Based on Table 4, it can be seen that the MSE value for Single Linkage method has indicated the smallest MSE value, that is 122.089. So, Single Linkage method constitutes the best performance method.

Tabel 4. MSE Values for Manhattan Distance

Method	MSE
Single	122.0089
Complete	191.3422
Average	203.7983
Ward	184.9664
Diana	193.0648

Finally in considering with Table 3 and Table 4 it can be concluded that the province-based clustering according to the purpose of accessing internet in 2017 but calculated by Euclidean and Manhattan distances, it is better to separate the 3 clusters using Single Linkage method.

4. CONCLUSION

The range of similarity measures between Euclidean distance, 14.2 – 78.7, and Manhattan distance, 43 – 240, has shown that Euclidean distance has demonstrated a better performance as indicated by its smaller calculation results than the one produced by Manhattan distance. Of course it is confirmed since the smaller the distance, the more similar the variables measured.

Based on the MSE value for grouping into 2 clusters, it shows that the single linkage method is the method that has the best performance in the clustering process, both for the Euclid distance measure and for the Manhattan distance measure. For grouping into 3 clusters, the single linkage method has also indicated the best performance with respect to the measures of Euclidean and Manhattan distance.

REFERENCES

[1] Timm, N.H., Applied Multivariate Analysis, New York, Springer-Verlag. Inc., 2002.

[2] Everitt, B, Cluster Analysis, Third Edition, New York, Halsted Press, 1993

[3] Gan, G., Ma, C., and Wu, J., Data Clustering: Theory, Algorithms, and Applications, ASA SIAM series on Statistics and Applied Probability, Philadelphia, 2007.

[4] Wishart, D., “K-Means Clustering with Outliers Detection, Mixed Variables and Missing Values”, Explanatory Data Analysis in Empirical Research, pages 216 – 226, New York, Springer, 2002