

# **Hierarchical Cluster Analysis**

## Some Basics and Algorithms

Nethra Sambamoorthi  
CRMportals Inc.,  
11 Bartram Road,  
Englishtown, NJ 07726

(NOTE: Please use always the latest copy of the document. [Click on this line for the latest copy of the document](#), if you are connected to the internet to download)

## 1. Introduction

### *1.1 What is cluster analysis?*

Cluster analysis is a collection of statistical methods, which identifies groups of samples that behave similarly or show similar characteristics. In common parlance it is also called look-a-like groups. The simplest mechanism is to partition the samples using measurements that capture similarity or distance between samples. In this way, clusters and groups are interchangeable words. Often in market research studies, cluster analysis is also referred to as a segmentation method. In neural network concepts, clustering method is called unsupervised learning (refers to discovery as against prediction – even discovery in loose sense may be called prediction, but it does not have predefined learning sets to validate the knowledge). Typically in clustering methods, all the samples within a cluster is considered to be equally belonging to the cluster (as against belonging with certain probability). If each observation has its unique probability of belonging to a group(cluster) and the application is interested more about these probabilities than we have to use (binomial) multinomial models.

### *1.2 What is a segmentation method and how is it different or the same as cluster analysis?*

In general terms, segmentation method could be interpreted as a collection of methods, which identifies groups of entities or statistical samples (consumers/customers, markets, organizations, which generally do not have a good application definition to classify entities in one of many groups with out significant errors) that share certain common characteristics such as attitudes, purchase propensities, media habits, and lifestyle etc. The sample characteristics are used to group the samples. Grouping can be arrived at, either hierarchically partitioning the sample or non-hierarchically partitioning the samples. Thus, segmentation methods include probability-based grouping of observations and cluster (grouping) based observations. It includes hierarchical (tree based method – divisive) and non-hierarchical (agglomerative) methods. Segmentation methods are thus very general category of methodology, which includes clustering methods also.

There are three stages to cluster analysis; namely, partitioning/similarity (what defines the groups), interpretation of clusters (how to use groups), and profiling the characteristics of similar/partitioned groups (what explains the groups).

*1.3 The natural questions that need to be attended are:*

What defines inter-object or inter-sample similarity?

What defines a distance between two samples, or clusters?

How to find groups? What are the different methods to identify these groups?

What is the difference between hierarchical and non-hierarchical clustering methods?

When do we stop further and further partitioning or identifying groups (in a divisive approach) and when do we stop further and further joining samples (agglomerative methods)?

What are the right data elements to use in finding clusters in a business problem where we may have hundreds of variables (variable selection method)?

What are the limitations of cluster analysis?

As in many methods common among different areas of study, cluster analysis is also called differently in different specializations. Cluster analysis is called Q-analysis (finding distinct ethnic groups using data about beliefs and feelings<sup>1</sup>), numerical taxonomy (biology), classification analysis (sociology, business, psychology), typology<sup>2</sup> and so on. In marketing it is a very useful technique as will be shown at the end.

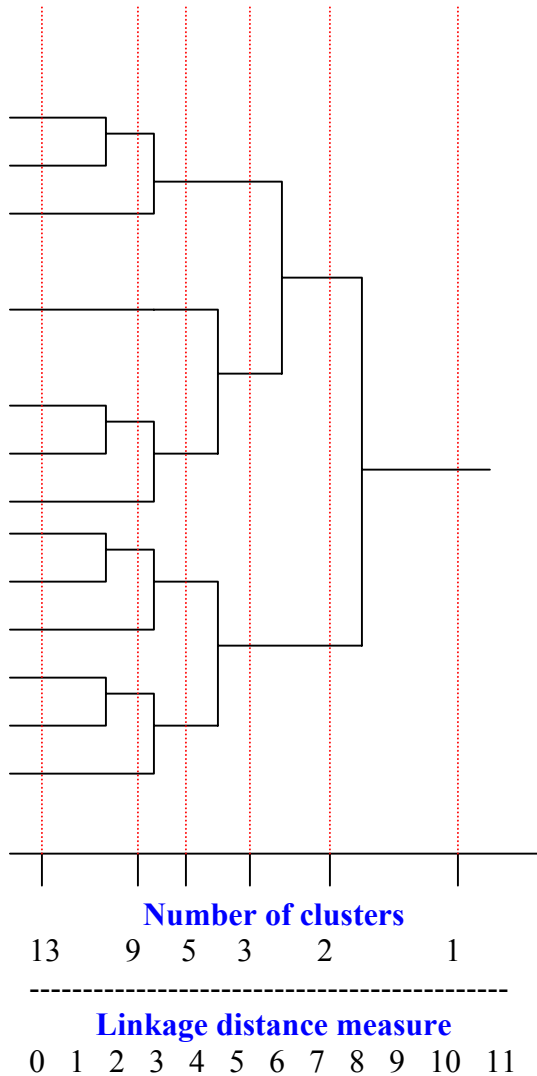
---

<sup>1</sup> [http://www.dodccrp.org/1999CCRTS/pdf\\_files/track\\_3/woodc.pdf](http://www.dodccrp.org/1999CCRTS/pdf_files/track_3/woodc.pdf)

<sup>2</sup> Science of classifying stone tools by form, techniques and technological traits. Must include duplication of the technique by first observing the intentional form, then reconstructing or replicating the tool in the exact order of the aboriginal workman. Shows elements of culture. Typology cannot be based on function." (Crabtree 1982:57)

## 2. Hierarchical and non-hierarchical clustering methods:

The clustering algorithms are broadly classified into two namely hierarchical and non-



hierarchical algorithms. In the hierarchical procedures, we construct a hierarchy or tree-like structure to see the relationship among entities (observations or individuals). In the non-hierarchical method a position in the measurement is taken as central place and distance is measured from such central point (seed). Identifying a right central position is a big challenge and hence non-hierarchical methods are less popular.

### 2.1 Hierarchical Clustering

There is a concept of ordering involved in this approach. The ordering is driven by how many observations could be combined at a time or what determines that the distance is not statistically different from 0 between two observations or two clusters. The clusters could be arrived at either from weeding out

dissimilar observations (divisive method) or joining together similar observations (agglomerative method). Most common statistical packages use agglomerative method and the most popular agglomerative methods are (1) single linkage (nearest neighbor approach), (2) complete linkage (furthest neighbor), (3) average linkage, (4) Ward's method, and (5) Centroid method. All these differ in the definition of distance and what defines largest distance as statistically no-distance or zero-distance. Most of the time, the

distance is based on Euclidean distance in the sample axes (Mahalanobis distance is for non-orthogonal sample).

## *2.2 Points to articulate*

- a) How could clustering methods be used for identifying outlier(s)? – note that outlier(s) by itself will be a cluster. Think of an example of a tree diagram which will point out few outliers; how the grouping pattern and the stem will be represented by a cluster of outliers
- b) The relationship between the linkage distance measure and the number of clusters
- c) Why number of clusters may not be a simple continuously increasing numbers and why for example there may not be one to one relationship between the linkage distance and the number of clusters?
- d) How does variable selection play a role in cluster analysis; what method to use?
- e) How will the above diagram including the measurements and clusters look different if complete linkage measure is used?
- f) Why linkage distance is inversely related to the number of clusters in general?
- g) What happens if similarity measure is used instead of distance measure?
- h) What is meant by similarity or distance measures when we have qualitative data?
- i) What is the major problem with non-hierarchical method? (Hint: start point of the seed or center of the cluster)
- j) Why do we have to standardize the data when we do cluster analysis?
- k) How to use the dendrogram? (tree structure for identifying the distance “between clusters” and what observations belong to which cluster. – a graphical representation issue.) – Hint, the length of the stem and coding the records at the left most stems.
- l) Various factors affect the stability of clustering solution, such as variable selection, distance/ similarity measure used, different significance levels, type of method (divisive vs. agglomerative). Articulate a method that converges to the right solution in the midst of the above mentioned parameters

### 2.3 Simple algorithmic concepts:

**Single linkage:** At each step we agglomerate one additional observation to form a cluster. Thus, first cluster is one with two observations that have the shortest distance. A third observation, which has the next least distance, is added to the two observation cluster to create a three observation cluster or a new two observation cluster is formed. The algorithm continues until all the observations are in one cluster. “The distance between any two clusters is the shortest distance from any point in one cluster to any point in the second cluster. Two clusters are merged at any single stage by the single shortest or strongest link between them”<sup>3</sup>  
Explain: What is meant by snake-like chain of clusters? why does it happen in single linkage methods. This procedure is also known as nearest-neighbour method in the neural network methodologies.

“The distance between any two clusters is the shortest distance from any point in one cluster to any point in the other cluster”<sup>3</sup>.

**Complete linkage:** This is similar to single linkage except that this is based on maximum distance not minimum distance. The maximum distance between any two individuals in a cluster represents the smallest (minimum diameter) sphere that can enclose the cluster<sup>3</sup>. **The advantage here is that this does not create one cluster for “chain observations”. This happens in single linkage distance where the whole collection of data becomes a cluster, though the first and the last observation will be at the maximum distance for the entire sample space!.**

**Average linkage:** Here we use the average distance from samples in one cluster to samples in other clusters.

**Ward Distance:** This method is distinct from all other methods because it uses an analysis of variance approach to evaluate the distances between clusters. In short, this method attempts to minimize the Sum of Squares (SS) of any two

---

<sup>3</sup> J.F. Hair, Jr., Rolph E. Anderson, and Ronald L. Tatham, Multivariate Data Analysis with Readings, 2<sup>nd</sup> Edition, MacMillan Publishing Co., New York, PP 303

(hypothetical) clusters that can be formed at each step<sup>4</sup>. Typical of properties of variance for statistical decision-making, this tends to create too many clusters or clusters of small sizes because the more the observations scattered, the sum of squares makes the distance bigger.

**Centroid (Mean) Method:** Here Euclidean distance measured between centroids of two clusters.

measure each observation's distance from the other observations  
 $S_i$  is the distance between  $i$ th observation with the rest of the observations

	s1	s2	s3	...	s25
1					
2					
3					
4					
5					
6					
7					
8					
9					
10					
11					
12					
13					
14					
15					
16					
17					
18					
19					
20					
21					
22					
23					
24					
25					

SAMPLE

There are 25 obs (samples)  
 Diagonal elements have zero values and the matrix is symmetric by property of distance  
 The least distance in the who matrix identifies the first cluster of two elements if the that distance qualifies as a significant distance from zero.  
 Next consider the next least distance measure.  
 Join this with the first cluster(min of 2 distances with the new third obs) if its distance is statistically same as zero or this might form a new 2 individual cluster.  
 Thus given a distance, we will be able to form clusters. Now increase the distance measure and see how the different clusters are changing. Since increasing the distance should mean less and less number of cluster, the algorithm comes to the end when all the observations are in one cluster.  
 In the process we create the dedrogram (tree-based graphics representing the clusters)

<sup>4</sup> <http://www.statsoftinc.com/textbook/stcluan.html>

### *2.3 An Example:*

Suppose a credit card company wants to introduce different incentive schemes for card usage for different groups of merchants whose marketing strategies and positioning the card can play a significant role in promoting the card. It is a non-trivial problem, especially if many of the merchants sell many things and are not correlated very well with the SIC code. Plus what matters is outcome variable of profitability and revenue rather than type of product produced or sold. Thus, grouping them based on SIC code cannot reveal and may not serve the purpose of providing incentives to the merchants, which is to increase usage of the card products across whole collection of merchandize in a store. So the decision variable is increased card usage. The data elements are store related information. There could be hundreds of variables. Now cluster (group) the stores so that we will all these merchants grouped in to say around 10 actionable groups so that a new incentive system could be developed.


### *Think and Articulate:*

1. Redefine a problem where we need to use card users as groups, not the merchandisers
2. How is this grouping different from a traditional scoring model for the merchandisers or consumers, assuming share of card or share of wallet accordingly as the decision variable (as against yes or no type decision variable)?



## Appendix

Consider the following nice graphics<sup>5</sup>.



Do we know how to group these people? Do similar groups for an outcome exist? How many groups are there?

We need observations to group these people. Typically we collect observations about these people which do not change. (Think why we may collect and under what conditions it is good to use such data that changes over time)

A picture is worth thousand words and our translation of these pictures in to data not only should capture all the information about these people but also such data elements should be correlated to a relevant outcome variable

<sup>5</sup> <http://obelia.jde.aca.mmu.ac.uk/multivar/ca.htm>

