Clustering in Data Mining

Clustering is an unsupervised Machine Learning-based Algorithm that comprises a group of data points into clusters so that the objects belong to the same group.

Clustering helps to splits data into several subsets. Each of these subsets contains data similar to each other, and these subsets are called clusters. Now that the data from our customer base is divided into clusters, we can make an informed decision about who we think is best suited for this product.

A good clustering algorithm aims to obtain clusters whose:

- The intra-cluster similarities are high, It implies that the data present inside the cluster is similar to one another.
- The inter-cluster similarity is low, and it means each cluster holds data that is not similar to other data.

What is a Cluster?

- A cluster is a subset of similar objects
- A subset of objects such that the distance between any of the two objects in the cluster is less than the distance between any object in the cluster and any object that is not located inside it.

- A connected region of a multidimensional space with a comparatively high density of objects.

What is clustering in Data Mining?

- Clustering is the method of converting a group of abstract objects into classes of similar objects.
- Clustering is a method of partitioning a set of data or objects into a set of significant subclasses called clusters.
- It helps users to understand the structure or natural grouping in a data set and used either as a stand-alone instrument to get a better insight into data distribution or as a pre-processing step for other algorithms

Important points:

- Data objects of a cluster can be considered as one group.
- We first partition the information set into groups while doing cluster analysis. It is based on data similarities and then assigns the levels to the groups.
- The over-classification main advantage is that it is adaptable to modifications, and it helps single out important characteristics that differentiate between distinct groups.

Applications of cluster analysis in data mining:

- In many applications, clustering analysis is widely used, such as data analysis, market research, pattern recognition, and image processing.
- It assists marketers to find different groups in their client base and based on the purchasing patterns. They can characterize their customer groups.
- It helps in allocating documents on the internet for data discovery.
- Clustering is also used in tracking applications such as detection of credit card fraud.
- As a data mining function, cluster analysis serves as a tool to gain insight into the distribution of data to analyze the characteristics of each cluster.
- In terms of biology, It can be used to determine plant and animal taxonomies, categorization of genes with the same functionalities and gain insight into structure inherent to populations.
- It helps in the identification of areas of similar land that are used in an earth observation database and the identification of house groups in a city according to house type, value, and geographical location.

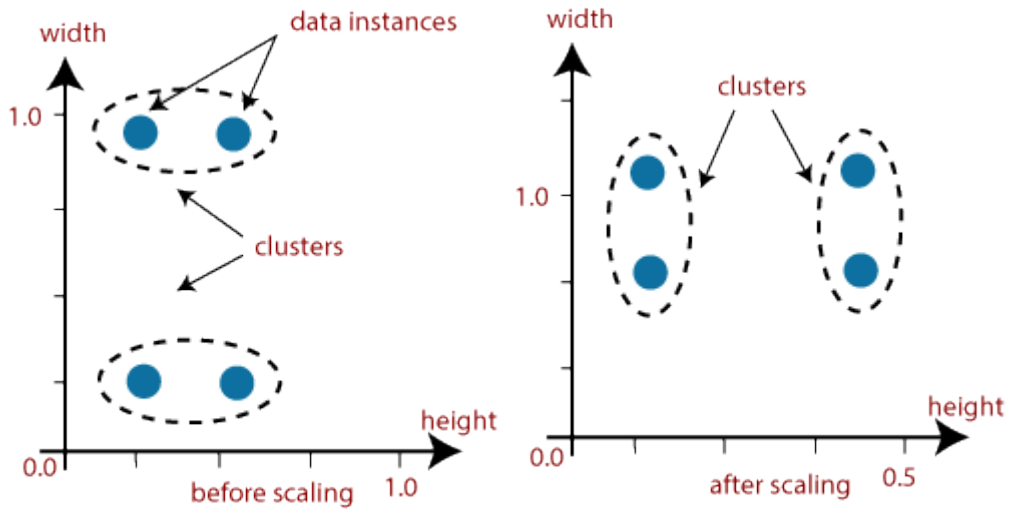
Why is clustering used in data mining?

Clustering analysis has been an evolving problem in data mining due to its variety of applications. The advent of various data clustering tools in the last few years and their comprehensive use in a broad range of applications, including image processing, computational biology, mobile communication, medicine, and economics, must contribute to the popularity of these algorithms. The main issue with the data clustering algorithms is that it cant be standardized. The advanced algorithm may give the best results with one type of data set, but it may fail or perform poorly with other kinds of data set. Although many efforts have been made to standardize the algorithms that can perform well in all situations, no significant achievement has been achieved so far. Many clustering tools have been proposed so far. However, each algorithm has its advantages or disadvantages and can't work on all real situations.

1. Scalability:

Scalability in clustering implies that as we boost the amount of data objects, the time to perform clustering should approximately scale to the complexity order of the algorithm. For example, if we perform K- means clustering, we know it is O(n), where n is the number of objects in the data. If we raise the number of data objects 10 folds, then the time taken to cluster them should also approximately

increase 10 times. It means there should be a linear relationship. If that is not the case, then there is some error with our implementation process.



Showing example where scalability may leads to wrong result

Data should be scalable if it is not scalable, then we can't get the appropriate result. The figure illustrates the graphical example where it may lead to the wrong result.

2. Interpretability:

The outcomes of clustering should be interpretable, comprehensible, and usable.

3. Discovery of clusters with attribute shape:

The clustering algorithm should be able to find arbitrary shape clusters. They should not be limited to only distance measurements that tend to discover a spherical cluster of small sizes.

4. Ability to deal with different types of attributes:

Algorithms should be capable of being applied to any data such as data based on intervals (numeric), binary data, and categorical data.

5. Ability to deal with noisy data:

Databases contain data that is noisy, missing, or incorrect. Few algorithms are sensitive to such data and may result in poor quality clusters.

6. High dimensionality:

The clustering tools should not only able to handle high dimensional data space but also the low-dimensional space.