Programme	**:** Bachelor of Science
Subject	**:** Computer Science
Semester	**:** V
Course Code	**:** CSD102
Course Title	**:** Data Mining
Unit VII	**:** Cluster Analysis
Module Name	**:** Major Clustering Methods - Partitioning Methods, Hierarchical Methods, Density Based Methods.

Notes

Major Clustering Methods

Clustering methods can be classified into different categories.

Following are commonly used methods:

1. Partitioning Method
2. Hierarchical Method
3. Density Based Method

1] Partitioning Method

Partitioning method permits division of data objects into non-overlapping subsets (clusters) such that each data object is in exactly one subset.

Suppose we are given a database of ‘n’ objects and the partitioning method constructs ‘k’ partition of data. Each partition will represent a cluster and k ≤ n. It means that it will classify the data into k groups, which satisfy the following requirements:

1. Each group contains at least one object.
2. Each object must belong to exactly one group.

For a given number of partitions (say k), the partitioning method will create an initial partitioning of k clusters. Then it uses the iterative relocation technique to improve the partitioning by moving objects from one group to other. The general criterion of a good partitioning is that objects in the same cluster are close or related to each other, whereas objects of different clusters are far apart or very different.

Simple illustration of partitioning method is as follows:

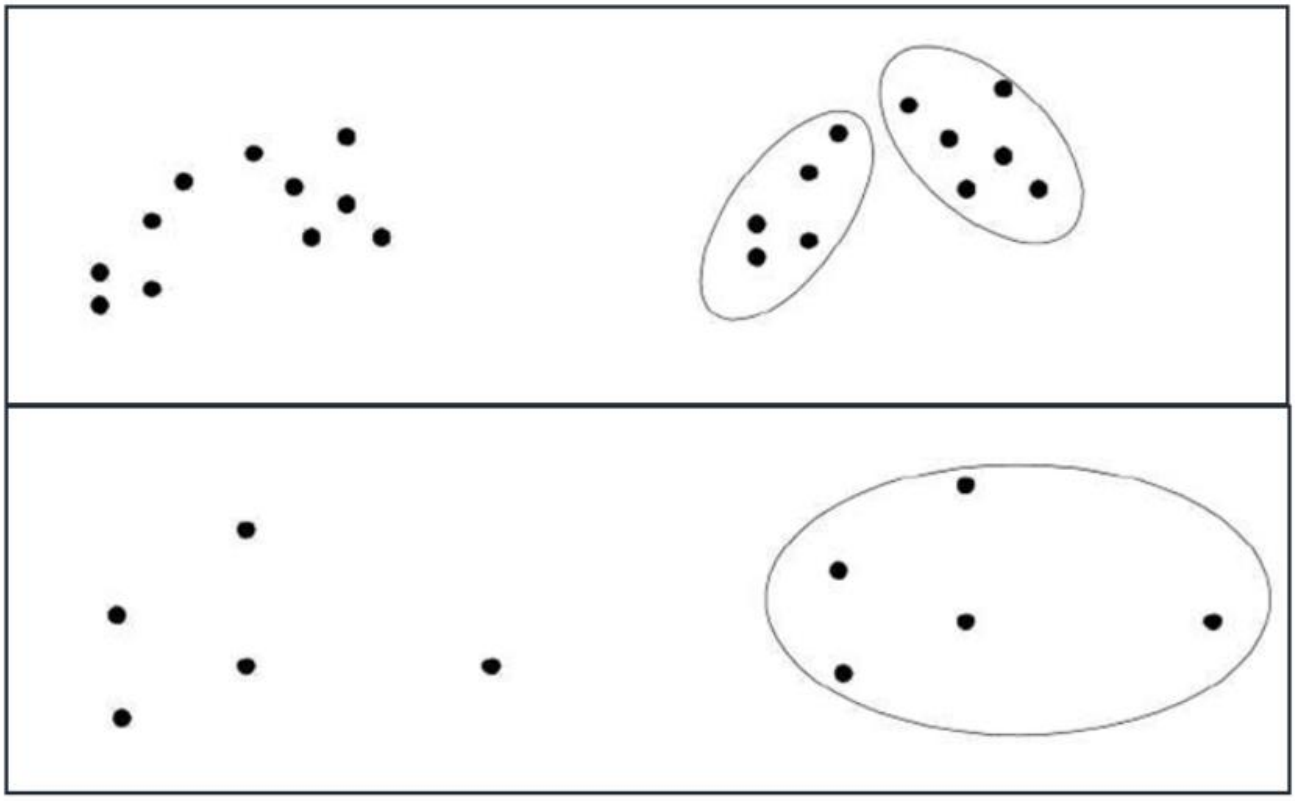

Figure: Partitioning Clustering

2] Hierarchical Method

While partitioning methods meet the basic clustering requirement of organizing a set of objects into a number of exclusive groups, in some situations we may want to partition our data into groups at different levels such as in a hierarchy.

A hierarchical clustering method works by grouping data objects into a hierarchy or “tree” of clusters. i.e., it creates a hierarchical decomposition of the given set of data objects. Example - Consider handwritten character recognition as another example.

A set of handwriting samples may be first partitioned into general groups where each group corresponds to a unique character. Some groups can be further partitioned into subgroups since a character may be written in multiple substantially different ways. If necessary, the hierarchical partitioning can be continued recursively until a desired granularity is reached.

Hierarchical clustering methods can be further classified as either agglomerative or divisive, depending on whether the hierarchical decomposition is formed in a bottom-up (merging) or top-down (splitting) fashion.

There are two approaches are −

A. **Agglomerative Approach**
B. **Divisive Approach**

A. Agglomerative Approach

This approach is also known as the bottom-up approach. In this, we start with each object forming a separate group. It keeps on merging the objects or groups that are close to one another. It keeps on doing so until all of the groups are merged into one or until the termination condition holds. An example for the same is as follows:

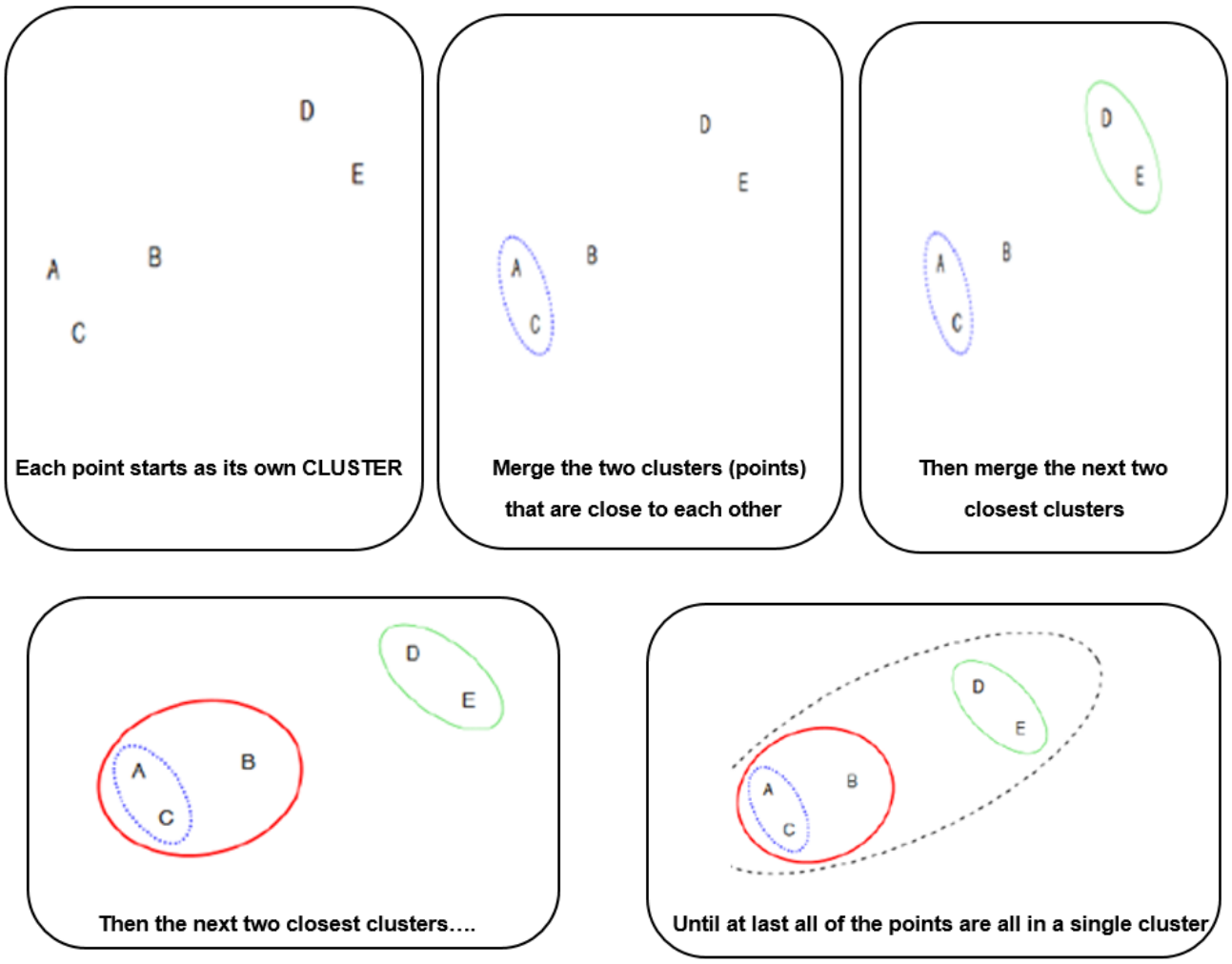

Figure: Agglomerative Approach of Hierarchical Clustering Method

B. Divisive Approach

This approach is also known as the top-down approach and it is just the reverse of Agglomerative Hierarchical approach. In this, we start with all of the objects in the same cluster. In the continuous iteration, a cluster is split up into smaller clusters. It is down until each object in one cluster or the termination condition holds. This method is rigid, i.e., once a merging or splitting is done, it can never be undone.

Figure below shows the comparison of the sequence of steps carried out in Agglomerative and Divisive approaches –

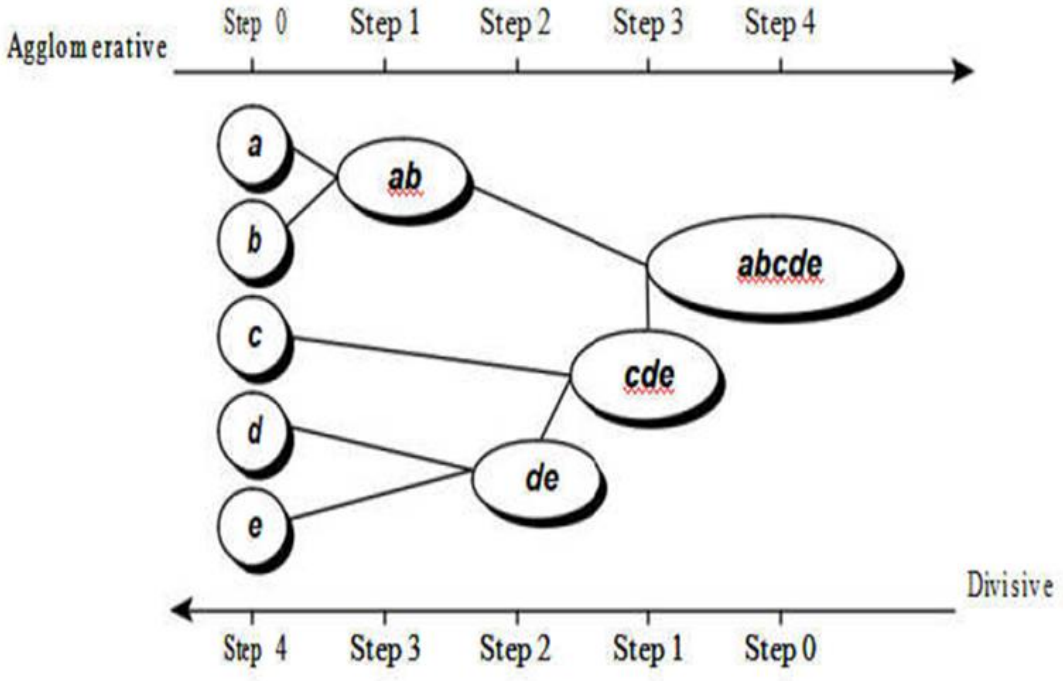

Figure: Agglomerative and Divisive Hierarchical Clustering on data objects a, b, c, d and e

Approaches to Improve Quality of Hierarchical Clustering

The two approaches that are used to improve the quality of hierarchical clustering are:

1. Perform careful analysis of object linkages at each hierarchical partitioning.

2. Integrate hierarchical agglomeration by first using a hierarchical agglomerative algorithm to group objects into micro-clusters, and then performing macro-clustering on the micro-clusters.

3] Density Based Method

Most partitioning methods cluster objects based on the distance between objects. Such methods can find only spherical-shaped clusters and encounter difficulty at discovering clusters of arbitrary shape. To find clusters of arbitrary shape, alternatively, we can model clusters as dense regions in the data space, separated by sparse regions. This is the main strategy behind density-based clustering methods, which can discover clusters of non-spherical shape.

Density based clustering locates regions of high density that are separated from one another by regions of low density. Their general idea is to continue growing the given cluster as long as the density in the neighbourhood exceeds some threshold; that is, for each data point within a given cluster, the neighbourhood of a given radius has to contain at least a minimum number of points. Such a method can be used to filter out noise (outliers) and discover clusters of arbitrary shape.

The two parameters generally used for generating the clusters are:

1. Maximum radius of the neighbourhood (r).
2. Minimum number of points in an (r) neighbourhood of a point (pts).

Figure below shows the clusters of arbitrary shape generated with density based clustering.

Figure: Density Based Method