UNIT IV

CLUSTERING AND APPLICATIONS

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
- Hierarchical Method
- Density-based Method
- Grid-Based Method
- Model-Based Method
- Constraint-based Method

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 neighborhood 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

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

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.

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
- Ratio

And those combined called as mixed-type variables.

CATEGORIZATION OF MAJOR CLUSTERING METHODS

PARTITIONING METHODS

Partitioning Method:

This clustering method classifies the information into multiple groups based on the characteristics and similarity of the data.

Its the data analysts to specify the number of clusters that has to be generated for the clustering methods.

In the partitioning method when database(D) that contains multiple(N) objects then the partitioning method constructs user-specified(K) partitions of the data in which each partition represents a cluster and a particular region.

There are many algorithms that come under partitioning method some of the popular ones are K-Mean, PAM(K-Medoids), CLARA algorithm (Clustering Large Applications) etc.

The working of K Mean algorithm in detail. K-Mean (A centroid based Technique):

The K means algorithm takes the input parameter K from the user and partitions the dataset containing N objects into K clusters so that resulting similarity among the data objects inside the group (intracluster) is high but the similarity of data objects with the data objects from outside the cluster is low (intercluster).

The similarity of the cluster is determined with respect to the mean value of the cluster.

It is a type of square error algorithm.

At the start randomly k objects from the dataset are chosen in which each of the objects represents a cluster mean(centre).

For the rest of the data objects, they are assigned to the nearest cluster based on their distance from the cluster mean.

The new mean of each of the cluster is then calculated with the added data objects.

Algorithm: K mean:

Input:

K: The number of clusters in which the dataset has to be divided

D: A dataset containing N number of objects

Output:

A dataset of K clusters

Method:

1. Randomly assign K objects from the dataset(D) as cluster centres(C)
2. (Re) Assign each object to which object is most similar based upon mean values.
3. Update Cluster means, i.e., Recalculate the mean of each cluster with the updated values.
4. Repeat Step 2 until no change occurs.

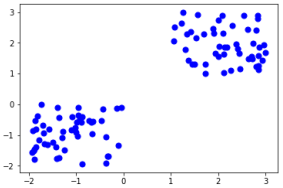


Figure – K-mean ClusteringFlowchart:

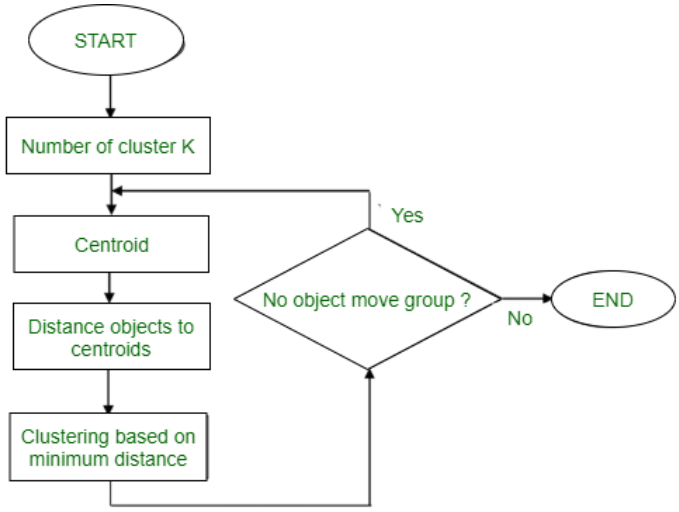


Figure – K-mean ClusteringExample: Suppose we want to group the visitors to a website using just their age as follows:

16, 16, 17, 20, 20, 21, 21, 22, 23, 29, 36, 41, 42, 43, 44, 45, 61, 62, 66

INITIAL CLUSTER:

K=2

Centroid(C1) = 16 [16]

Centroid(C2) = 22 [22]

Note: These two points are chosen randomly from the dataset. Iteration-1:

C1 = 16.33 [16, 16, 17]

C2 = 37.25 [20, 20, 21, 21, 22, 23, 29, 36, 41, 42, 43, 44, 45, 61, 62, 66]

Iteration-2:

C1 = 19.55 [16, 16, 17, 20, 20, 21, 21, 22, 23]

C2 = 46.90 [29, 36, 41, 42, 43, 44, 45, 61, 62, 66]

Iteration-3:

C1 = 20.50 [16, 16, 17, 20, 20, 21, 21, 22, 23, 29]

C2 = 48.89 [36, 41, 42, 43, 44, 45, 61, 62, 66]

Iteration-4:

C1 = 20.50 [16, 16, 17, 20, 20, 21, 21, 22, 23, 29]

C2 = 48.89 [36, 41, 42, 43, 44, 45, 61, 62, 66]

No change Between Iteration 3 and 4, so we stop. Therefore we get the clusters (16-29) and (36-66) as 2 clusters we get using K Mean Algorithm.

HIERARCHICAL CLUSTERING

A Hierarchical clustering method works via grouping data into a tree of clusters. Hierarchical clustering begins by treating every data point as a separate cluster. Then, it repeatedly executes the subsequent steps:

1. Identify the 2 clusters which can be closest together, and
2. Merge the 2 maximum comparable clusters. We need to continue these steps until all the clusters are merged together.

In Hierarchical Clustering, the aim is to produce a hierarchical series of nested clusters.

A diagram called Dendrogram (A Dendrogram is a tree-like diagram that statistics the sequences of merges or splits) graphically represents this hierarchy and is an inverted tree that describes the order in which factors are merged (bottom-up view) or clusters are broken up (top-down view).

Hierarchical clustering is a method of cluster analysis in data mining that creates a hierarchical representation of the clusters in a dataset.

The method starts by treating each data point as a separate cluster and then iteratively combines the closest clusters until a stopping criterion is reached.

The result of hierarchical clustering is a tree-like structure, called a dendrogram, which illustrates the hierarchical relationships among the clusters.

Hierarchical clustering has a number of advantages over other clustering methods, including:

1. The ability to handle non-convex clusters and clusters of different sizes and densities.
2. The ability to handle missing data and noisy data.
3. The ability to reveal the hierarchical structure of the data, which can be useful for understanding the relationships among the clusters. However, it also has some drawbacks, such as:

4. The need for a criterion to stop the clustering process and determine the final number of clusters.
5. The computational cost and memory requirements of the method can be high, especially for large datasets.
6. The results can be sensitive to the initial conditions, linkage criterion, and distance metric used.

   In summary, Hierarchical clustering is a method of data mining that groups similar data points into clusters by creating a hierarchical structure of the clusters.

7. This method can handle different types of data and reveal the relationships among the clusters. However, it can have high computational cost and results can be sensitive to some conditions.

1. AGGLOMERATIVE:

Initially consider every data point as an individual Cluster and at every step, merge the nearest pairs of the cluster. (It is a bottom-up method).

At first, every dataset is considered an individual entity or cluster.

At every iteration, the clusters merge with different clusters until one cluster is formed.

The algorithm for Agglomerative Hierarchical Clustering is:

- Calculate the similarity of one cluster with all the other clusters (calculate proximity matrix)
- Consider every data point as an individual cluster
- Merge the clusters which are highly similar or close to each other.
- Recalculate the proximity matrix for each cluster
- Repeat Steps 3 and 4 until only a single cluster remains.

Let’s see the graphical representation of this algorithm using a dendrogram.

Note: This is just a demonstration of how the actual algorithm works no calculation has been performed below all the proximity among the clusters is assumed.

Let’s say we have six data points A, B, C, D, E, and F.

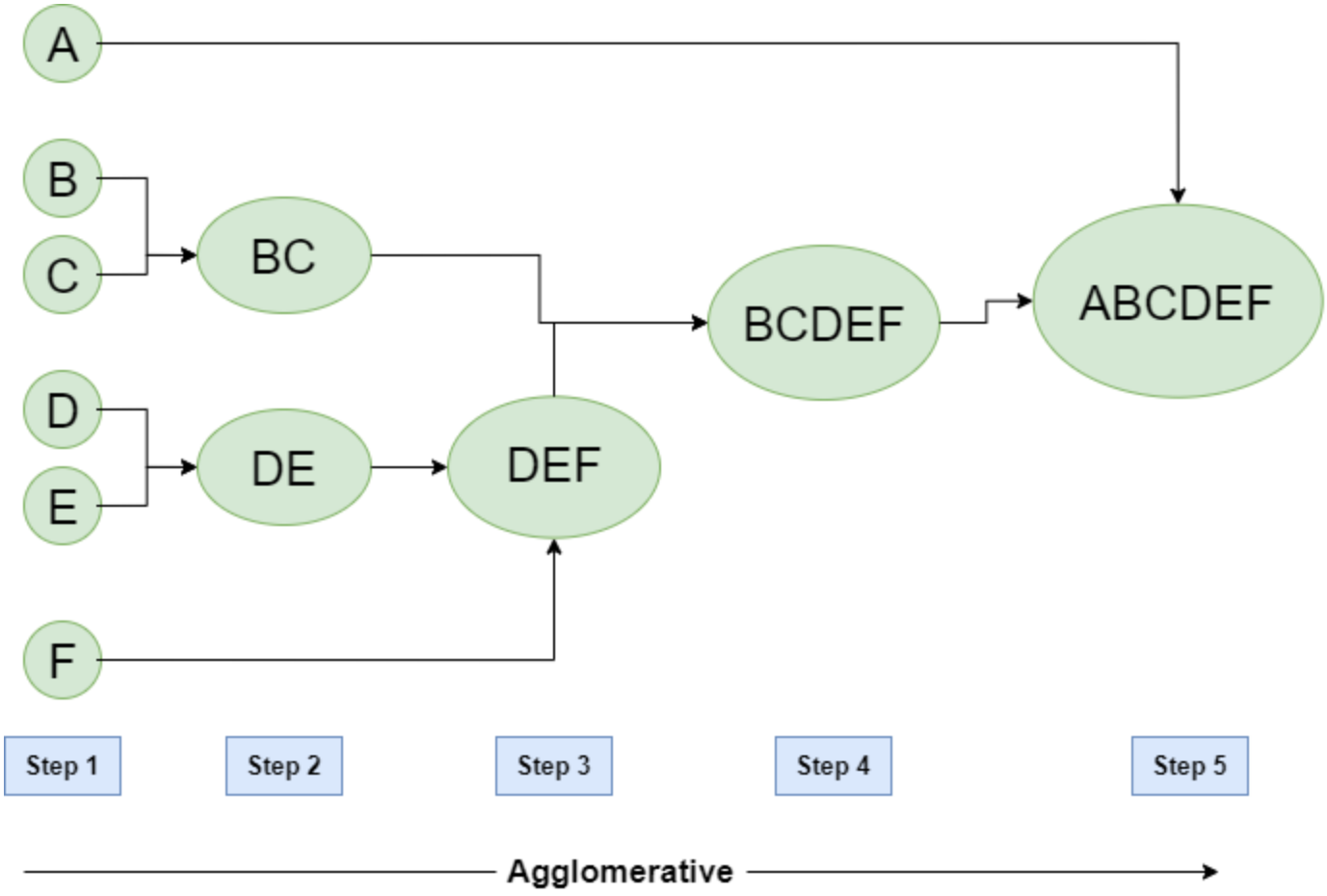


Figure – Agglomerative Hierarchical clustering

- **Step-1:** Consider each alphabet as a single cluster and calculate the distance of one cluster from all the other clusters.
- **Step-2:** In the second step comparable clusters are merged together to form a single cluster. Let's say cluster (B) and cluster (C) are very similar to each other therefore we merge them in the second step similarly to cluster (D) and (E) and at last, we get the clusters [(A), (BC), (DE), (F)]
- **Step-3:** We recalculate the proximity according to the algorithm and merge the two nearest clusters([(DE), (F)]) together to form new clusters as [(A), (BC), (DEF)]
- **Step-4:** Repeating the same process; The clusters DEF and BC are comparable and merged together to form a new cluster. We're now left with clusters [(A), (BCDEF)].
- **Step-5:** At last the two remaining clusters are merged together to form a single cluster [(ABCDEF)].

2. DIVISIVE:

We can say that Divisive Hierarchical clustering is precisely the opposite of Agglomerative Hierarchical clustering.

In Divisive Hierarchical clustering, we take into account all of the data points as a single cluster and in every iteration, we separate the data points from the clusters which aren't comparable.

In the end, we are left with N clusters.

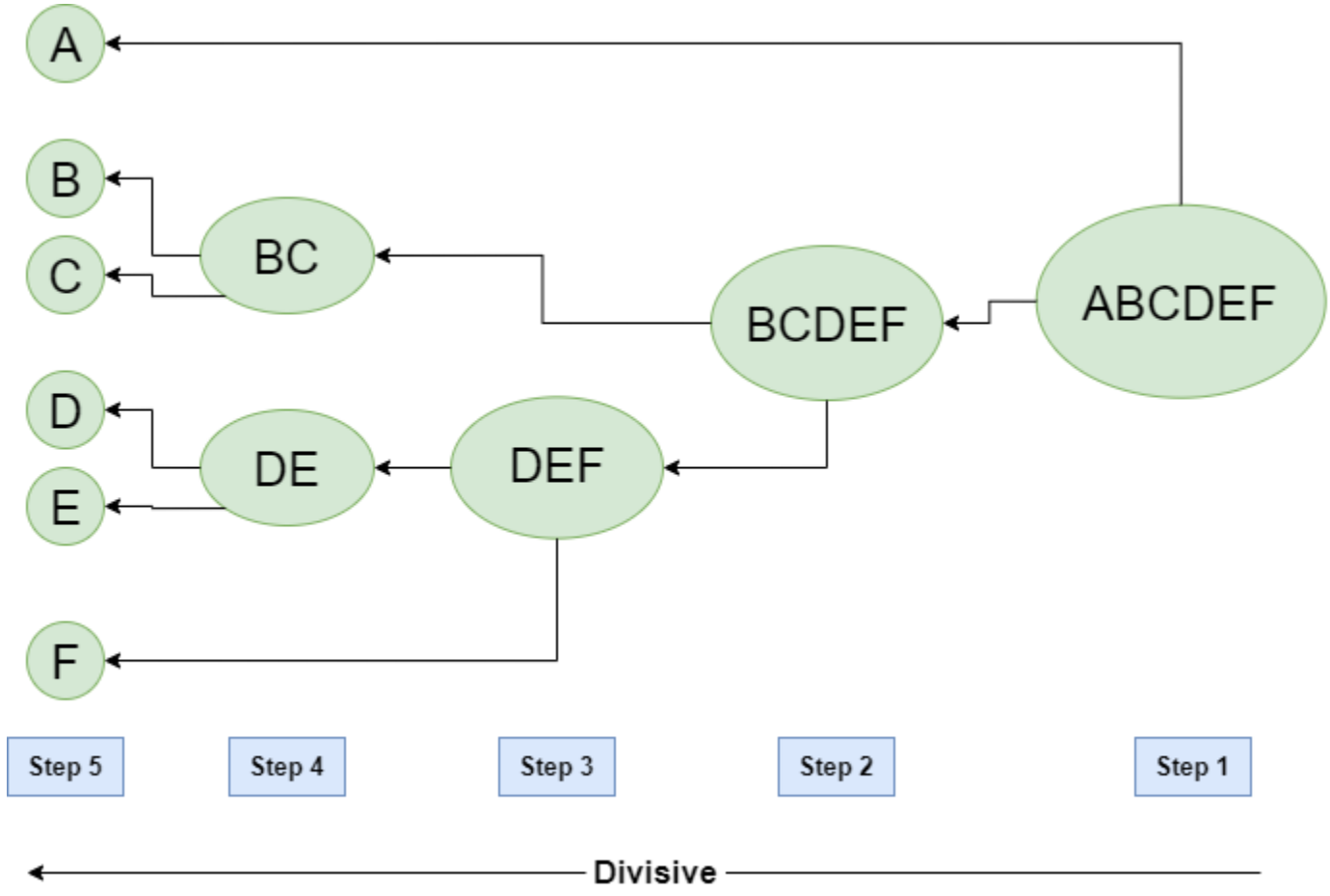


Figure – Divisive Hierarchical clustering

DENSITY-BASED CLUSTERING

Density-based clustering refers to a method that is based on local cluster criterion, such as density connected points.

Density-Based Clustering refers to one of the most popular unsupervised learning methodologies used in model building and machine learning algorithms. The data points in the region separated by two clusters of low point density are considered as noise. The surroundings with a radius ε of a given object are known as the ε neighborhood of the object. If the ε neighborhood of the object comprises at least a minimum number, MinPts of objects, then it is called a core object.

Density-Based Clustering - Background

There are two different parameters to calculate the density-based clustering

E_{PS}: It is considered as the maximum radius of the neighborhood.

MinPts: MinPts refers to the minimum number of points in an Eps neighborhood of that point.

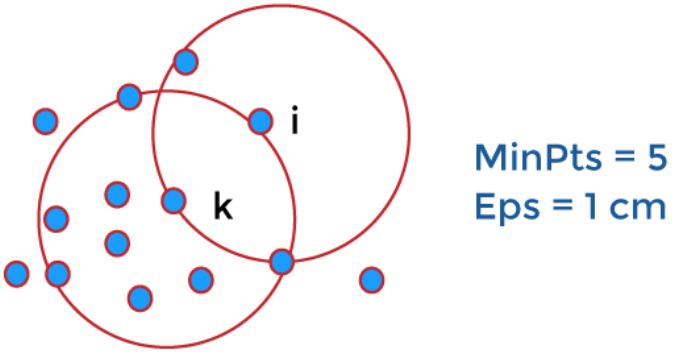
NEps (i) : { k belongs to D and dist (i,k) < = Eps}

DIRECTLY DENSITY REACHABLE:

A point i is considered as the directly density reachable from a point k with respect to Eps, MinPts if i belongs to NEps(k)

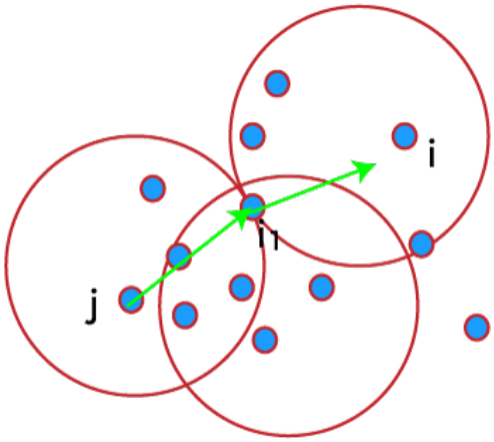
CORE POINT CONDITION:

NEps (k) >= MinPts



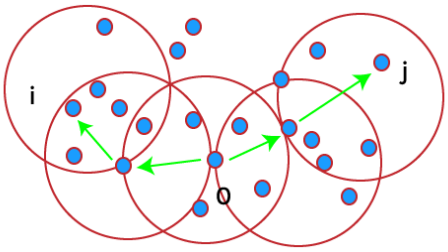
DENSITY REACHABLE:

A point denoted by i is a density reachable from a point j with respect to Eps, MinPts if there is a sequence chain of a point i1,…., in, i1 = j, pn = i such that i_i + 1 is directly density reachable from i_i.



DENSITY CONNECTED:

A point i refers to density connected to a point j with respect to Eps, MinPts if there is a point o such that both i and j are considered as density reachable from o with respect to Eps and MinPts.



WORKING OF DENSITY-BASED CLUSTERING

Suppose a set of objects is denoted by D', we can say that an object I is directly density reachable form the object j only if it is located within the ε neighborhood of j, and j is a core object.

An object i is density reachable form the object j with respect to ε and MinPts in a given set of objects, D' only if there is a sequence of object chains point i1,…., in, i1 = j, pn = i such that i_i + 1 is directly density reachable from i_i with respect to ε and MinPts.

An object i is density connected object j with respect to ε and MinPts in a given set of objects, D' only if there is an object o belongs to D such that both point i and j are density reachable from o with respect to ε and MinPts.

MAJOR FEATURES OF DENSITY-BASED CLUSTERING

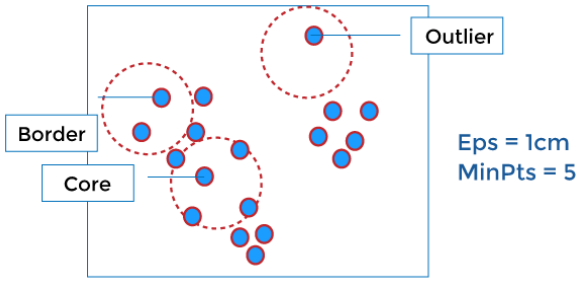
The primary features of Density-based clustering are given below.

- It is a scan method.
- It requires density parameters as a termination condition.
- It is used to manage noise in data clusters.
- Density-based clustering is used to identify clusters of arbitrary size.

DENSITY-BASED CLUSTERING METHODS

DBSCAN

DBSCAN stands for Density-Based Spatial Clustering of Applications with Noise. It depends on a density-based notion of cluster. It also identifies clusters of arbitrary size in the spatial database with outliers.



OPTICS

OPTICS stands for Ordering Points to Identify the Clustering Structure.

It gives a significant order of database with respect to its density-based clustering structure.

The order of the cluster comprises information equivalent to the density-based clustering related to a long range of parameter settings.

OPTICS methods are beneficial for both automatic and interactive cluster analysis, including determining an intrinsic clustering structure.

DENCLUE

Density-based clustering by Hinnebirg and Kiem.

It enables a compact mathematical description of arbitrarily shaped clusters in high dimension state of data, and it is good for data sets with a huge amount of noise.

STING is a Grid-Based Clustering Technique. In STING, the dataset is recursively divided in a hierarchical manner. After the dataset, each cell is divided into a different number of cells. And after the cell, the statistical measures of the cell are collected, which helps answer the query as quickly as possible.

GRID-BASED METHOD IN DATA MINING:

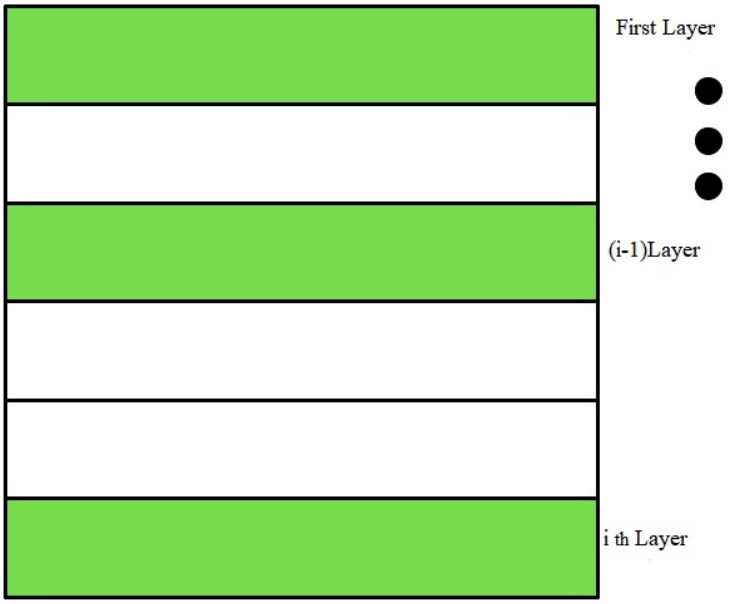
In Grid-Based Methods, the space of instance is divided into a grid structure. Clustering techniques are then applied using the Cells of the grid, instead of individual data points, as the base units. The biggest advantage of this method is to improve the processing time.

STATISTICAL INFORMATION GRID(STING):

A STING is a grid-based clustering technique. It uses a multidimensional grid data structure that quantifies space into a finite number of cells. Instead of focusing on data points, it focuses on the value space surrounding the data points.

In STING, the spatial area is divided into rectangular cells and several levels of cells at different resolution levels. High-level cells are divided into several low-level cells.

In STING Statistical Information about attributes in each cell, such as mean, maximum, and minimum values, are precomputed and stored as statistical parameters. These statistical parameters are useful for query processing and other data analysis tasks.



The statistical parameter of higher-level cells can easily be computed from the parameters of the lower-level cells.

HOW STING WORK:

Step 1: Determine a layer, to begin with.

Step 2: For each cell of this layer, it calculates the confidence interval or estimated range of probability that this is cell is relevant to the query.

Step 3: From the interval calculate above, it labels the cell as relevant or not relevant.

Step 4: If this layer is the bottom layer, go to point 6, otherwise, go to point 5.

Step 5: It goes down the hierarchy structure by one level. Go to point 2 for those cells that form the relevant cell of the high-level layer.

Step 6: If the specification of the query is met, go to point 8, otherwise go to point 7.

Step 7: Retrieve those data that fall into the relevant cells and do further processing. Return the result that meets the requirement of the query. Go to point 9.

Step 8: Find the regions of relevant cells. Return those regions that meet the requirement of the query. Go to point 9.

Step 9: Stop or terminate.

ADVANTAGES:

- Grid-based computing is query-independent because the statistics stored in each cell represent a summary of the data in the grid cells and are query-independent.
- The grid structure facilitates parallel processing and incremental updates.

DISADVANTAGE:

- The main disadvantage of Sting (Statistics Grid) as we know, all cluster boundaries are either horizontal or vertical, so no diagonal boundaries are detected.

Whenever we talk about data analysis, the term outliers often come to our mind.

As the name suggests, "outliers" refer to the data points that exist outside of what is to be expected.

The major thing about the outliers is what you do with them.

If you are going to analyze any task to analyze data sets, you will always have some assumptions based on how this data is generated.

If you find some data points that are likely to contain some form of error, then these are definitely outliers, and depending on the context, you want to overcome those errors.

The data mining process involves the analysis and prediction of data that the data holds. In 1969, Grubbs introduced the first definition of outliers.

DIFFERENCE BETWEEN OUTLIERS AND NOISE

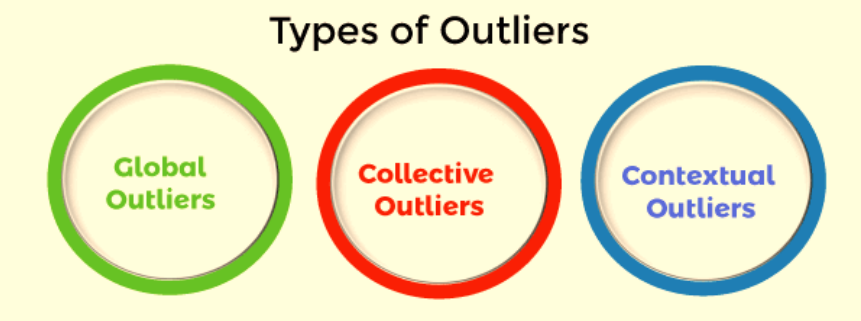
Any unwanted error occurs in some previously measured variable, or there is any variance in the previously measured variable called noise.

Before finding the outliers present in any data set, it is recommended first to remove the noise.

Types of Outliers

Outliers are divided into three different types

1. Global or point outliers
2. Collective outliers
3. Contextual or conditional outliers



GLOBAL OUTLIERS

Global outliers are also called point outliers. Global outliers are taken as the simplest form of outliers. When data points deviate from all the rest of the data points in a given data set, it is known as the global outlier. In most cases, all the outlier detection procedures are targeted to determine the global outliers. The green data point is the global outlier.