



Exploring Algorithms and Mechanics of Fuzzy and Hard Clustering in Universal Cluster Analysis

ABSTRACT

A cluster is a collection of data objects which can be based on similarity to one another within the same cluster or dissimilar to the objects in other clusters. Therefore, cluster analysis refers to the grouping a set of data objects into clusters. It involves organizing data into classes such that there is both high intra class similarity as well as low inter-class similarity. Cluster analysis is a group of multivariate techniques whose purpose is to group objects (for example, test items, respondents, products, or other entities) based on the characteristics they possess. It is a means of grouping such objects based upon attributes that make them similar. If plotted geometrically, the objects within the clusters will be close together, while the distance between clusters will be farther apart. Statisticians conceptualize clustering as unsupervised classification where there are no predefined classes; psychologists conceptualize it as sorting by psychologists while marketing experts consider it as segmentation. It is the algorithmic process of finding natural groupings among objects. Cluster variate represents a mathematical representation of the selected set of variables which compares the object's similarities. Clustering may specify the relationship of the clusters to each other, for example, a hierarchy of clusters embedded in each other. There are two types of clustering thus: (1) Soft clustering (which is also known as fuzzy clustering) in which each object belongs to each cluster to a certain degree (for instance, a likelihood of belonging to the cluster) and, (2) Hard clustering in which each object belongs to a cluster or not. The paper explores the algorithms and mechanics of carrying out clustering in both hard and soft clustering.

Kimani Chege Gabriel

Morendat Institute of Oil & Gas

Corresponding author:

Email: kimani.chege@kpc.co.ke



Conceptual issues

Cluster analysis refers to the grouping a set of data objects into clusters. It involves organizing data into classes such that there is both high intra class similarity as well as low inter-class similarity. Cluster analysis is a group of multivariate techniques whose purpose is to group objects (for example, test items, respondents, products, or other entities) based on the characteristics they possess. It is a means of grouping such objects based upon attributes that make them similar.

Cluster analysis compared to discriminant analysis

In discriminant analysis the expert should know group membership for the cases used to derive classification rules while in cluster analysis, the expert doesn't know the number of groups, does not know who or what belongs to which group.

Cluster Analysis compared to Factor Analysis

According to Aldenderfer and Blashfield (1984), cluster analysis is the grouping of objects based on the distance (proximity) while factor analysis refers to grouping that is based on patterns of variation and correlation. Typically, in factor analysis, the expert forms a group of variables based on the several examinee's responses to those variables while in contrast to Cluster analysis, we group examinees based on their responses to several variables.



Criteria for good clustering

This includes the following:

- i. A good clustering method will produce high quality clusters which have (a) low inter-class similarity and, (b) high intra-class similarity.
- ii. The quality of a clustering method is also measured by its ability to discover some or all of the hidden patterns and take them into account during formation of clusters.
- iii. The quality of a clustering result depends on both the similarity measure used by the method and its implementation.

Common applications of clustering

We ought to mention that clustering can be applied either as (a) As a stand-alone tool to get insight into data distribution or as (b) as a preprocessing step for other algorithms.

The specific applications include (but are not limited to) the following:

Psychographic profiles: it is applied in psychographics which is a qualitative methodology used to describe consumers based on their psychological attributes. A psychographic profile is a complete profile of a person or group's psychographic make-up which is constructed using clustering techniques. These psychographic profiles are used in the design of advertisements and the process of market segmentation during which the clustering of categories of psychographic factors used in market segmentation such as activity, interest, opinion (AIOs), attitudes, values and behavior is carried out.

Applications in Psychiatry involve the application of cluster analysis in the characterization of patients on the basis of clusters of symptoms which in turn is useful in the identification of an appropriate form of therapy to be adopted in the management of the patient's condition.

In Biology, cluster analysis is used to find groups of genes that have similar functions. This involves investigating the taxonomy of species being considered. Geneticists can collect a dataset of different animals and note different attributes of their phenotypes (as influenced by the genotype). Thereafter, cluster analysis is applied in grouping those phenotypic observations into a series of clusters and from this the expert builds a taxonomy of groups and subgroups of similar animals. In Medicine, we use clustering to identify the diagnostic clusters. For this purpose, a diagnostic questionnaire has to be devised in such a manner that it includes possible symptoms (for example, in psychology, anxiety, depression etc.) and then cluster analysis can help to identify groups of patients that have similar symptoms.

Climate applications – clustering is employed to facilitate understanding the Earth's climate which requires finding patterns in the atmosphere and ocean. Specifically, cluster analysis is applied to find patterns in the atmospheric pressure of Polar Regions and areas of the ocean that have a significant impact on land climate.

Information retrieval also utilizes clustering. For instance, the World Wide Web consists of billions of web pages, and when a query is made to any of the search engines, the results can return thousands of pages. To make the search meaningful, clustering can be used to group these search results into small number of clusters, each of which captures a particular aspect (key words etc.) of the query.

In **Education**, researchers may measure psychological, aptitude, and achievement characteristics of learners then apply cluster analysis to identify what homogeneous groups exist among students (for instance, students that excel in certain subjects but fail in others, learners with special needs, exceptional ones or high achievers in all subjects) and prescribe relevant educational provisions in terms of materials, programs and processes.

TYPICAL CLUSTER MODELS

According to Bailey (1994) typical cluster models include:

1. Centroid models are clustering models which represent each cluster by a single mean vector, for example, the k-means algorithm.
2. Group models are models whose algorithms do not provide a refined model for their results and so they just provide the grouping information.
3. Connectivity models are clustering models which builds models based on distance connectivity, for example, hierarchical clustering.
4. Distribution models: here, clusters are modeled using statistical distributions, such as multivariate normal distributions used by the expectation-maximization algorithm.
5. Density models are models which define clusters as connected dense regions in the data space.
6. Subspace models which are also known as two-mode-clustering or co-clustering are applied in biclustering in which clusters are modeled with both cluster members and relevant attributes.
7. Graph-based models – in these models, a clique, that is, a subset of nodes in a graph such that every two nodes in the subset are connected by an edge can be considered as a prototypical form of cluster. An example is the Highly Connected Subgraphs (HCS) clustering algorithm where it is possible for us to relax the complete connectivity requirement and therefore a fraction of the edges can be missing (i.e. quasi-cliques).
8. In the Signed graph models, every path in a signed graph has a sign from the product of the signs on the edges. Edges may change sign and result in a bifurcated graph based on the assumptions of balance theory. The weaker "clusterability axiom" (no cycle has exactly one negative edge) yields results with more than two clusters, or sub-graphs with only positive edges.



9. Neural models: are models that can be characterized as similar to one or more of the above models, including subspace models when neural networks implement a form of Principal Component Analysis or Independent Component Analysis which normally operate as unsupervised neural network is the self-organizing map (Agrawal, Gehrke, Gunopulos, and Raghavan, 1998).

COMMON ROLES THAT CLUSTER ANALYSIS CAN PLAY

Generally speaking, cluster analysis can play the following roles:

- Data reduction**
A psychometrician may be faced with a large number of observations and data that can be meaningless unless classified into manageable groups. Cluster analysis can perform this data reduction procedure objectively by reducing the information from an entire population (or sample) to information about specific groups.
- Hypothesis generation**
Cluster analysis is also useful when a psychometrician wishes to develop hypotheses concerning the nature of the data or to examine hypotheses that were stated previously.

Objectives of cluster analysis

Cluster analysis used with the objective of carrying out:

- Taxonomy description** that is identifying groups within the research, experimental or psychometric data
- Data simplification** which is the ability to analyze groups of similar observations instead of all individual observations and,
- Relationship Identification** - The simplified structure from cluster analysis portrays relationships not revealed otherwise.

It is important to note that theoretical, conceptual and practical considerations must be observed when selecting clustering variables for cluster analysis: (i) Only variables that relate specifically to objectives of the cluster analysis are included and, (ii) variables selected characterize the individuals (objects) being clustered

Measuring the Quality of Clustering

The following are quality indicators in cluster analysis:

- The similarity/dissimilarity metric which is sometimes expressed in terms of a distance function, which is typically metric: $d(i, j)$
- There is (apart from the cluster analysis model) a separate 'quality' function that measures the 'goodness' of a cluster that has been constructed or developed.
- The definitions of distance functions are usually very different for interval-scaled, Boolean, categorical, and ordinal variables.
- Cluster analysis weights should be associated with different variables based on data semantics and applications.
- Cluster analysis is usually very subjective and therefore it can be quite hard to define what is 'good enough' or 'similar enough'.

How cluster analysis works

As mentioned earlier, the goal of cluster analysis is to define the structure of the data or objects by placing the most similar observations into groups. In order for this task to be accomplished, we must first answer the following three fundamental questions:

- How can we measure similarity among the observations in the data sets?
- How do we form clusters that are good enough?
- How many groups do we form?

Let us then descriptively answer those questions here below.

1. Measuring similarity among the observations

The question at this point is, what measure of inter-object similarity is to be used and how is each variable to be 'weighted' in the construction of such a summary measure? Here, similarity represents the degree of correspondence among objects across all of the characteristics used in the analysis. It is in essence a set of rules that serve as criteria for grouping or separating items and it can be established using any of the following methods:

(a) Correlational measures - which involves the computation of correlation coefficients between objects where large values (and statistically significant coefficients) of r 's indicate similarity. However, this method is not used frequently.

(b) Distance Measures.

Distance measures (determination of $d(i, j)$) is the most often used as a measure of similarity, with higher values representing greater dissimilarity (distance between cases) as opposed to similarity between objects.

Two distance graphs may have the same $r = 1$, which implies to have a same pattern but the distances (d 's) may not be equal.

Several distance measures are available and each method has specific characteristics. They include:

i. The Euclidean distance which is recognized as the straight-line distance and it is the most commonly used distance measure. In this method, the distance between two objects is given by the formula:

$$d_{\text{Euclidean}}(B, C) = \sqrt{(x_B - x_C)^2 + (y_B - y_C)^2}$$

In other notations, the Euclidean distance function measures the 'as-the-crow-flies' distance and the simpler version of the formula for this distance between a point $X (X_1, X_2, \text{etc.})$ and a point $Y (Y_1, Y_2, \text{etc.})$ is:

$$d = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$$

ii. Squared Euclidean distance which is a metric that uses the same equation as the Euclidean distance metric, but does not take the square root. Therefore clustering with the Euclidean Squared distance metric is faster than clustering with the regular Euclidean distance. The consequences of using this metric is that it affects the output of hierarchical clustering which is likely to change but the output of K-Means and the Jarvis-Patrick clustering is not affected if Euclidean distance is replaced with Euclidean squared.

iii. Manhattan (city- block) distance. This Euclidean distance uses the sum of the variables' absolute differences. The Manhattan distance function calculates the distance that would be traveled to get from one data point to the other if a grid-like path is followed. The Manhattan distance between two items is the sum of the differences of their corresponding components. The formula for this distance between a point $X = (X_1, X_2, \text{etc.})$ and a point $Y = (Y_1, Y_2, \text{etc.})$ is:

$$d = \sum_{i=1}^n |x_i - y_i|$$

Alternatively we can say:

$$d_{\text{City-block}}(B, C) = |x_B - x_C| + |y_B - y_C|$$

The diagram below illustrates the comparative 2-D paths for the Manhattan distance and Euclidean distance

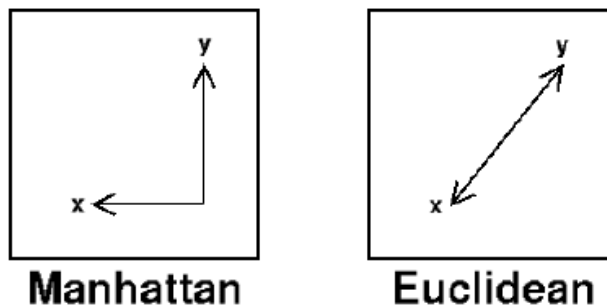


Figure 32: Comparative 2-D paths for the Manhattan distance and Euclidean distance

iv. Chebychev distance. This is the maximum of the absolute difference in the clustering variables' values and it is frequently used when working with metric (or ordinal) data. The distance is computed using the formula:

$$d_{\text{Chebychev}}(B, C) = \max(|x_B - x_C|, |y_B - y_C|)$$

v. Mahalanobis distance (D_2 or d_m) is a generalized distance measure that accounts for the correlations among variables in a way that weights each variables equally.

The equation below shows the Mahalanobis distance for a two dimensional vector with **no covariance**.

$$d(x, c) = \sqrt{\frac{(x_1 - c_1)^2}{\sigma_1^2} + \frac{(x_2 - c_2)^2}{\sigma_2^2}}$$

However, the general equation for the Mahalanobis distance uses the full covariance matrix, which includes the covariances between the vector components. Below is the general equation for the Mahalanobis distance between two vectors, x and y , where S is the covariance matrix.

$$d_M(x, y) = \sqrt{(x - y)^T S^{-1} (x - y)}$$

The computation and the measures of the distance between objects is expected to reflect the following properties:

Symmetry i.e. $D(A, B) = D(B, A)$ Constancy of Self-Similarity i.e. $D(A, A) = 0$ Positivity (Separation) i.e. $D(A, B) = 0$ If $A = B$ Triangular Inequality i.e. $D(A, B) \leq D(A, C) + D(B, C)$ From the above properties we can derive the following intuitions $D(A, B) = D(B, A)$ that is Symmetry Otherwise one could claim 'Logan looks like Merc, but Merc looks nothing like Logan.' $D(A, A) = 0$ that is Constancy of self-similarity otherwise one could claim 'Logan looks more like Merc, than Merc does.' $D(A, B) = 0$ If $A = B$ that is Positivity (Separation) otherwise there are objects in our world that are different, but we cannot tell apart $D(A, B) \leq D(A, C) + D(B, C)$ that is Triangular Inequality. Otherwise we could claim 'Logan is very like Merc, and Logan is very like Carl, but Merc is very unlike Carl.' The following diagram illustrates three distance metrics that were described earlier.

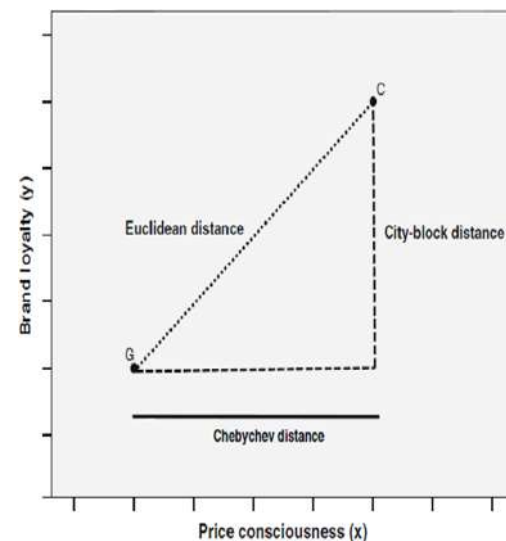


Figure 33: Three Euclidean distance metrics in cluster analysis



2. Formation of clusters (second question)

The second question is, after inter-subject similarities are obtained, how are the classes formed? This addresses the issue of class formation. Stated differently, what summary measures of each cluster are appropriate in a descriptive sense; that is, how are the clusters to be defined?

Normally one should identify the two most similar (closest) observations not already in the same cluster and combine them. This rule should be applied repeatedly to generate a number of cluster solutions, starting with each observation as its own 'cluster' and then combining two clusters at a time until all observations are in a single cluster. This process is a hierarchical procedure because it moves in a stepwise fashion to form an entire range of cluster solutions. It can also be viewed as an agglomerative method because clusters are formed by combining existing clusters (Theodoridis and Koutroubas 1999).

3. The number of groups to be formed (Third question)

The decision about the number of groups to be formed depends on the **Overall Similarity Measure** (Average Within-Cluster Distance) OSM changes in the output table of the agglomerative process and cluster solution.

We scrutinize the cluster solution to identify the step that has the most appropriate cluster solution which possibly gives us a number of equally sized clusters and few outlying observations. This approach is particularly useful in identifying outliers and it also depicts the relative size of varying clusters, although it becomes unwieldy when the number of observations increases.

GRAPHICAL PORTRAYS OF THE EXTRACTED GROUPS OR CLUSTERS

A. Nested groupings

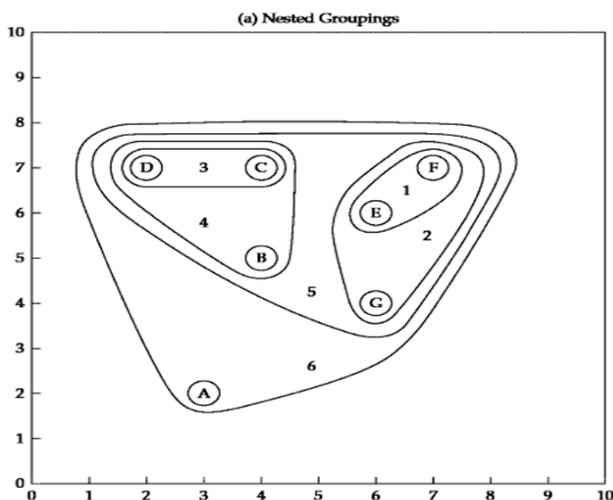


Figure 34: Graphical portrayal of nested groupings

A. The dendrogram

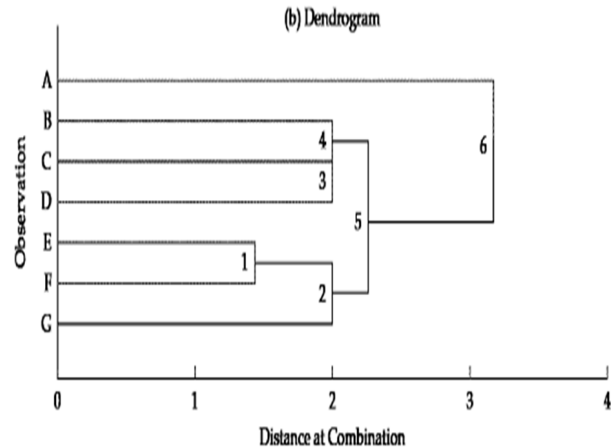


Figure 35: Graphical portrayal of a dendrogram

The dendrogram is a graphical representation (tree graph) of the results of a hierarchical procedure in clustering. Starting with each object as a separate cluster, the dendrogram shows graphically how the clusters are combined at each step of the procedure until all are contained in a single cluster.

Managing outliers during clustering (Remove or to retain them)

It is important to deal with outliers because they can severely distort the representativeness of the results if they appear as structure (clusters) inconsistent with the objectives. They can either be removed or retained under the following circumstances:

- (a) They should be removed if the outliers represent: Abberant observations that are not representative of the population, or, Observations of small or insignificant segments within the population and of no interest to the research objectives
- (b) Outliers should be retained if there is under-sampling or a poor representation of relevant groups in the population. A good practice is to augmented the sample in order to ensure representation of these groups.

How are outliers identified or detected during clustering?

We use similarity measures to identify the outliers by:

- (a) Identifying or finding the objects or observations that have large distances from all other observations.
- (b) Graphic profile diagrams which highlight all the outlying cases (we can spot them from the graphs as points falling away from the cluster).
- (c) Their appearance in cluster solutions as single – member or small clusters.

Sample size concerns in cluster analysis

The psychometrician should ensure that the sample size is large enough to provide sufficient representation of all relevant groups of the population. The relevant sample size calculation formulas should be applied as recommended for highly quantitative work such as cluster analysis. The psychometrician must therefore be confident that the obtained sample is representative of the population.

Standardizing cluster analysis data

Fayyad, Piatetsky-Shapiro, Smuth and Uthurusamy (2008) posit that in the operations of clustering it is important to consider data standardization in managing the effects of scales. Clustering variables that have scales using widely differing numbers of scale points or that exhibit large differences in standard deviations should be standardized. The most common standardization conversion is Z score (with mean equals to 0 and standard deviation of 1).

DERIVING CLUSTERS (CLUSTER ANALYSIS METHODS)

There are different methods that can be used to derive clusters in the process of carrying out a cluster analysis; these methods can be classified as follows:

- Partitioning algorithms (Non-hierarchical clustering) which usually construct various partitions and then evaluate them by some criterion.
- Hierarchy algorithms on the other hand create a hierarchical decomposition of the set of data (or objects) using some criterion.
- Grid-based are a clustering approach based on a multiple-level granularity structure
- Density-based approach to cluster analysis is based on connectivity and density functions
- Model-based clustering approach involves hypothesizing a model for each of the clusters and the idea is to find the best fit of that model to each other.

The most commonly used clustering approaches include:

- Partitional algorithms or Non-hierarchical Cluster Analysis
- Hierarchical Cluster Analysis
- Combination of Both Methods

Presented here-below is a description of the said approaches that are commonly used.

1. HIERARCHICAL ALGORITHMS OR HIERARCHICAL CLUSTER ANALYSIS

Hierarchical algorithms usually create a hierarchical decomposition of the set of objects using some criterion. HCA are preferred when the sample size is moderate (under 300 – 400,

not exceeding 1000). Hierarchical clustering is a set of nested clusters that are organized as a tree as shown in the diagram below.

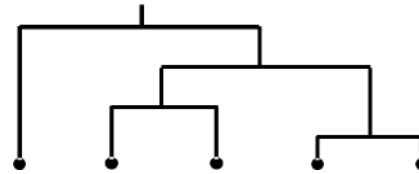


Figure 36: Hierarchical nested clusters organized as a tree

The size of the dendrograms is determined in terms of the number of its leaves. The number of dendrograms with n leaves = $(2n-3)! / [(2(n-2)) (n-2)!]$. Since we cannot test all possible trees we will have to heuristic search of all possible trees.

There are two basic forms of Hierarchical algorithms (HCA), that is (a) Agglomerative algorithm and, (b) the Divisive algorithm. An algorithm refers to a set of rules or Starting with each item in its own cluster, protocol that defines how similarity is defined between multiple – member clusters in the clustering process. A brief description of the two basic forms of HCA algorithms is as follows:

Agglomerative algorithm (Bottom-Up algorithm): this algorithm involves starting with each object point (with each item in its own cluster) as individual clusters and, at each step, find the best pair to merge into a new cluster and eventually merge the closest pair of clusters. This should be repeated until all clusters are fused together.

Divisive algorithm (Top-Down algorithm): Starting with all the data in a single cluster, consider every possible way to divide the cluster into two and then choose the best division and recursively operate on both sides. In other words, one should start with one, all-inclusive cluster and, at each step, split a cluster until only singleton clusters of individual points remain. The dendrogram below illustrates this.

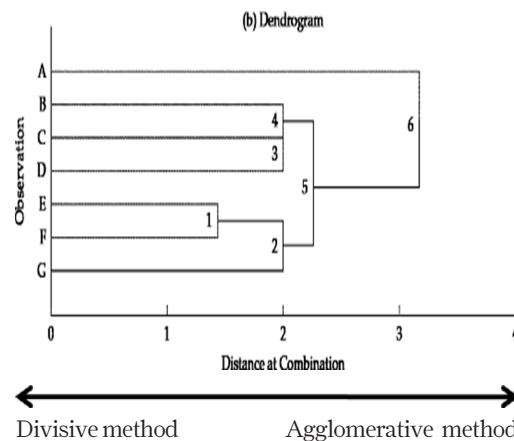




Figure 37: Dendrogram illustrating divisive algorithm

Note: The Divisive Hierarchical clustering Technique is not much used in the real world and that is why not much discursive attention will be given to it in this book.

Agglomerative algorithm of Hierarchical algorithms (HCA) There are numerous clustering approaches under the agglomerative HCA but the five most popular agglomerative algorithms are:

i. Single – Linkage type of agglomerative HCA is also known as the nearest – neighbor method and it defines similarity between clusters as the shortest distance from any object in one cluster to any object in the other. Operationally it is based on grouping clusters in bottom-up at each step combining two clusters that contain the closest pair of elements not yet belonging to the same cluster as each other. This is illustrated in the figure below.

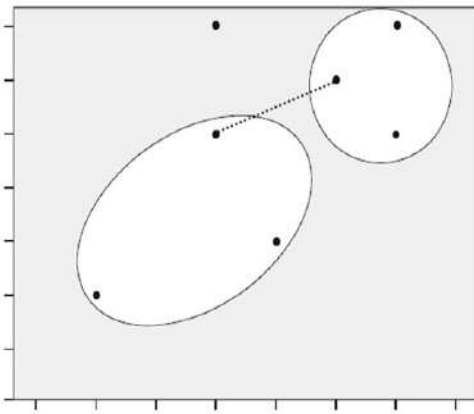


Figure 38: Single – Linkage type of agglomerative HCA

ii. Complete – Linkage algorithm (or the farthest – neighbor method). In this method of hierarchical clustering, we merge in the members of the clusters in each step, which provide the smallest maximum pairwise distance. It is based on the assumption that the distance between two clusters is based on the maximum distance between any two members in the two clusters. This is illustrated by the figure below.

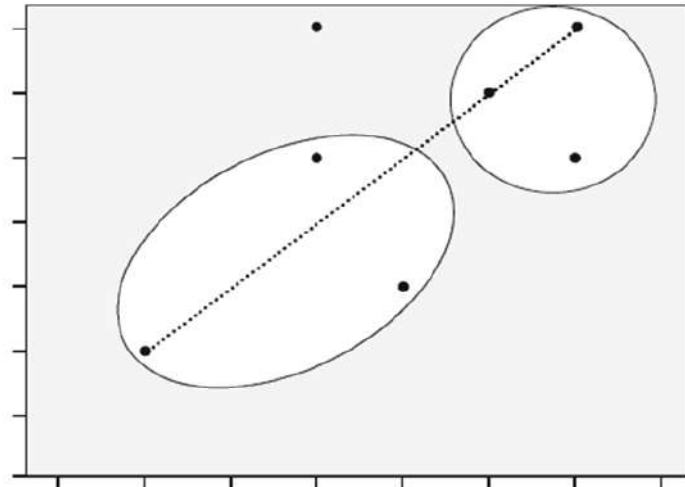


Figure 39: Complete – Linkage algorithm in HCA

iii. Average – Linkage method (i.e. Average between groups) is sometimes referred to as Unweighted Pair Group Method with Arithmetic Mean (UPGMA) in which the distance between two clusters is calculated as the average distance between all pairs of subjects in the two clusters. Most psychometricians consider this as a fairly robust method and it is illustrated by the figure below.

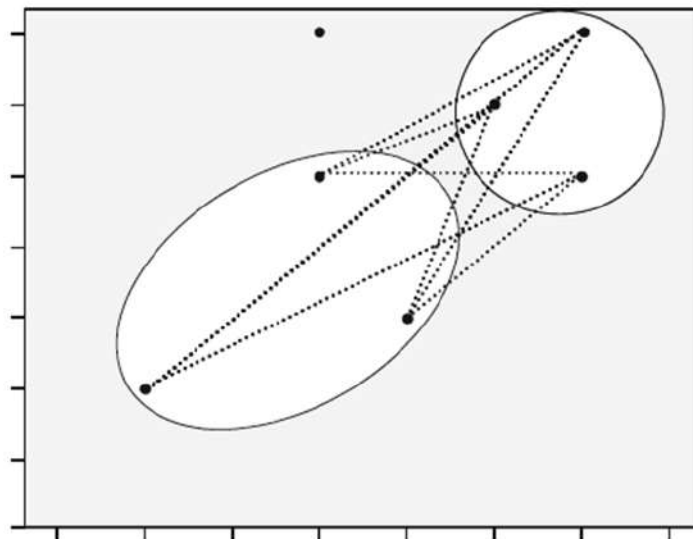


Figure 40: Average – Linkage method algorithm in HCA

iv. Centroid Method. Cluster Centroids – are the mean values of the observation on the variables of the cluster. In this type of agglomerative

HCA, the centroid (mean value for each variable) of each cluster is calculated and the distance between centroids is used. Practically, the distance between the two clusters equals the distance between the two centroids. The clustering action involves merging clusters whose centroids are closest together. This is illustrated by the figure below.

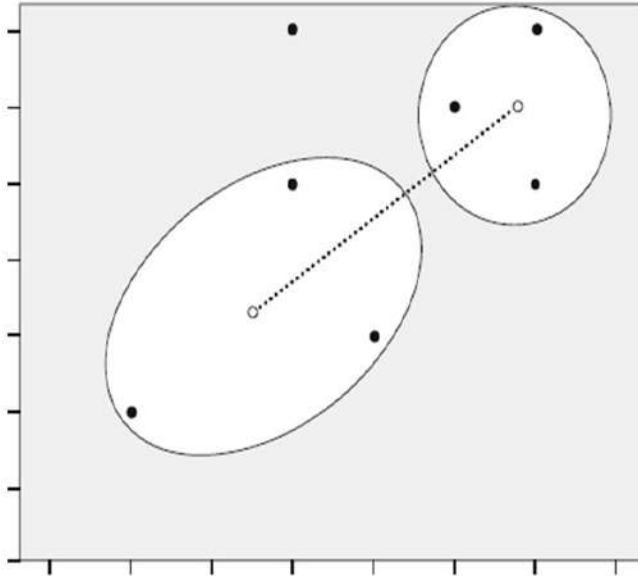


Figure 41: Centroid cluster Method

v. Ward's Method

Ward's method (also known as Ward's Minimum Variance Clustering Method or the Minimum variance method) is an alternative to single-link clustering. In this method all possible pairs of clusters are combined and the sum of the squared distances within each cluster is calculated which is then summed over all clusters. The combination that gives the lowest sum of squares is chosen and as a result, the method produces clusters of nearly equal size which is not always desirable. The method is quite sensitive to outliers.

Ward's method starts with n clusters, each containing a single object and these n clusters are combined to make one cluster containing all objects (MacQueen, 1967). At each step, the process makes a new cluster that minimizes variance, measured by the sum of squares index which is denoted by E . At each step, the following algorithmic calculations are made to find E :

- Find the mean of each cluster.
- Calculate the distance between each object in a particular cluster, and that cluster's mean.
- Square the differences from Step ii.
- Sum (add up) the squared values from Step iii.
- Add up all the sums of squares from Step iv.

Every possible combination of clusters must be considered in order to select a new cluster at each step.

Summary of hierarchical cluster analysis

The Hierarchical Cluster Analysis provides an excellent

framework with which to compare any set of cluster solutions. This method helps in judging how many clusters should be retained or considered. In summary, in Hierarchical Cluster Analysis

- There is usually no need to specify the number of clusters in advance.
- The hierarchical nature of the analysis maps perfectly well onto human intuition for some domains
- Hierarchical Cluster Analysis approaches do not scale well considering time complexity of at least $O(n^2)$, where n is the number of total objects.
- Like any heuristic search algorithms, local optima in Hierarchical Cluster Analysis are a problem.
- In Hierarchical Cluster Analysis, the interpretation of results is (very) subjective.

Advantages of Hierarchical Cluster Analysis

- Speed. Hierarchical Cluster Analysis have the advantage of generating an entire set of clustering solutions in an expedient manner.
- Measures of Similarity. Hierarchical Cluster Analysis can be applied to almost any type of research question.
- Simplicity. With the development of Dendrogram, the Hierarchical Cluster Analysis so afford the researcher with a simple, yet comprehensive portrayal of clustering solutions.

Disadvantages of Hierarchical Cluster Analysis

- Hierarchical Cluster Analysis is not amenable to analyze large samples.
- To reduce the impact of outliers, the researcher may wish to cluster analyze the data several times, each time deleting problem observations or outliers.

2. PARTITIONAL ALGORITHMS OR NON-HIERARCHICAL CLUSTER ANALYSIS (NHCA)

These algorithms are simply a division of the set of data objects into non-overlapping subsets (clusters) such that each data object is in exactly one subset. It involves the construction of various partitions and then evaluating them based on some criterion (Everitt, Landau and Leese, 2001). The partitioned clusters resulting are as illustrated in the figure below where each box represents a distinct cluster.

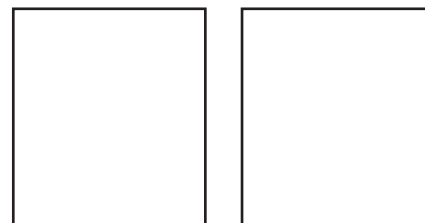


Figure 42: Non-overlapping or partitioned clusters



In contrast to Hierarchical Method, the NHCA do not involve the treelike construction process. Instead, these algorithms assign data objects into clusters once the number of clusters is specified.

Nonhierarchical cluster analysis (NHCA) is executed in two steps thus:

- 1.) We first specify Cluster Seed – that is, the expert identifies the starting points
- 2.) Assignment follows where we assign each observation to one of the cluster seeds.

The Nonhierarchical cluster analysis (NHCA) algorithm includes:

- i. Sequential Threshold Method
- ii. Parallel Threshold Method
- iii. Optimizing Procedures

All of the above algorithms belong to a group of clustering algorithm known as K – means. The name K-means as used for Nonhierarchical cluster analysis derives from the fact that in each instance the data objects or observations are placed in exactly one of K non-overlapping clusters. In NHCA the expert has to input the desired number of clusters since only one set of clusters is output.

K - means clustering method

The term K - means is a term that is used commonly is used by some experts to refer to Nonhierarchical cluster analysis in general. Operationally, the K-means algorithm assigns each point to the cluster whose center (this center is called the centroid) is nearest. The center is the average of all the points in the cluster – that is, its coordinates are the arithmetic mean for each dimension separately over all the points in the cluster.

Algorithms of the K-means clustering

K-means clustering is a centroid-based clustering method that identifies a specified number of non-overlapping clusters within data objects or observations. It requires the psychometrician to pre-specify the number of clusters and then place each examinee into one of them. It is

important to note that the actual profile (i.e., means on the variables used to cluster) of the clusters is not pre-specified, but only the number is (Jain, Murty and Flynn, 1999).

The K-means clustering algorithm is based on the following steps.

1. The psychometrician indicates the number of clusters.
 2. In step two, the initial cluster centroids are formed either through pre-specification of cluster centroids by the psychometrician or by using random selection for the K clusters.
 3. The squared Euclidean distance (ESS) is calculated based on the current cluster solution.
 4. Each individual (or observations or objects) is reassigned to the cluster to whose centroid it is closest.
 5. The cluster centroids are updated after each reassignment.
 6. Steps 3–5 are repeated until no further reassignment of individuals (or observations or objects) to clusters takes place, i.e., each individual is in the cluster with the nearest centroid.
- This method tries to minimize the sum of the within-cluster variances.

$$V_K = \sum_{k=1}^K \sum_{i=1}^n \delta_{ik} m_i d^2 (x_i - \bar{x}_k)$$

The indicator function δ_{ik} equals 1 if the observation x_i comes from cluster k , or 0 otherwise. We also must reckon that the element $\bar{x}_{kj}$ of the vector $\bar{x}_k$ is the mean value of the variable j in the cluster k :

$$\bar{x}_{kj} = \frac{1}{n_k} \sum_{i=1}^I \delta_{ik} m_i x_{ij}$$

We denote the mass of the cluster k with n_k , which is equal to the sum of the masses of all observations belonging to the cluster k

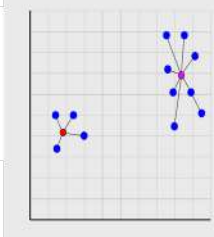
The above criterion of the K-means clustering method can be derived in a more direct way using the Maximum Likelihood approach assuming that there is normality and independence in

the testing populations.

Illustration: The data set has three dimensions and the cluster has two points: $X = (x_1, x_2, x_3)$ and $Y = (y_1, y_2, y_3)$. Then the centroid Z becomes $Z = (z_1, z_2, z_3)$, where $z_1 = (x_1 + y_1)/2$ and $z_2 = (x_2 + y_2)/2$ and $z_3 = (x_3 + y_3)/2$

The squared error in K-means algorithm is obtained using the formula:

$$se_{K_i} = \sum_{j=1}^m \|t_{ij} - C_k\|^2$$

$$se_K = \sum_{j=1}^k se_{K_j}$$


Objective function

Strength of the K-means method

1. The clustering method is relatively efficient: $O(tkn)$, where n is # objects, k is # clusters, and t is # iterations. Normally, $k, t \ll n$.
2. The K-means method often terminates at a local optimum. The global optimum may be found using techniques such as: deterministic annealing and genetic algorithms.

Weakness of the K-means method

1. K-means clustering is not suitable to discover clusters with non-convex shapes
2. There is need to specify k , the number of clusters, in advance and it is unable to handle noisy data and outliers
3. It is applicable only when mean is defined (ratio and interval data only) and therefore it is not applicable to categorical data.

Quality threshold clustering (QT clust) algorithm

Quality threshold clustering is an alternative method of partitioning data which was invented for gene clustering in genetics. The technical difference is that It requires more computing power than k-means, but does not require specifying the number of clusters a priori, and always returns the same result when run several times (Karypis, Han and Kumar, 1999).

The Quality threshold clustering algorithm is:

- The psychometrician chooses a maximum diameter for clusters.
- Build a candidate cluster for each point by including the closest point, the next closest, and so on, until the diameter of the cluster surpasses the threshold.
- Save the candidate cluster with the most points as the first true cluster, and then remove all points in the cluster from further consideration.
- Re-curse with the reduced set of points.

Thereafter the distance between a point and a group of points is computed using complete linkage, i.e. as the maximum distance from the point to any member of the group.

Fuzzy c-means clustering

Fuzzy c-means (FCM) clustering is a method of clustering which allows one piece of data to belong to two or more clusters. The centroid of a cluster is calculated as the mean of all points, weighted by their degree of belonging to the cluster and the degree, to which an element belongs to a given cluster, is a numerical value varying from 0 to 1. FCM clustering allows each point the degree of belonging to clusters, as in fuzzy logic, rather than belonging completely to just one cluster.

Thus, points on the edge of a cluster may be in the cluster to a lesser degree than points in the center of cluster and in essence, this algorithm lumps the two clusters with natural shapes but close boundaries into a large cluster. The Fuzzy c-means (FCM) clustering algorithm is more suited to data that is more or less evenly distributed around the cluster centers and it is frequently used in pattern recognition. The FCM algorithm is different from k-means and k-medoid clustering, where each object is affected exactly to one cluster. K-means and k-medoids clustering are known as hard or non-fuzzy clustering.

In the process of pattern recognition the FCM algorithm is based on minimization of the following objective function:

$$J_m = \sum_{i=1}^N \sum_{j=1}^C u_{ij}^m \|x_i - c_j\|^2$$

$$1 \leq m \leq \infty$$

where m is any real number greater than 1, u_{ij} is the degree of membership of x_i in the cluster j , x_i is the i th of d -dimensional measured data, c_j is the d -dimension center of the cluster, and $\|*\|$ is any norm expressing the similarity between any measured data and the center.

Fuzzy partitioning is carried out through an iterative optimization of the objective function shown above, with the update of membership u_{ij} and the cluster centers c_j by:

$$u_{ij} = \frac{1}{\sum_{k=1}^C \left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|} \right)^{\frac{2}{m-1}}}$$

$$c_j = \frac{\sum_{i=1}^N u_{ij}^m \cdot x_i}{\sum_{i=1}^N u_{ij}^m}$$

$$\max_{ij} \left\{ \left| u_{ij}^{(k+1)} - u_{ij}^{(k)} \right| \right\} < \varepsilon$$

When is attained, then this iteration should be stopped. Here, χ is a termination criterion between 0 and 1, whereas k are the iteration steps. This procedure converges to a local minimum or a saddle point of J_m .

Huang (1997) opines that fuzzy clustering algorithms are divided into two areas: (a) classical fuzzy clustering and (b) shape-based fuzzy clustering.

A. Classical fuzzy clustering algorithms.

There are three subtypes of classical fuzzy clustering algorithms as explained below:

- Fuzzy C-Means algorithm (FCM). It is identical to the K-Means algorithm and it is the most widely-used algorithm. An observation or a data point can theoretically belong to all groups, with a membership function (also called a membership grade) between 0 and 1, where: 0 is where the data point is at the farthest possible point from a cluster's center and 1 is where the data point is closest to the center. Examples include and Possibilistic Fuzzy C-Means (PFCM) algorithm, the Possibilistic C-Means (PCM) algorithm and the Fuzzy Possibilistic C-Means (FPCM) algorithm.
- Gaussian Mixture Decomposition algorithm or the Gath-Geva algorithm which is similar to FCM, but its clusters can have any shape.
- Gustafson-Kessel (GK) algorithm normally associates a data point with a cluster and a matrix. The Gustafson-Kessel (GK) algorithm has elliptical-shaped clusters while the fuzzy C-means assumes the clusters are spherical.

B. Shape-based fuzzy clustering algorithms.

There are three subtypes of shape-based fuzzy clustering algorithms as explained below:

- The Elliptical shaped fuzzy clustering algorithms which is an algorithm that constrains points to elliptical shapes and it is used in the Gustafson-Kessel algorithm.
- Generic shaped fuzzy clustering is an algorithm in which most real life objects are considered neither circular not elliptical and functionally the generic algorithm allows clusters to take any shape.
- Circular-shaped (CS) algorithms are what constrains data point to a circular shape and technically we end up with a CS-FCM when this algorithm is incorporated into Fuzzy C-Means.

**Desirable properties of a clustering algorithm**

As part of an evaluation checklist or a guide to the selection of an algorithm, the expert should consider the following properties:

- i. The algorithm should be able to deal with noise and outliers
- ii. It should have the ability to deal with different data types
- iii. Consider the scalability issues (in terms of both time and space)
- iv. The algorithm should be insensitive to order of input records
- v. It should allow for incorporation of user-specified constraints
- vi. It should be easy to both use and interpret the algorithm
- vii. It is desired that a good algorithm have minimal requirements for domain knowledge to determine input parameters

Advantages of nonhierarchical cluster analysis (NHCA)

- i. Non Hierarchical Cluster Analysis can analyze extremely large data sets which makes it widely applicable.
- ii. The results of NHCA are less susceptible to outliers in the data.

Disadvantages of nonhierarchical cluster analysis (NHCA)

- i. Nonhierarchical methods are also not so efficient when there is a large number of potential cluster solutions. In contrast to the hierarchical techniques that generate all possible cluster solutions in a single analysis, in the NHCA methods each cluster solution is a separate analysis. The consequence is that the nonhierarchical techniques are not the best for use in exploring a wide range of solutions based on varying elements such as similarity measures, observations included, and potential seed points.
- ii. NHCA tends to deliver different final solutions for each set of specified seed points because even in a non-random starting solution there is no guarantee that an optimal clustering of observations will be obtained.

How should the expert interpret the resulting clusters?

According to Han and Kamber, (2001) interpretation of the resulting cluster results involves examining the distinguishing characteristics of each cluster's profile and identifying substantial differences between clusters. Specific attention should be paid to the cluster centroid and a mean profile of the cluster on each clustering variable.

The psychometrician should examine the cluster solutions that fail to show substantial variation from other cluster solutions. The cluster centroid should also be assessed for correspondence with the psychometrician's prior expectations based on theory or practical experience.

Validating clusters

Validation is essential in cluster analysis because the clusters are descriptive of structure and require additional support for their relevance. Therefore in the final activities of clustering the expert should conduct this exercise. Validation is achieved through various means thus:

Validation can be achieved by examining differences on variables not included in the cluster analysis but for which a theoretical and relevant reason enables the expectation of variation across the clusters.

Secondly, cross - validation can also be employed. It empirically validates a cluster solution by randomly splitting the sample to create two sub-samples and then comparing the two cluster solutions for consistency with respect to the number of clusters and the clusters profiles.

We should also mention that clustering research into group-based orientation is commonly assessed using K-means clustering. Much as this method (K-means) has been shown to be useful and effective it does not allow the psychometrician to account for any overlaps among the clusters. The only remedy that could help address the overlap problem is the use of fuzzy clustering methods.



References



- Aldenderfer, M.S. and Blashfield, R. (1984). *Cluster Analysis*, Sage Publications, London, Thousand Oaks, New Delhi, 88 p.
- Agrawal, R., Gehrke J., Gunopulos D., and Raghavan P. (1998). Automatic subspace clustering of high dimensional data for data mining applications. In *Proc. ACM SIGMOD Int. Conf. on Management of Data*, 1998, pp. 94–105.
- Bezdeck, J.C., Ehrlich, R., and Full, W. (1984). Fuzzy C-Means algorithm. *Comput. Geosci.*, 10(2–3):191–203, 1984.
- Bailey, K. D., (1994). *Typologies and Taxonomies: An Introduction to Classification Techniques*, Sage Publications, London, Thousand Oaks, New Delhi, 96 p.
- Ester, M., Kriegel, H.-Peter., Sander, J., and Xu X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proc. 2nd Int. Conf. on Knowledge Discovery and Data Mining*, 1996, pp. 226–231.
- Everitt, Brian S., (1993). *Cluster Analysis*, 3rd ed., Halsted Press, London, New York, 170 p. Gordon, A.D., 1999, *Classification*, Chapman & Hall/CRC, Boca Raton, 256 p.
- Everitt, B.S., Landau, S., and Leese, M. (2001) *Cluster Analysis*. Hodder Arnold, London, UK
- Fayyad, U.M., Piatetsky-Shapiro, G., Smuth, P., and Uthurusamy, R. (2008). *Advances in Knowledge Discovery and Data Mining*. AAAI Press, Menlo Park, CA
- Han, J. and Kamber, M. (2001). *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers, San Fransisco, CA, 2001.
- Huang, Z. A (1997). Fast clustering algorithm to cluster very large categorical data sets in data mining. In *Proc. ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, 1997.
- Jain, A.K., Murty, M.N., and Flynn P.J. (1999). Data clustering: a review. *ACM Computing Surv.*, 31(3):264–323
- Karypis, G., Han E.-H., and Kumar, V. (1999). CHAMELEON: a hierarchical clustering algorithm using dynamic modeling. *IEEE Computer.*, 32(8):68–75
- MacQueen, J.B. (1967). Some methods for classification and analysis of multivariate observations. In *Proc. 5th Berkeley Symp. on Mathematical Statistics and Probability*, vol. 1, pp. 281–297
- Romesburg, C.H., (1984), *Cluster Analysis for Researchers*, Lifetime Learning Publications Belmont, CA, 334 p.
- Theodoridis, S. and Koutroubas, K. (1999). *Pattern Recognition*. Academic Press, New York
- Vazirgiannis M., Halkidi M., and Gunopulos, D. (2003). *Uncertainty Handling and Quality Assessment in Data Mining*. Springer, New York
- Zhang, T., Ramakrishnan, R., and Linvy, M. (1996). BIRCH: an efficient method for very large databases. In *Proc. ACM SIGMOD Int. Conf. on Management of Data*, 1996, pp. 103–114.