Cluster Analysis

Cluster analysis is a set of mathematical methods for exploring the quantitative properties of statistical data for sample items drawn from the population, by classifying them into groups within clusters, so that the items within each group are similar to each other with regard to variables or characteristics, and where the groups are different from each other. In other words, the goal of cluster analysis is to collect the sample items and classify them into groups that are internally homogeneous and externally disparate from each other.

In general, cluster analysis is divided into two basic types:

- Hierarchical cluster analysis.
- Non- Hierarchical cluster analysis.

Cluster analysis

Cluster analysis is one of the branches of multivariate statistical analysis. It is concerned with grouping the items of the research community in the form of a cluster that begins with a branch and ends with a single branch. The grouping is done either on the basis of the observations themselves in light of the characteristics of the variables, or on the basis of...

The variables themselves.

The idea of cluster analysis begins without prior knowledge of the number of groups or which items belong to this or that group. The method is exploratory approach.

Some basic concepts

Element

The element (xi) is a vector in scale space of n dimensions

$$\underline{Xi} = (X_{i1}, X_{i2}, \ldots\ldots\ldots, X_{in})$$

Elements are numerical values of measurable quantities (properties).

Distance (D).

It is the space or space separating two elements. The relationship between similarity and distance is an inverse relationship, and cluster analysis can be performed based on either of them. And its mathematical formula

$$D_m(\underline{X}_i, \underline{X}_j) = \left(\sum_w (X_{iw} - X_{jw})^m \right)^{1/m}$$

whereas :

$\underline{Xi}, \underline{Xj}$ are the two elements between which the Euclidean distance was calculated

Xiw It is the component w of element I in the measurement space with (n) dimensions

Cluster

It is a group of fairly homogeneous elements that describe what is within one cluster and is different from the elements within other clusters. It is also known as a group of adjacent objects for a statistical population (such as a family).

Tree

It is the hierarchical shape resulting from the clustering process, and it can be accessed in two ways

1. Agglomerative method; This method consists of a series of steps, each of which is linked clusters and elements are grouped together based on the similarity factor or the distance factor.

2. division method; You start by separating the large group in which the elements are located into small parts until you reach the last group, which contains two elements that have been separated into components.

In both cases, the result shown by the two methods is a hierarchical tree, and the beginning of the branch is called the root and the branch points are called the nodes.

The final or final node on the tree has no branches. It is called leaves, and it represents the elements that have come together. Each of the nodes in the tree, including the root, represents a qualitative set of all the things that can be accessed at that node towards the front and through the tree.

Classification

It is the arrangement of things based on their similarities or differences, or these things may be arranged according to more than one method, meaning that it is possible placing more than one arrangement of elements or things according to the similarity or difference of interest.

Clustering Steps

1. Calculate the distance matrix, correlation matrix, or similarity matrix.
2. The two elements with the shortest distance between them are connected within the matrix calculated in (a), and in the event that there are equal distances, it is possible to perform the connection process for more than two elements in one stage (both elements together).
3. A new distance matrix is calculated that takes into account the changes that occurred in (b). The degree of the new distance matrix will decrease by the number of linking operations performed in stage (b).

d. The linking process continues until the cluster tree is reached.

It is worth noting that data conversion may be performed, such as taking the logarithm or converting to the standard degree, before performing the above operation.

Agglomerate Active Methods

In this section, the two most important clustering methods will be discussed, which are:

Single Linkage method

This method is based on considering that the two elements that are most similar among the elements will form the nucleus of the cluster, then the rest of the units are added to this nucleus sequentially and according to the degree of similarity with the elements of the cluster nucleus, where the most similar are added, then the least, and gradually, and in the case of linking a group of clusters to each other, this is done. Based on the closest distances between cluster elements and according to the formulas the following:

$$D_{I}\,,_{J} = \min\{D_{ij}\} \qquad i \in I \quad , \quad j \in J$$

Where i and j represent the elements of clusters I and J, respectively

Complete Linkage method

The principle of operation of this method is completely opposite to the principle of operation of the previous method. The distance between it and any of the cluster elements must be greater than the distance between any other element and any of the cluster elements. The elements of the clusters are linked according to the following formula:

$$DI,J = Max\{Dij\} \qquad i \in I \quad , \quad j \in J$$