

Implementation of Agglomerative Hierarchical Clustering Using Single Linkage

Dr. G. Vijaya Lakshmi

Assistant Professor, Department of Computer Science, Vikrama Simhapuri University, Nellore, Andhra Pradesh, India

ARTICLE INFO

Article History:

Accepted : 14 Jan 2025

Published: 17 Jan 2025

Publication Issue

Volume 11, Issue 1

January-February-2025

Page Number

724-727

ABSTRACT

Classifying and grouping similar type of data objects in one segment is known as clustering. There are different types of techniques to create the clusters namely partitioning methods, hierarchical methods, density-based methods, and grid-based methods. . In this paper, we focus on agglomerative hierarchical techniques which is also known as bottom-up approach. In this approach the objects are successively form clusters into one until termination condition holds. In this paper, python we construct distance matrix by measuring the distance between the two data points. There after a dendrogram is prepared by providing a visual representation that helps to the find the optimal number of clusters.

Keywords: Data mining, Agglomerative, linkage, Hierarchal, distance matrix.

Introduction

Clustering is an unsupervised machine learning techniques ^[4] which is used to form a similar types of objects. There are three types of categories namely partitioning methods, hierarchical methods, density-based methods, and grid-based methods. The main objective of clustering is to partition a set of homogeneous data points from a heterogeneous dataset. It evaluates the similarity based on a metric like Cosine similarity, Manhattan distance, Euclidean distance, etc. and then group the data points with max, min or average similarity score together. This type of clustering can be used in various application namely

on Image processing, Document clustering, Market basket analysis, information retrieval etc. The Agglomerative algorithm ^[6] approach is one of the hierarchal clustering that uses bottom up approach . It starts with considering each observation as one cluster and then iteratively merges clusters until all the data points forms a single cluster. There are three types of linkage methods used to cluster namely single linkage, complete linkage, average linkage^{[10][7]}.

a) Types of Linkages:

- Single linkage: In this approach it computes the minimum distance between clusters before merging them.

- Complete linkage: In this method distance between two clusters is computed by considering the maximum distance between two points in each cluster.
- Average linkage: in this method the distance between each pair of datasets is the average distance between each data point in one cluster to the another cluster

In this paper we focus on how to form clusters using single linkage algorithm by using Euclidean distance and implementing it in python.

Literature Review

Manish Verma, Mauly Srivastava, Neha ^[11] compared various clustering techniques and analysed the that implementing the clustering algorithm using any software will draw the same result even when changing any of the attributes. since most of the clustering software uses the similar procedure in implementing any algorithm. F. Murtagh^{[12][6]} discussed in design of algorithms for hierarchical clustering it takes the nearest neighbor problem as a more primitive task. R. Raj Kumar ^{[2] [5]} explains the idea of Agglomerative Hierarchical Model^[3] and Divisive Model with its algorithmic implementation procedures with the data analysis. Yogita Rani ^[1] elaborates the concept hierarchical clustering algorithm ie, CURE (Clustering using representatives), BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies), ROCK (Robust Clustering using links), CHEMELEOM Algorithm, Linkage Algorithms, Leaders-Subleaders and Bisecting K-Means with the challenges of existing algorithms. Johnson and Wichern ^[8] confer that in agglomerative hierarchical and divisive hierarchical methods for clustering clusters form by considering that each object is a cluster. Later two data objects that have minimum distance are merged into single cluster. The process repeats till the end till it forms a cluster containing all data objects in one cluster.^[9]

Illustrative Example

a) Given dataset :

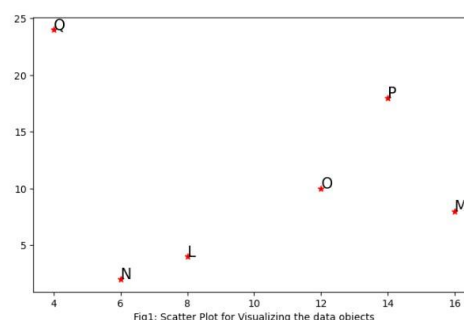
Single linkage clustering prepare clusters by calculating minimum distance between data point.

Consider there are 6 data points on the 2 dimensional space and each have two measured features X and Y. These data points are L,M,N,O,P,Q are shown in table 1. The procedure for the single linkage clustering is as follows:

Table1: Datasets with 6 Objects

Objects	X	Y
L	8	4
M	16	8
N	6	2
O	12	10
P	14	18
Q	4	24

Figure1 depicts the visualisation of proximity between objects by means of a scattered plot.



Step 1: First we begin by considering every datapoint as its one cluster. Then compute a distance matrix between every pair of objects that you need to join cluster. A distance matrix is a symmetric (because the distance between Land M is correspond to distance between M and L) and the diagonal will contains zeroes because every object is distance zero from itself.

In table2 we calculate a distance matrix from each object to all other using Euclidean distance (E.D)

$$E.D = d(L,M) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

Table2: Euclidean Distance for each object

objects	L	M	N	O	P	Q
L	0					
M	8.9	0				
N	2.82	11.6	0			
O	7.21	4.47	10	0		
P	13.4	10.1	17.8	8.24	0	
Q	20.3	20	22.09	16.12	11.66	0

Here only the lower triangle is considered, because the upper triangle can be filled in by reflection.

Step2: scan the minimum distance between the data points . The data points are L and N. So merge them together to form single cluster as shown in Table 3.

Step3: In single linkage, proximity of two cluster is measured as by taking minimum distance between two clusters .

So, calculating distance $\{(L,N),M\} = \text{Min}\{\text{distance}(L,M), \text{distance}(N,M)\}$

$$= \text{Min}(8.9, 11.6)$$

$$= 8.9.$$

Similarly calculating the distance of $[(L,N) O], [(L,N) P], [(L,N) Q]$.The result is shown as below in Table3.

Table3: Distance matrix (M,O)

Objects	L,N	M	O	P	Q
L,N	0				
M	8.9	0			
O	7.21	4.47	0		
P	13.4	10.1	8.24	0	
Q	20.3	20	16.12	11.66	0

In step 4, we merge O,M objects and form them as one cluster . we again recalculate the Distance of $[(L,N)(M,O)], [(M,O) P], [(M,O) Q]$ as shown in table4.

Table4: Distance matrix from (L,N)(M,O)

objects	L,N	M,O	P	Q
L,N	0			
M,O	7.21	0		

objects	L,N	M,O	P	Q
P	13.4	8.24	0	
Q	20.3	16.12	11.66	0

In step 5, we find minimum distance from each data point and form cluster as (M,O,L,N). And we calculate the distance between $\{(L,N,M,O), (P)\}, \{(L,M,N,O), Q\}$ as shown in table 5.

Table5: Distance matrix from (L,N, M,O),P

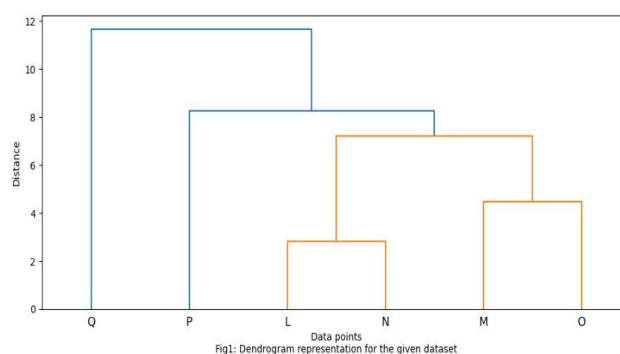
objects	L,N, M,O	P	Q
L,N, M,O	0		
P	8.24	0	
Q	16.12	11.66	0

In step 6, we merge L,N, M,O,P, Q and form as one cluster as shown in table 6.

Table 6: Distance matrix from (L,N, M,O,P)(Q)

objects	L,N, M,O,P	Q
L,N, M,O,P	0	
Q	8.24	0

b) **Dendrogram** : Based on the above matrices, a dendrogram chart is represented to find the optimal number of clusters.



From the above fig2, we can visualize that when we combine L, N the distance is 2.5, when we combine M,O the distance is 4.5, combining L,N,M,O the distance is 7.5, combining the P,L,N,M,O the distance is 8. Merging P,L,N,M,O the distance is 11.5. Later we find the optimal number of clusters based on their threshold value .For example if threshold value is

9. From the above dendrogram the optimal number of clusters is four since we have four bars below the horizontal line.

Conclusion

Single linkage clustering involves various key steps. In this paper, using python, we initially visualize the data objects. Again a distance matrix is calculated using Euclidean distance. Later depending on the shortest distance clusters are formed. Once each cluster is formed, the process iteratively repeated in updating the distance matrix to incorporate new distances. Finally all objects are clustered until revealing patterns in the data objects. Later we visualized the same using dendrogram to find the optimal number of clusters.

References

- [1]. Yogita Rani & Dr. Harish Rohil, "A Study of Hierarchical Clustering Algorithm", International Journal of Information and Computation Technology. ISSN 0974-2239 Volume 3, Number 11 (2013), pp. 1225-1232.
- [2]. R. Raj Kumar et al, Manish Verma, Mauly Srivastava, et al "A Comparative Study of Various Clustering Algorithms in Data Mining", International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622, Vol. 2, Issue 3, May-Jun 2012, pp.1379-1384.
- [3]. Osama Abu Abbas, "Comparison Between Data Clustering Algorithms", The International Arab Journal Of Information Technology, vol.5, No.3, July 2008.
- [4]. Han J. and Kamber M., Data Mining: Concepts and Techniques, Morgan Kaufmann publishers, 2001.
- [5]. K Kiruba, Dr B Rosiline Jeetha, "A Comparative Study on Hierarchical Clustering in Data Mining", 2014, International Journal Of Engineering Sciences & Research Technology, ISSN: 2277-9655, Feb 2014.
- [6]. Sunila Godara, Amita Verma, "Analysis of Various Clustering Algorithms", International Journal of Innovative Technology and Exploring Engineering (IJITEE), ISSN: 2278-3075 (Online), Volume-3 Issue-1, June 2013
- [7]. Adji Achmad Rinaldo Fernandes et al, "Comparison of the Use of Linkage in Cluster Integration With Path Analysis Approach, study uses the cluster integration with path analysis method by comparing the linkage.", Front. Appl. Math. Stat., 23 August 2022., <https://doi.org/10.3389/fams.2022.790010>.
- [8]. Johnson RA, dan Wichern DW. Applied Multivariate Analysis. Upper Saddle River, NJ: Prentice Hall (2007).
- [9]. Fernandes S, Rinaldo AARAA. The mediating effect of service quality and organizational commitment on the effect of management process alignment on higher education performance in Makassar, Indonesia. J Organ Chang Manag. (2018) 31:410–25. doi: 10.1108/JOCM-11-2016-0247.
- [10]. Krishna K. Mohbey et al, An Experimental Survey on Single Linkage Clustering, "International Journal of Computer Applications (0975 – 8887) Volume 76– No.17, August 2013.
- [11]. Manish Verma, Mauly Srivastava, "A Comparative Study of Various Clustering Algorithms in Data Mining", International Journal of Engineering Research and Applications (IJERA), ISSN: 2248-9622, Vol. 2, Issue 3, May-Jun 2012, pp.1379-1384.
- [12]. F. Murtagh, "A Survey of Recent Advances in Hierarchical Clustering Algorithms", The Computer Journal, Vol. 26, No. 4, 1983,