Cluster analysis was originated in anthropology by Driver and Kroeber in 1932. It is the task of grouping a set of objects in such a way that objects in the same group or cluster are more similar to each other than to those in other groups or clusters. It is a common technique for statistical data analysis.

Cluster analysis can be achieved by various algorithms that might differ significantly. Modern notions of clusters include groups with small distances among the cluster members, dense areas of the data space, intervals or particular statistical distributions. Therefore, cluster analysis as such is not a trivial task. It is an interactive multi-objective optimization that involves trial and error. It is also often necessary to modify data preprocessing and model or algorithms parameters until the result achieves the desired characteristics.

Therefore, in cluster analysis, the clustering of subjects or variables are made from similarity measures or dissimilarity (distance) between two subjects initially, and later between two clusters. These groups can be done using hierarchical or non-hierarchical techniques.

Similarity and Dissimilarity

The identification of natural clusters of subjects or variables requires that the similarity between these have to be measured explicitly. There are several similarity measures (or proximity) or dissimilarity (or distance) that can be used depending on the variable type (interval, frequency or nominal). In cluster analysis, the most common measures are:

Euclidean distance: is the distance between two points (p, q) in any dimension of the space and is the most common use of distance. When data is dense or continuous, this is the best proximity measure. Euclidean distance measure is given by:

$$d(p,q) = \sqrt{\sum_{k=1}^{n} (p_k - q_k)^2}$$

Minkowski distance: is a metric in a normed vector space, which can be considered a generalization of the Euclidean distance. The Minkowski distance measured between two points (p, q) is given by:

$$d(p,q) = \left(\sum_{k=1}^{n} |p_k - q_k|^c\right)^{\frac{1}{c}}$$

With $c = 1$ and $c = 2$, the Minkowski metric becomes equal to the Manhattan and Euclidean metrics respectively.

Cosine Similarity: it is often used when comparing two documents against each other. It measures the angle between two vectors. If the value is zero, the angle between the two vectors is 90 degrees, and they share no terms. If the value is one, the two vectors are the same except for magnitude. Given two vectors of attributes, u and v, the cosine similarity, $cos\theta$, is represented as

$$cos\theta = \frac{u \cdot v}{\|u\| \|v\|} = \frac{\sum_{i=1}^{n} u_i v_i}{\sqrt{\sum_{i=1}^{n} u_i^2} \sqrt{\sum_{i=1}^{n} v_i^2}}$$

where u_i and v_i are components of vector u and v, respectively.

Jaccard similarity: is a standard index for binary variables. It is defined as the quotient between the intersection and the union of the pairwise compared variables between two objects.

The Jaccard distance between the objects i and j is given by

$$d(i,j) = \frac{M_{11}}{M_{01} + M_{10} + M_{11}}$$

M_{11} represents the total number of attributes where both data objects have a 1; and M_{10}, M_{01} represent the total number of attributes where one data object has a 1, and the other has a 0. The total matching attributes are then divided by the total non-matching attributes, plus the matching ones. A perfect similarity score would then be 1.

Similarity measures for variables: when cluster analysis aims to group variables (and not subjects or items), the appropriate similarity measures are the sample correlation coefficients. In case of continuous variables, Pearson correlation coefficient is the most suitable. For ordinal variables the Spearman correlation coefficient should be used. Finally, for nominal variables, the reader should use the phi coefficient, $\phi = \sqrt{\frac{X^2}{N}}$, where X^2 is the chi-square statistic.

Hierarchical Clustering

Hierarchical techniques appeal to successive steps of aggregation of the considered subjects, individually. Thus, given a set of N items to be clustered, and a $N \times N$ distance (or similarity) matrix, the primary process of hierarchical clustering is:

1. Start by assigning each item to its cluster, so that if it has N items, it now has N clusters, each containing just one item. Let the distances (similarities) between the clusters equal the distances (similarities) between items they contain.
2. Find the closest (most similar) pair of clusters and merge them into a single cluster, so that now it has 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.

Hierarchical methods of clusters mostly differ in how these distances (in step 3) are calculated. The methods most frequently used are:

Single-linkage clustering

Single linkage (also called connectedness or minimum method) is one of the simplest agglomerative hierarchical clustering methods. In single linkage, the distance between groups is defined as the distance between the closest pair of objects, where only pairs consisting of one object from each group are considered.

In single linkage method, $D(r, s)$ is computed as $D(r, s) = Min\big(d(i, j)\big)$, where i is in cluster r and j is in cluster s. Thus, the distance between two clusters is given by the value of the shortest link between the clusters.

Complete linkage clustering

In complete linkage (also called farthest neighbor), the clustering method is the opposite of single linkage. The distance between groups is defined as the distance between the most distant pair of objects, one from each group.

In complete linkage method, $D(r, s)$ is computed as$D(r, s) = Max\big(d(i, j)\big)$, where i is in cluster r and object j is in cluster s. Thus, the distance between two clusters is given by the value of the longest link between clusters.

Average group linkage

With average group linkage, the groups formed are represented by their mean values for each variable (i.e., their mean vector and inter-group distance is defined regarding the distance between two such mean vectors).

In average group linkage method, the two clusters, r, and s, are merged such that the average pairwise distance within the newly formed cluster is minimum. Suppose the new cluster formed by combining clusters r and s is labeled as t. Then the distance between clusters r and s, $D(r, s)$, is computed as $D(r, s) = Average\big(d(i, j)\big)$, where observations i and j are in cluster t, the cluster formed by merging clusters r and s.

At each stage of hierarchical clustering, the r and s clusters for which $D(r, s)$ is minimum, are merged. In this case, those two clusters are merged such that the newly formed cluster, on average, will have minimum pairwise distances between the points.

Average linkage within groups

Average linkage within groups is a technique of cluster analysis in which clusters are combined in order to minimize the average distance between all individuals or cases in the resulting cluster. Also, the distance between two clusters is defined as the average distance between all possible pairs of individuals in the cluster that would result if they were combined.

Centroid clustering

A cluster centroid is the middle point of a cluster. A centroid is a vector containing one number for each variable, where each number is the mean of a variable for the observations in that cluster.
The reader can use the centroid as a measure of cluster location. For a particular cluster, the average distance from the centroid is the average of the distances between observations and the centroid. The maximum distance from the centroid is the maximum of these distances.

Ward method

It is an alternative approach for performing cluster analysis. Essentially, it looks at cluster analysis as an analysis of variance problem, instead of using distance metrics or measures of association.
This method involves an agglomerative clustering algorithm. It will start out at the leaves and work its way to the trunk. It looks for groups of leaves that it forms into branches, the branches into limbs and eventually into the trunk. Ward's method starts out with n clusters of size 1 and continues until all the observations are included into one cluster.
This method is the most appropriate for quantitative variables and not binary variables.

As there are several available methods, the existence of advantages and disadvantages in using each one of them is visible. Since the "best" method of performing hierarchical clustering does not exist, some authors (Marôco, 2011) suggest the use of various methods simultaneously. Hence, if all methods produce similarly interpretable solutions, it is possible to conclude that data matrix has natural groupings.

Non-hierarchical cluster analysis

Non-hierarchical clustering methods are intended in grouping items (and not variables) in a set of clusters whose number is defined a-priori. These methods quickly apply to arrays of large data because it is not necessary to calculate and store a new dissimilarity matrix in each step of the algorithm.
There are various non-hierarchical methods that differ primarily in the way it unfolds the first aggregation of items in clusters and how the new distances between the centroids of the clusters and the item are calculated. One of the standard methods in most statistical software is the K-means.

K-means

The procedure follows a straightforward and easy way to classify a given data set with a specified number of clusters (assume k clusters) fixed a-priori. The main idea is to determine k centers, one for each cluster. These centers should be placed in a cunning way because of different location causes a different result.
Thus, the better choice is to place them far away from each other, as much as possible. The next step is to take each point belonging to a given data set and associate it to the nearest center. When no point is pending, the first phase is completed, and an early group is done. At this point, the re-calculation of k new centroids as barycenter of the clusters resulting from the previous step is done. After this, a new binding has to be done between the same data set points and the nearest new center. A loop has been generated. As a result of this loop, the reader may notice that k centers change their location, step by step, until no more changes are done or, in other words, centers do not move anymore.

REFERENCES

Driver, H. E., and Kroeber, A. L. (1932). Quantitative expression of cultural relationships. University of California Publications in American Archeology and Ethnology, 31 (4), pp. 211-256.

Marôco, J. (2011). Análise Estatística com o SPSS Statistics. 5th Edition. Pero Pinheiro: Report Number, pp. 7-61.