Types of cluster analysis

There are two types of cluster analysis used to classify data:

1. Hierarchical cluster analysis

. Hierarchal Cluster Analysis 2. Non-hierarchical cluster analysis [averages method]

K-Means

Hierarchical cluster analysis

This method is one of the preferred methods in cluster analysis, in which 6 items are clustered sequentially into m clusters, where m is the weakest cluster and m is the strongest.

Hierarchical clustering methods:

.(1) The adivisive technique

The work of this method can be summed up on the assumption that there is one cluster of vocabulary, then this cluster is divided into partial clusters, then the clusters are divided into smaller clusters, and so we continue until each vocabulary has its own cluster.

(2) Clustering method. Agglomerative technique

The principle of operation of this method is completely different from the method above, as it is assumed that each word describes its own sub-cluster, then similar sub-clusters are grouped into more comprehensive sub-clusters, and the process is repeated until a single cluster that includes all the items is obtained.

Stages of implementing hierarchical cluster analysis

1. Formation of the proximities matrix

Formation of the proximity matrix (proximities matrix), which is a square matrix of symmetric degree n, whose elements are a measure that expresses the distance between each pair of data (cases), and is symbolized by the symbol D, and is known as below:

$$\boldsymbol{D} = \begin{bmatrix} \boldsymbol{d_{11}} & \boldsymbol{d_{12}} & \cdots & \boldsymbol{d_{1n}} \\ \vdots & \vdots & \cdots & \vdots \\ \boldsymbol{d_{n1}} & \boldsymbol{d_{n2}} & \cdots & \boldsymbol{d_{nn}} \end{bmatrix}$$

There are a number of ways to measure the distance between the elements of the kinship matrix, including:

1 - Euclidean Distance. -

2. Minkowsski metric formula

3- City Block distance

المسافة الاقليدية Euclidean	$d_e = \left[\sum_{i=1}^{P} (x_{ij} - x_{ik})^2\right]^{\frac{1}{2}}$
مسافة المقاطع City Block	$d_{cb} = \sum_{i=1}^{P} \lvert x_{ij} - x_{ik} \rvert$
مسافة (تشيبتشيف) Chebychev	$d_{ch} = Max \lvert x_{ij} - X_{ik} \rvert$
مسافة (مينكوفسكي) Minkowski	$d_m = \left[\sum_{i=1}^{P} (x_{ij} - x_{ik})^m\right]^{\frac{1}{m}}$
المسافة التربيعية Q_s Mahalanobis	$d_q = \sum (X_{ij} - X_{ik})^2 S^{-1} (X'_j - X_k)$

Example

To classify 5 students into clusters according to only two variables: x_1 expenditures on food for the student, and x_2 expenditures on communications by the student, according to the data shown in the following table:

Code	x_1	x_2
1 = a	2	4
2 = b	8	2
3 = c	9	3
4 = d	1	5
5 = e	8.5	1

Let us first assume that each of the five students alone forms a special cluster, then we calculate the Euclidean distances between each pair according to the relationship:

$$d_{jk} = \sqrt{(x_{1j} - x_{1k})^2 + (x_{2j} - x_{2k})^2}$$

We find the distance between students 1 and 2

$$d_{12} = \sqrt{(2-8)^2 + (4-2)^2} = 6.325$$

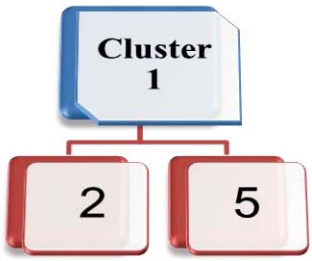
We find the distance between students 1 and 3

$$d_{13} = \sqrt{(2-9)^2 + (4-3)^2} = 7.071$$

Continuing to calculate the rest of the Euclidean distances, it is possible to formulate the distance matrix as follows:

$$D = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 0 & 6{,}325 & 7.071 & 1.414 & 7.159 \\ & 0 & 1.414 & 7.616 & 1.116 \\ & & 0 & 8.246 & 2.062 \\ & & & 0 & 8.500 \\ & & & & 0 \end{bmatrix} \end{array}$$

We note that the smallest element of this matrix is 1.116, which corresponds to students 2 and 5, because they are the most similar in expenses, so we can create the first cluster from them.



Therefore, the previous matrix will be updated by merging the components of the first cluster together into one column and one row, and the values included in the new column will be updated according to the following equation:

$$d_{(u,v)_j} = \frac{1}{2}\left[d_{uj} + d_{vj}\right]$$

or

$$D(P_k, (P_i, P_j)) = \frac{1}{2}\left((P_k, P_i) + (P_k, P_j)\right)$$

Accordingly, the following points will be updated:

$$D(P_1,(P_2,P_5)), D((P_2,P_5),P_3), D((P_2,P_5),P_4).$$

The new matrix will be as follows: