# Hierarchical Cluster Analysis: Comparison of Single linkage,Complete linkage, Average linkage and Centroid Linkage Method

**Research** · February 2020

DOI: 10.13140/RG.2.2.11388.90240

CITATIONS
2

READS
4,703

**1 author:**

Angur Mahmud Jarman
Georgia Southern University
**5** PUBLICATIONS **27** CITATIONS

SEE PROFILE

# Hierarchical Cluster Analysis: Comparison of Single linkage,Complete linkage, Average linkage and Centroid Linkage Method.

Angur Mahmud Jarman

Dept of Computer Science, Georgia Southern University

**Abstract:**

In unsupervised learning to understand the underlying structure of data, hierarchical clustering is widely used technique. The success of hierarchical clustering algorithm depends on the linkage methods used to determine the similarity or dissimilarity between the clusters. This paper reviews the theories of hierarchical clustering analysis, application and focuses on different linkage methods and especially the centroid linkage approach in hierarchical clustering analysis, compares the results, limitations and time complexities. Among the four methods discussed in the paper, three of them has limitations except the average linkage clustering which is the most widely used in hierarchical clustering algorithms.

**Key Words:** Hierarchical, Clustering, Single , Complete, Average, Centroid, Linkage

## Introduction:

Sorting similar items together, comparing them, trying to get some sense out of the properties of those items is very common in our everyday life. Doing such monotonous work manually is a time consuming task. Building some automated methods to group similar items and bringing some order in chaos was necessary for researchers. It is very essential to find similarities between data to come out with statistical meaningful group and those importance have gave birth to the topic of cluster analysis in some disciplines. The primary purpose of cluster analysis is to build a classification system to group cases that are relatively homogeneous within themselves. The group of those homogeneous cases are referred to as a cluster. There are some existing popularly used clustering algorithms yet the choice in algorithms and measurement technique

that specifies to the successive merging of similar cases is a complex process. Clustering algorithm should be chosen after a rigorous research on the particular data set. While one technique or algorithm can successfully classify one type of dataset can fail in another type of data set.

Initial interest in clustering analysis was started in the 1960s and resulted in the development of some new algorithms that showed the possibility of cluster analysis and during this period researchers began to research on various innovative tools in their statistical analysis to uncover the structure in various datasets. In 1963 [1] in the disciplines of biology and ecology cluster analysis was initially used. This technique has been used in social sciences also but did not get widespread popularity. In the 1970s several algorithms were developed for clustering analysis. Clustering analysis became most popular in health and social sciences in the last decades.

In this paper some applications and general technique of hierarchical cluster analysis along with their computational complexity will be described in brief. Three typical distance measure techniques and another centroid based technique, comparison of each with their pros and cons will be described to understand clustering techniques.

**Applications:**

So many years have been spent by biologists to create a taxonomy by hierarchical classification of all living things such as finding the kingdom, family, genus, and species. Much of the early work in cluster analysis was to create a discipline of mathematical taxonomy that could automatically find classification structures in the biodiversity. Large amounts of genetic information that are now available are recently applied by biologist to analyze the groups of genes that have similar functions.

In this new era there are billions of web pages which are source of huge information and are being used to find specific information from a query in the search engines by cluster analysis.

Clustering algorithms are being widely used to return the most relevant results from those billions of pages by grouping these search results into small number of clusters. To predict atmospheric condition and to know whether or not tomorrow is gonna be snowing or raining it is very common to use weather forecast apps which are made based on cluster analysis. It requires finding patterns in the atmosphere and the ocean and cluster analysis is the main tool to find the patterns of atmospheric pressure of popular regions and areas of land and ocean that controls weather impacting on climate. In medical science, clustering analysis techniques brought fortune and speed. Every illness or disease has a number of variations, and cluster analysis can be used to identify these different subcategories. Now a days clustering analysis is being used in higher degree to identify types and pattern of diseases.

Modern businesses including banks, share markets, online shops and social medias are now dependent on clustering algorithms. Businesses collect large amounts of information on current and potential customers from their own platforms or from social media and are being used to segment customers into a small number of groups for additional analysis and marketing activities.

**Hierarchical Cluster Analysis :**

A unique set of nested categories or clusters are produced by sequentially pairing variables or clusters in hierarchical cluster analysis. Starting with the correlation matrix, clusters and unclustered variables at each step tried in all possible pairs, and that pair producing the highest average intercorrelation within the trial cluster is chosen as the new cluster. Like the taxonomic dendrogram graph of the biological systematist, shows the relations between clusters and the value of the clustering criterion associated with each [2]. The complete definition for clustering, however, doesn't come to an agreement, and a classic one is described as follows [3]:

1. Instances, in the same cluster, must be similar as much as possible;

2. Instances, in the different clusters, must be different as much as possible;
3. Measurement for similarity and dissimilarity must be clear and have the practical meaning;

There is no ideal clustering algorithm that solves all the problems out there, but as it was noted, "clustering is in the eye of the beholder." [4] No formula exists that can tell about the most appropriate clustering algorithm for a particular problem. Lot of experimentation and mathematical reasoning needed before choosing a successful clustering algorithm. An algorithm that is designed for one kind of model will generally fail on a data set that contains a radically different kind of model [4].

Given a set of N items to be clustered, and an N square distance (or similarity) matrix, the basic process of hierarchical clustering steps are as follows [5]:

1. Start by assigning each item to its own cluster, so that if you have N items, you now have N clusters, each containing just one item. Let the distances (similarities) between the clusters equal the distances (similarities) between the items they contain.
2. Find the closest (most similar) pair of clusters and merge them into a single cluster, so that now you have one less cluster.
3. Compute distances (similarities) between the new cluster and each of the old clusters.
4. Repeat steps 2 and 3 until all items are clustered into a single cluster of size N.

The core idea of hierarchical clustering is to group objects having more related to nearby objects than to objects farther away. It connects "objects" and forms "clusters" based on distance. In hierarchical clustering the number of clusters changes in every iteration. In divisive clustering the number of clusters increases: all data instances start in one cluster, and splits are performed in each iteration, resulting in a hierarchy of clusters. In agglomerative clustering, the number of clusters decreases in each iteration which is a bottom-up approach: each instance is a cluster at the beginning, and clusters are merged in every iteration.

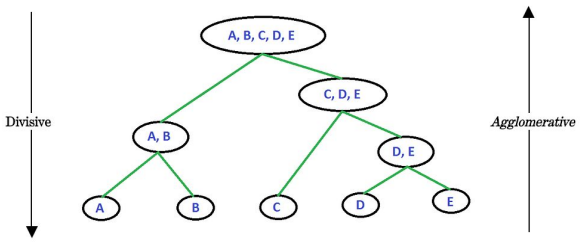



Fig 1: Divisive vs Agglomerative Clustering

**Distance and similarity Measure:**

Cluster analysis is inherently linked to the similarity concept. Determining the right statistic to measure similarity or distance among the cases is the first step that a researcher should take. It might seem that both calculations is the same because it is obvious to observe that if the distance decreases then the similarity increases and vise versa [5]. First important step in hierarchical clustering analysis is to choose the right distance measurement technique between the cases or variables. Distance and similarity measurements are the basis for the process of clustering analysis. Distance is used when the data points are quantitative, on the other hand similarity is used when the data points are qualitative [3].

Most widely used method to calculate the distance between continuous data points or cases is called squared euclidean distance. In this method the distance between all the variables in two cases are calculated and reflected as a single value which then considered as the distance between two whole cases. This method is used in each step in the clustering procedure and is shown by a proximity matrix. The cases who has the calculated smallest distance value among them are merged together and becomes a single case or cluster [8]. That is why a hierarchical clustering technique is time consuming. In a plane with $p1$ at $(x1, y1)$ and $p2$ at $(p2, y2)$, it is

$$\sqrt{(x1 - x2)^2 + (y1 - y2)^2}.$$

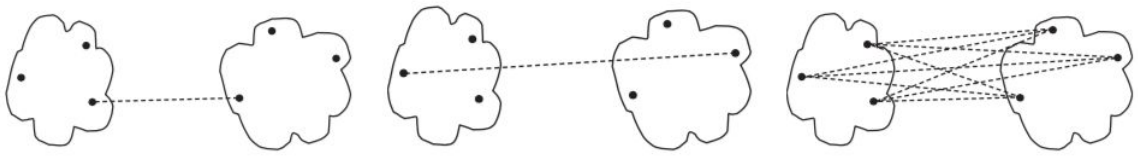
In $N$ dimensions, the Euclidean distance between two points $p$ and $q$ is

$$\sqrt{\sum_{i=1}^{N}(pi - qi)^2}$$

where $pi, qi$ is the coordinate of $p, q$ in dimension $i$.

**Linkage Measure (Single, Complete & Average):**

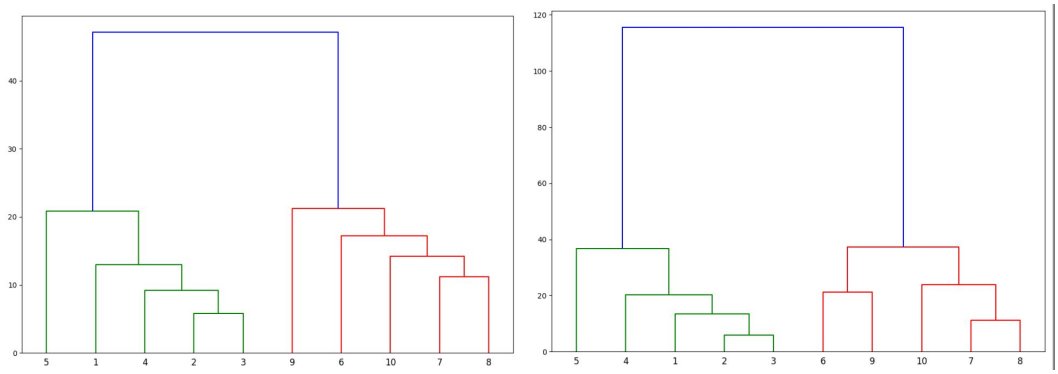
It seems very straightforward to find the euclidean distance between clusters when there is no more than one case in each cluster. However, where there are many cases in each cluster then calculating the euclidean distance between only two cases belongs to each is not sufficient. Since the final goal is to find the clusters that are most similar or nearest among all other clusters, it is important to find the right method of calculating the linkage between two clusters that have more than one cases or data points. Researchers have found many techniques to find that linkage between clusters which are very important to study before starting any hierarchical clustering analysis in a dataset. Each linkage measure technique works in their own way. Whether one linkage measure can work well for a particular type of dataset, it might not work well in another type of dataset. The most commonly known linkage methods are single, complete and average linkage methods. These are all based on the euclidean distance but the main difference between them is the selection of the data points that are considered to take as a final criterion on which the similarity or distance depends. In general, single linkage method depends on the smallest distance between two points where each point belongs to each cluster from the pair of clusters [9]. Complete linkage and average linkage method takes the farthest distance and average of the distances respectively and shown in the following figures [10].



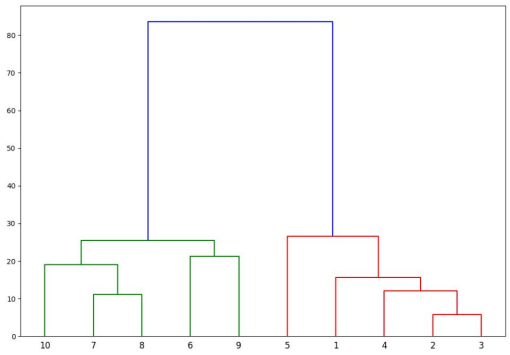
(a) MIN (single link.) (b) MAX (complete link.) (c) Group average.

Figure 2: Linkage criterion of clustering methods

If among the pair of clusters the first one have points p, q, r and the second one has the points w, x, y then the distance method between them in a single link method would be the minimum calculated distance found between the following point pairs: (p,w) (p,x) (p,y) (q,w) (q,x) (q,y) (r,w) (r,x) (r,y). The main problem that remains in the single link clustering is that some clusters may merged together just because one of their data points is closest to another point in another cluster whereas most of the other points resided in significantly larger distance. This is called chaining effect and this effect has a negative impact on the overall result of the cluster analysis if there exists noise in the dataset. Single link technique is used in some data points and the resulting dendrogram is shown in Figure 3(a). In complete link method the distance between them would be the maximum calculated distance found between the following point pairs: (p,w) (p,x) (p,y) (q,w) (q,x) (q,y) (r,w) (r,x) (r,y). Complete link method also can give a disastrous result if noise is present in the dataset. Resulting dendrogram of complete linkage clustering applied in the same example dataset is presented in the following figure 3(b) . In [11] a new method was proposed to overcome the limitation of single link and complete link clustering which is called the average linkage method. The strategy is to take the average of the distance between the point pairs: (p,w) (p,x) (p,y) (q,w) (q,x) (q,y) (r,w) (r,x) (r,y) as the distance between the clusters. This method is widely used in hierarchical clustering analysis. The resulting dendrogram is shown in figure 3(c) after applying the average link method in the same example dataset applied. The height of the dendrogram is the distance between the clusters.




(a)Single linkage (b) Complete linkage




(c): Average linkage

Figure 3: Dendrogram resulting from single, complete and average linkage clustering

**Centroid based linkage approach :**

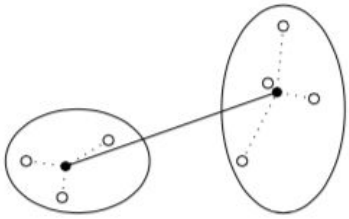
The basic idea of centroid linkage method is to take the distance between the centroids of the data points in clusters. If among the pair of clusters the first one have points p, q, r, s, t and the second one has the points w, x, y, z then to find the distance between the clusters would be the distance between the centroid found for the data points (p,q,r,s,t) and centroid found for the data points (w,x,y,z). Unlike above single, complete and average linkage method, the distance is calculated once rather than between each and every points of the clusters which is shown in figure 4(a). The formula to calculate the centroid of a finite set of k points x1,x2,x3,..xn is straightforward: [12]

$$C = \frac{x1+x2+x3+.....Xn}{k}$$

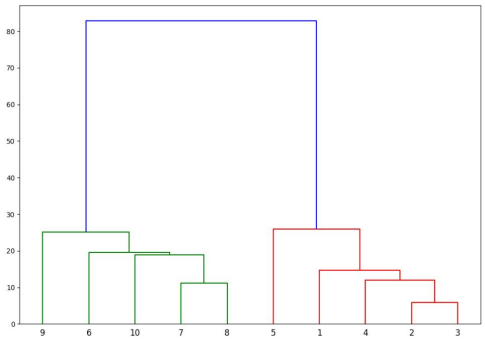
In most cases, the points will be n-dimensional and the centroid should be calculated as taking the points as vertices in a simplex and the formula would be if the n vertices are: v1,v2,v3...vn which are vectors

$$C = \frac{1}{n+1} \sum_{i=0}^{n} V i.$$

Centroid linkage clustering results somewhat similar to average linkage clustering but centroid linkage method has a bad characteristics: possibility of inversion [13] and that's why it is dangerous to use in hierarchical clustering which needs further research. The resulting dendrogram found after applying centroid linkage method is shown in figure 4(b)



4(a)




4(b)

Figure 4: Centroid linkage clustering

If there are N data points in the dataset in cluster analysis, then the time complexity to the above algorithms is shown in table 1.

| Method | Linkage criteria | Time complexity | Comment |
|---|---|---|---|
| Single Link | Smallest distance between two points | $O(n^2)$ | Chaining effect |
| Complete Link | Longest Distance between two points | $O(n^2 logn)$ | Sensitive to noise |
| Average Link | Average Distance of all the data points | $O(n^2 logn)$ | Best choice for most applications |
| Centroid | Distance between Centroid | $O(n^2 logn)$ | Inversion can occur |

Table 1: Time complexity of linkage methods

**Conclusion:**

Hierarchical clustering analysis is a technique that allows researchers to find the meaning and understand the pattern of data from a large unstructured dataset. A theoretical background of Hierarchical clustering, its application and commonly used linkage methods are discussed in this paper. Among the single linkage, complete linkage and average linkage methods only average linkage method is used for most applications because both the single and complete linkage methods results can be greatly affected by noise. Centroid linkage method also works like other linkage methods but has a serious limitation of inversion possibility and that's why it is not used widely in applications.

**References:**


[1] Sokal, R.R. and Sneath, P.H.A. (1963) Principles of Numerical Taxonomy. W.H. Freeman & Co., New York.

[2]Cecil C. Bridges, Jr. (1966). *Hierarchical Cluster Analysis*. SAGE Journals, DOI 10.2466/pr0.1966.18.3.851

[3] Dongkuan XuYingjie Tian (2015) ,"*A Comprehensive Survey of Clustering Algorithms*". Ann. Data. Sci. 2(2):165–193, DOI 10.1007/s40745-015-0040-1

[4] Vladimir Estivill-Castro (2002). "*Why so many clustering algorithms: a position paper*". ACM SIGKDD Explorations Newsletter Homepage archive. Volume 4 Issue 1, pp 65-75

[5] Johnson,S.C. (1967), "*Hierarchical Clustering Schemes*" Psychometrika, 2:241-254.

[6] Jane. Clatworthy Deanna. Buick Matthew. Hankins John. Weinman Robert. Horne (2005), *The use and reporting of cluster analysis in health psychology: A review*. British Journal of health psychology. Volume 10 Issue 3. DOI 10.1348/135910705X25697

[7] William H. E. Day, Herbert Edelsbrunner (1984). *"Efficient algorithms for agglomerative hierarchical clustering methods*". Journal of classification, Volume 1, Issue 1, pp 7–24

[8] Odilia Yim, Kylee T. Ramdeen (2015). "*Hierarchical Cluster Analysis: Comparison of Three Linkage Measures and Application to Psychological Data*". Scientific journal Quantitative Methods for psychology, vol 11 no 1.

[9] Zubrzchi K. Florek, J. Lukaszewiez, J. Perkal, H. Steinhaus, and S. Zubrzchi. (1951) "Sur la liason et la division des points d'un ensemble fini. Colloquium Methematicum, 2".

[10] Mario Mazzocchi (2008).*"Statistics for Marketing and Consumer Research". Chapter 12*. DOI 10.4135/9780857024657.

[10] FrédéricRosa, SergeGuillaume(2019). "*A hierarchical clustering algorithm and an improvement of the single linkage criterion to deal with noise*". Elsevier

[11] Sokal, R.R. and Michener, C.D. (1958) A Statistical Methods for Evaluating Relationships. University of Kansas Science Bulletin, 38, 1409-1448.

[12] Protter, Murray H.; Morrey, Jr., Charles B. (1970), College Calculus with Analytic Geometry (2nd ed.), Reading: Addison-Wesley, LCCN 76087042

[13] https://nlp.stanford.edu/IR-book/html/htmledition/centroid-clustering-1.html