Data Mining – Cluster Analysis

INTRODUCTION:

Cluster analysis, also known as clustering, is a method of data mining that groups similar data points together. The goal of cluster analysis is to divide a dataset into groups (or clusters) such that the data points within each group are more similar to each other than to data points in other groups. This process is often used for exploratory data analysis and can help identify patterns or relationships within the data that may not be immediately obvious. There are many different algorithms used for cluster analysis, such as k-means, hierarchical clustering, and density-based clustering. The choice of algorithm will depend on the specific requirements of the analysis and the nature of the data being analyzed.

Cluster Analysis is the process to find similar groups of objects in order to form clusters. It is an unsupervised machine learning-based algorithm that acts on unlabelled data. A group of data points would comprise together to form a cluster in which all the objects would belong to the same group.

The given data is divided into different groups by combining similar objects into a group. This group is nothing but a cluster. A cluster is nothing but a collection of similar data which is grouped together.

For example, consider a dataset of vehicles given in which it contains information about different vehicles like cars, buses, bicycles, etc. As it is unsupervised learning there are no class labels like Cars, Bikes, etc for all the vehicles, all the data is combined and is not in a structured manner.

Now our task is to convert the unlabelled data to labelled data and it can be done using clusters.

The main idea of cluster analysis is that it would arrange all the data points by forming clusters like cars cluster which contains all the cars, bikes clusters which contains all the bikes, etc.

Simply it is the partitioning of similar objects which are applied to unlabelled data.

Properties of Clustering :

1. Clustering Scalability: Nowadays there is a vast amount of data and should be dealing with huge databases. In order to handle extensive databases, the clustering algorithm should be scalable. Data should be scalable, if it is not scalable, then we can't get the appropriate result which would lead to wrong results.

2. High Dimensionality: The algorithm should be able to handle high dimensional space along with the data of small size.

3. Algorithm Usability with multiple data kinds: Different kinds of data can be used with algorithms of clustering. It should be capable of dealing with different types of data like discrete, categorical and interval-based data, binary data etc.

4. Dealing with unstructured data: There would be some databases that contain missing values, and noisy or erroneous data. If the algorithms are sensitive to such data then it may lead to poor quality clusters. So it should be able to handle unstructured data and give some structure to the data by organising it into groups of similar data objects. This makes the job of the data expert easier in order to process the data and discover new patterns.

5. Interpretability: The clustering outcomes should be interpretable, comprehensible, and usable. The interpretability reflects how easily the data is understood.

Clustering Methods:
The clustering methods can be classified into the following categories:

Partitioning Method: It is used to make partitions on the data in order to form clusters. If "n" partitions are done on "p" objects of the database then each partition is represented by a cluster and n < p. The two conditions which need to be satisfied with this Partitioning Clustering Method are:

One objective should only belong to only one group.
There should be no group without even a single purpose.

In the partitioning method, there is one technique called iterative relocation, which means the object will be moved from one group to another to improve the partitioning

Hierarchical Method: In this method, a hierarchical decomposition of the given set of data objects is created. We can classify hierarchical methods and will be able to know the purpose of classification on the basis of how the hierarchical decomposition is formed. There are two types of approaches for the creation of hierarchical decomposition, they are:

Agglomerative Approach: The agglomerative approach is also known as the bottom-up approach. Initially, the given data is divided into which objects form separate groups. Thereafter it keeps on merging the objects or the groups that are close to one another which means that they exhibit similar properties. This merging process continues until the termination condition holds.

Divisive Approach: The divisive approach is also known as the top-down approach. In this approach, we would start with the data objects that are in the same cluster. The group of individual clusters is divided into small clusters by continuous iteration. The iteration continues until the condition of termination is met or until each cluster contains one object.

Once the group is split or merged then it can never be undone as it is a rigid method and is not so flexible. The two approaches which can be used to improve the Hierarchical Clustering Quality in Data Mining are: –

One should carefully analyze the linkages of the object at every partitioning of hierarchical clustering.

One can use a hierarchical agglomerative algorithm for the integration of hierarchical agglomeration. In this approach, first, the objects are grouped into micro-clusters. After grouping data objects into microclusters, macro clustering is performed on the microcluster.

Density-Based Method: The density-based method mainly focuses on density. In this method, the given cluster will keep on growing continuously as long as the density in the neighbourhood exceeds some threshold, i.e, for each data point within a given cluster. The radius of a given cluster has to contain at least a minimum number of points.

Grid-Based Method: In the Grid-Based method a grid is formed using the object together,i.e, the object space is quantized into a finite number of cells that form a grid structure. One of the major advantages of the grid-based method is fast processing time and it is dependent only on the number of cells in each dimension in the quantized space. The processing time for this method is much faster so it can save time.

Model-Based Method: In the model-based method, all the clusters are hypothesized in order to find the data which is best suited for the model. The clustering of the density function is used to locate the clusters for a given model. It reflects the spatial distribution of data points and also provides a way to automatically determine the number of clusters based on standard statistics, taking outlier or noise into account. Therefore it yields robust clustering methods.

Constraint-Based Method: The constraint-based clustering method is performed by the incorporation of application or user-oriented constraints. A constraint refers to the user expectation or the properties of the desired clustering results. Constraints provide us with an interactive way of communication with the clustering process. The user or the application requirement can specify constraints.

Applications Of Cluster Analysis:

- It is widely used in image processing, data analysis, and pattern recognition.
- It helps marketers to find the distinct groups in their customer base and they can characterize their customer groups by using purchasing patterns.
- It can be used in the field of biology, by deriving animal and plant taxonomies and identifying genes with the same capabilities.
- It also helps in information discovery by classifying documents on the web.

Advantages of Cluster Analysis:

1. It can help identify patterns and relationships within a dataset that may not be immediately obvious.

2. It can be used for exploratory data analysis and can help with feature selection.

3. It can be used to reduce the dimensionality of the data.

4. It can be used for anomaly detection and outlier identification.

5. It can be used for market segmentation and customer profiling.

Disadvantages of Cluster Analysis:

1. It can be sensitive to the choice of initial conditions and the number of clusters.

2. It can be sensitive to the presence of noise or outliers in the data.

3. It can be difficult to interpret the results of the analysis if the clusters are not well-defined.

4. It can be computationally expensive for large datasets.

5. The results of the analysis can be affected by the choice of clustering algorithm used.

6. It is important to note that the success of cluster analysis depends on the data, the goals of the analysis, and the ability of the analyst to interpret the results.

Types Of Data Used In Cluster Analysis - Data Mining

Types Of Data Used In Cluster Analysis Are:

- **Interval-Scaled variables**
- **Binary variables**
- **Nominal, Ordinal, and Ratio variables**
- **Variables of mixed types**

Types Of Data Structures

First of all, let us know what types of data structures are widely used in cluster analysis.
Suppose that a data set to be clustered contains n objects, which may represent persons, houses, documents, countries, and so on.

Main memory-based clustering algorithms typically operate on either of the following two data structures.

Types of data structures in cluster analysis are

- **Data Matrix** (or object by variable structure)
- **Dissimilarity Matrix** (or object by object structure)

Data Matrix

This represents n objects, such as persons, with p variables (also called measurements or attributes), such as age, height, weight, gender, race and so on. The structure is in the form of a relational table, or n-by-p matrix (n objects x p variables)

The Data Matrix is often called a two-mode matrix since the rows and columns of this represent the different entities.

$$\begin{bmatrix} x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{i1} & \cdots & x_{if} & \cdots & x_{ip} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{n1} & \cdots & x_{nf} & \cdots & x_{np} \end{bmatrix}$$

Dissimilarity Matrix

This stores a collection of proximities that are available for all pairs of n objects. It is often represented by a n – by – n table, where d(i,j) is the measured difference or dissimilarity between objects i and j. In general, d(i,j) is a non-negative number that is close to 0 when objects i and j are higher similar or "near" each other and becomes larger the more they differ. Since d(i,j) = d(j,i) and d(i,i) =0, we have the matrix in figure.

This is also called as one mode matrix since the rows and columns of this represent the same entity.

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & & \\ d(n,1) & d(n,2) & \ldots & \ldots & 0 \end{bmatrix}$$

Types Of Data In Cluster Analysis Are:

Interval-Scaled Variables

Interval-scaled variables are continuous measurements of a roughly linear scale.

Typical examples include weight and height, latitude and longitude coordinates (e.g., when clustering houses), and weather temperature.

The measurement unit used can affect the clustering analysis. For example, changing measurement units from meters to inches for height, or from kilograms to pounds for weight, may lead to a very different clustering structure.
In general, expressing a variable in smaller units will lead to a larger range for that variable, and thus a larger effect on the resulting clustering structure.

To help avoid dependence on the choice of measurement units, the data should be standardized. Standardizing measurements attempts to give all variables an equal weight.
This is especially useful when given no prior knowledge of the data. However, in some applications, users may intentionally want to give more weight to a certain set of variables than to others.

For example, when clustering basketball player candidates, we may prefer to give more weight to the variable height.

Binary Variables

A binary variable is a variable that can take only 2 values.

For example, generally, gender variables can take 2 variables male and female.

Contingency Table For Binary Data

Let us consider binary values 0 and 1

	1	0	sum
1	a	b	$a+b$
0	c	d	$c+d$
sum	$a+c$	$b+d$	p

Let p=a+b+c+d

Simple matching coefficient (invariant, if the binary variable is symmetric):

$$d(i, j) = \frac{b + c}{a + b + c + d}$$

Jaccard coefficient (noninvariant if the binary variable is asymmetric):

$$d(i, j) = \frac{b + c}{a + b + c}$$

Nominal or Categorical Variables

A generalization of the binary variable in that it can take more than 2 states, e.g., red, yellow, blue, green.

Method 1: Simple matching

The dissimilarity between two objects i and j can be computed based on the simple matching.

m: Let m be no of matches (i.e., the number of variables for which i and j are in the same state).

p: Let p be total no of variables.

$$d(i, j) = \frac{p - m}{p}$$

Method 2: use a large number of binary variables

Creating a new binary variable for each of the M nominal states.

Ordinal Variables

An ordinal variable can be discrete or continuous.

In this order is important, e.g., rank.

It can be treated like interval-scaled

By replacing xif by their rank,

$$r_{if} \in \{1, \ldots, M_f\}$$

By mapping the range of each variable onto [0, 1] by replacing the i-th object in the f-th variable by,

$$z_{if} = \frac{r_{if} - 1}{M_f - 1}$$

Then compute the dissimilarity using methods for interval-scaled variables.

Ratio-Scaled Intervals

Ratio-scaled variable: It is a positive measurement on a nonlinear scale, approximately at an exponential scale, such as Ae∧Bt or A∧e-Bt.

Methods:

First, treat them like interval-scaled variables — not a good choice! (why?)

Then apply logarithmic transformation i.e.y = log(x)

Finally, treat them as continuous ordinal data treat their rank as interval-scaled.

Variables Of Mixed Type

A database may contain all the six types of variables
symmetric binary, asymmetric binary, nominal, ordinal, interval, and ratio.

And those combinedly called as mixed-type variables.

Summary

Data Types in Cluster Analysis

Interval-Scaled variables

Binary variables

Nominal, Ordinal, and Ratio variables

Variables of mixed types

this topic, we will learn what is K-means clustering algorithm, how the algorithm works, along with the Python implementation of k-means clustering.

What is K-Means Algorithm?

K-Means Clustering is an Unsupervised Learning algorithm, which groups the unlabeled dataset into different clusters. Here K defines the number of pre-defined clusters that need to be created in the process, as if K=2, there will be two clusters, and for K=3, there will be three clusters, and so on.

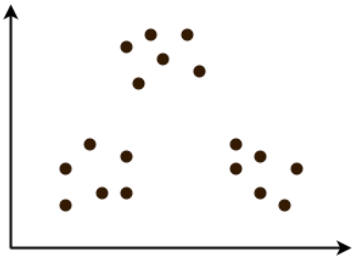
> It is an iterative algorithm that divides the unlabeled dataset into k different clusters in such a way that each dataset belongs only one group that has similar properties.

It allows us to cluster the data into different groups and a convenient way to discover the categories of groups in the unlabeled dataset on its own without the need for any training.

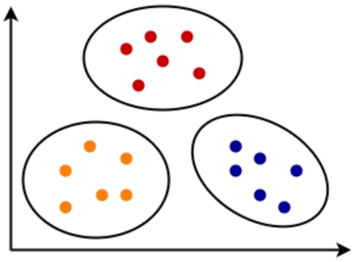
It is a centroid-based algorithm, where each cluster is associated with a centroid. The main aim of this algorithm is to minimize the sum of distances between the data point and their corresponding clusters.

K-Means Clustering-

K-Means clustering is an unsupervised iterative clustering technique.
It partitions the given data set into k predefined distinct clusters.
A cluster is defined as a collection of data points exhibiting certain similarities.



Before K-Means



After K-Means

It partitions the data set such that-

Each data point belongs to a cluster with the nearest mean.
Data points belonging to one cluster have high degree of similarity.
Data points belonging to different clusters have high degree of dissimilarity.

K-Means Clustering Algorithm-

K-Means Clustering Algorithm involves the following steps-

Step-01:

Choose the number of clusters K.

Step-02:

Randomly select any K data points as cluster centers.

Select cluster centers in such a way that they are as farther as possible from each other.

Step-03:

Calculate the distance between each data point and each cluster center.

The distance may be calculated either by using given distance function or by using euclidean distance formula.

Step-04:

Assign each data point to some cluster.

A data point is assigned to that cluster whose center is nearest to that data point.

Step-05:

Re-compute the center of newly formed clusters.

The center of a cluster is computed by taking mean of all the data points contained in that cluster.

Step-06:

Keep repeating the procedure from Step-03 to Step-05 until any of the following stopping criteria is met-

Center of newly formed clusters do not change

Data points remain present in the same cluster

Maximum number of iterations are reached

Advantages-

K-Means Clustering Algorithm offers the following advantages-

Point-01:

It is relatively efficient with time complexity O(nkt) where-

n = number of instances

k = number of clusters

t = number of iterations

Point-02:

It often terminates at local optimum.

Techniques such as Simulated Annealing or Genetic Algorithms may be used to find the global optimum.

Disadvantages-

K-Means Clustering Algorithm has the following disadvantages-

It requires to specify the number of clusters (k) in advance.

It can not handle noisy data and outliers.

It is not suitable to identify clusters with non-convex shapes.

PRACTICE PROBLEMS BASED ON K-MEANS CLUSTERING ALGORITHM-

Problem-01:

Cluster the following eight points (with (x, y) representing locations) into three clusters:

A1(2, 10), A2(2, 5), A3(8, 4), A4(5, 8), A5(7, 5), A6(6, 4), A7(1, 2), A8(4, 9)

Initial cluster centers are: A1(2, 10), A4(5, 8) and A7(1, 2).

The distance function between two points a = (x1, y1) and b = (x2, y2) is defined as-

$$P(a, b) = |x2 - x1| + |y2 - y1|$$

Use K-Means Algorithm to find the three cluster centers after the second iteration.

Solution-

We follow the above discussed K-Means Clustering Algorithm-

Iteration-01:

We calculate the distance of each point from each of the center of the three clusters.

The distance is calculated by using the given distance function.

The following illustration shows the calculation of distance between point A1(2, 10) and each of the center of the three clusters-

Calculating Distance Between A1(2, 10) and C1(2, 10)-

P(A1, C1)

$= |x2 - x1| + |y2 - y1|$

$= |2 - 2| + |10 - 10|$

$= 0$

Calculating Distance Between A1(2, 10) and C2(5, 8)-

P(A1, C2)

$= |x2 - x1| + |y2 - y1|$

$= |5 - 2| + |8 - 10|$

$= 3 + 2$

= 5

Calculating Distance Between A1(2, 10) and C3(1, 2)-

P(A1, C3)

= |x2 – x1| + |y2 – y1|

= |1 – 2| + |2 – 10|

= 1 + 8

= 9

In the similar manner, we calculate the distance of other points from each of the center of the three clusters.

Next,

We draw a table showing all the results.
Using the table, we decide which point belongs to which cluster.
The given point belongs to that cluster whose center is nearest to it.

Given Points	Distance from center (2, 10) of Cluster-01	Distance from center (5, 8) of Cluster-02	Distance from center (1, 2) of Cluster-03	Point belongs to Cluster
A1(2, 10)	0	5	9	C1
A2(2, 5)	5	6	4	C3
A3(8, 4)	12	7	9	C2
A4(5, 8)	5	0	10	C2
A5(7, 5)	10	5	9	C2
A6(6, 4)	10	5	7	C2
A7(1, 2)	9	10	0	C3
A8(4, 9)	3	2	10	C2

From here, New clusters are-

Cluster-01:

First cluster contains points-

A1(2, 10)

Cluster-02:

Second cluster contains points-

A3(8, 4)

A4(5, 8)

A5(7, 5)

A6(6, 4)

A8(4, 9)

Cluster-03:

Third cluster contains points-

A2(2, 5)

A7(1, 2)

Now,

We re-compute the new cluster clusters.

The new cluster center is computed by taking mean of all the points contained in that cluster.

For Cluster-01:

We have only one point A1(2, 10) in Cluster-01.

So, cluster center remains the same.

For Cluster-02:

Center of Cluster-02

= ((8 + 5 + 7 + 6 + 4)/5, (4 + 8 + 5 + 4 + 9)/5)

= (6, 6)

For Cluster-03:

Center of Cluster-03

= ((2 + 1)/2, (5 + 2)/2)

= (1.5, 3.5)

This is completion of Iteration-01.

Iteration-02:

We calculate the distance of each point from each of the center of the three clusters.

The distance is calculated by using the given distance function.

The following illustration shows the calculation of distance between point A1(2, 10) and each of the center of the three clusters-

Calculating Distance Between A1(2, 10) and C1(2, 10)-

P(A1, C1)

= |x2 – x1| + |y2 – y1|

= |2 – 2| + |10 – 10|

= 0

Calculating Distance Between A1(2, 10) and C2(6, 6)-

P(A1, C2)

= |x2 – x1| + |y2 – y1|

= |6 – 2| + |6 – 10|

= 4 + 4

= 8

Calculating Distance Between A1(2, 10) and C3(1.5, 3.5)-

P(A1, C3)

= |x2 – x1| + |y2 – y1|

= |1.5 – 2| + |3.5 – 10|

= 0.5 + 6.5

= 7

In the similar manner, we calculate the distance of other points from each of the center of the three clusters.

Next,

We draw a table showing all the results.

Using the table, we decide which point belongs to which cluster.

The given point belongs to that cluster whose center is nearest to it.

Given Points	Distance from center (2, 10) of Cluster-01	Distance from center (6, 6) of Cluster-02	Distance from center (1.5, 3.5) of Cluster-03	Point belongs to Cluster
A1(2, 10)	0	8	7	C1
A2(2, 5)	5	5	2	C3
A3(8, 4)	12	4	7	C2
A4(5, 8)	5	3	8	C2
A5(7, 5)	10	2	7	C2
A6(6, 4)	10	2	5	C2
A7(1, 2)	9	9	2	C3
A8(4, 9)	3	5	8	C1

From here, New clusters are-

Cluster-01:

First cluster contains points-

A1(2, 10)
A8(4, 9)

Cluster-02:

Second cluster contains points-

A3(8, 4)
A4(5, 8)
A5(7, 5)
A6(6, 4)

Cluster-03:

Third cluster contains points-

A2(2, 5)

A7(1, 2)

Now,

We re-compute the new cluster clusters.

The new cluster center is computed by taking mean of all the points contained in that cluster.

For Cluster-01:

Center of Cluster-01

= ((2 + 4)/2, (10 + 9)/2)

= (3, 9.5)

For Cluster-02:

Center of Cluster-02

= ((8 + 5 + 7 + 6)/4, (4 + 8 + 5 + 4)/4)

= (6.5, 5.25)

For Cluster-03:

Center of Cluster-03

= ((2 + 1)/2, (5 + 2)/2)

= (1.5, 3.5)

This is completion of Iteration-02.

After second iteration, the center of the three clusters are-

C1(3, 9.5)

C2(6.5, 5.25)

C3(1.5, 3.5)