THE NEW COLLEGE

(AN AUTONOUMOUS INSTITUTION AFFILIATED TO THE UNIVERSITY OF MADRAS
& ACCREDITED BY NAAC WITH 'A++' GRADE IN THE 4TH CYCLE)
SPONSORED BY : THE MUSLIM EDUCATIONAL ASSOCIATION OF SOUTHERN INDIA
(MEASI)

YA ALLAH
INCREASE MY KNOWLEDGE

DEPARTMENT OF COMPUTER APPLICATIONS

E-CONTENT UNIT V

Subject : **DATA WAREHOUSING & DATA MINING**

Class : **III BCA A**

E-mail : **abdulrasheedh@thenewcollege.edu.in**

Dr. J. Abdul Rasheedh M.C.A., M.Phil., Ph.D.
Assistant Professor,
P.G. Department of Computer Science,
The New College (Autonomous), Chennai-14.

UNIT – V:

Cluster Analysis – Types of Data – Categorization of Major Clustering Methods – K-means– Partitioning Methods – Hierarchical Methods – Density-Based Methods –Grid Based Methods – Model-Based ClusteringMethods – Clustering High Dimensional Data – Constraint – Based Cluster Analysis – Outlier Analysis – Data Mining Applications.

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.

Types of data that can be mined

1. Data stored in the database

A database is also called a database management system or DBMS. Every DBMS stores data that are related to each other in a way or the other. It also has a set of software programs that are used to manage data and provide easy access to it. These software programs serve a lot of purposes, including defining structure for database, making sure that the stored information remains secured and consistent, and managing different types of data access, such as shared, distributed, and concurrent.

A relational database has tables that have different names, attributes, and can store rows or records of large data sets. Every record stored in a table has a unique key. Entity-relationship model is created to provide a representation of a relational database that features entities and the relationships that exist between them.

2. Data warehouse

A data warehouse is a single data storage location that collects data from multiple sources and then stores it in the form of a unified plan. When data is stored in a data warehouse, it undergoes cleaning, integration, loading, and refreshing. Data stored in a data warehouse is

organized in several parts. If you want information on data that was stored 6 or 12 months back, you will get it in the form of a summary.

3. Transactional data

Transactional database stores record that are captured as transactions. These transactions include flight booking, customer purchase, click on a website, and others. Every transaction record has a unique ID. It also lists all those items that made it a transaction.

4. Other types of data

We have a lot of other types of data as well that are known for their structure, semantic meanings, and versatility. They are used in a lot of applications. Here are a few of those data types: data streams, engineering design data, sequence data, graph data, spatial data, multimedia data, and more.

Clustering Methods:

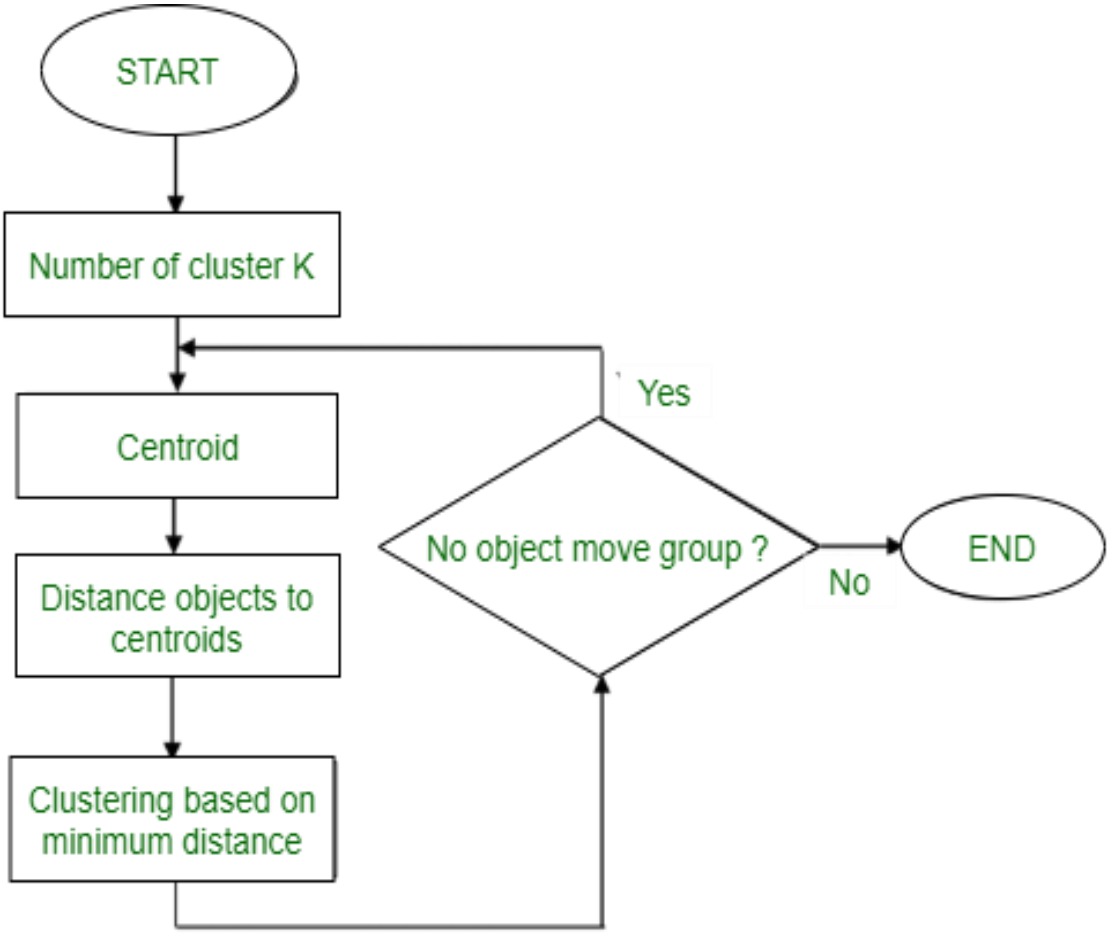
The clustering methods can be classified into the following categories:

- 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:

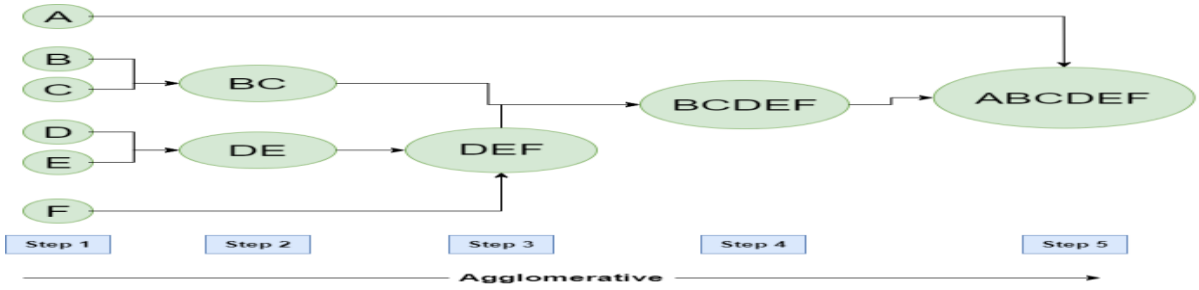
- One objective should only belong to only one group.
- There should be no group without even a single purpose.

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

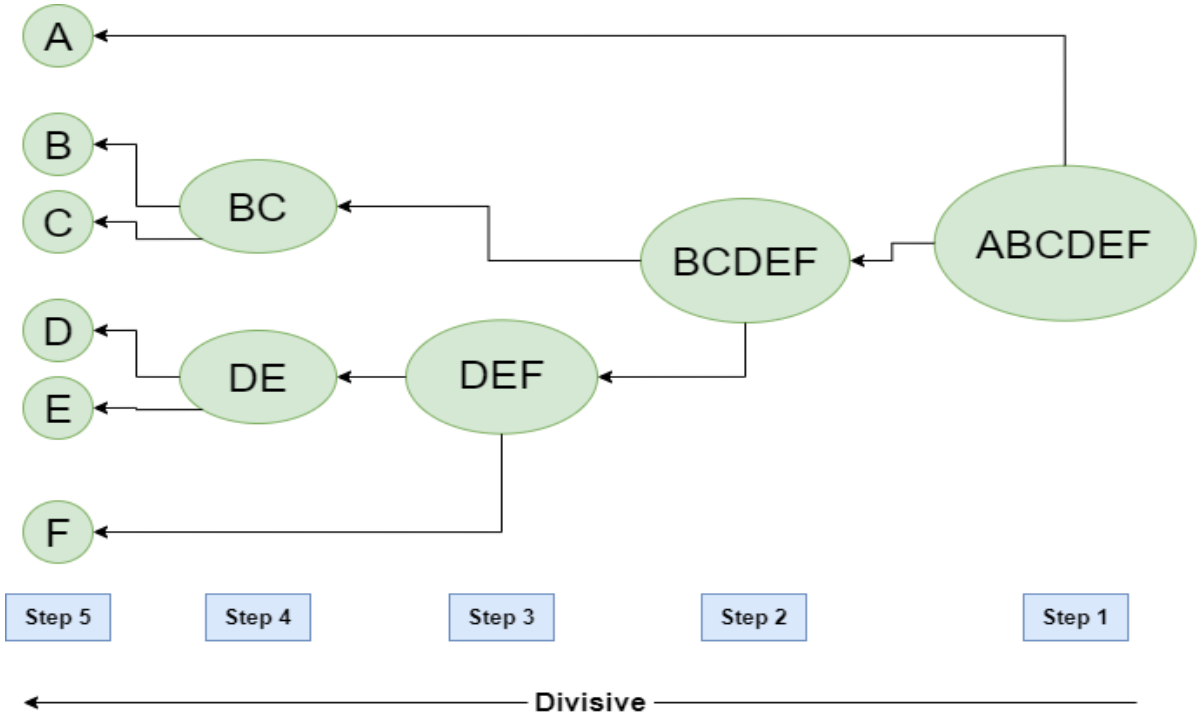


Hierarchical Method: In this method, a hierarchical decomposition of the given set of data objects is created. We can classify hierarchical methods and will be able to know the purpose of classification on the basis of how the hierarchical decomposition is formed. There are two types of approaches for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



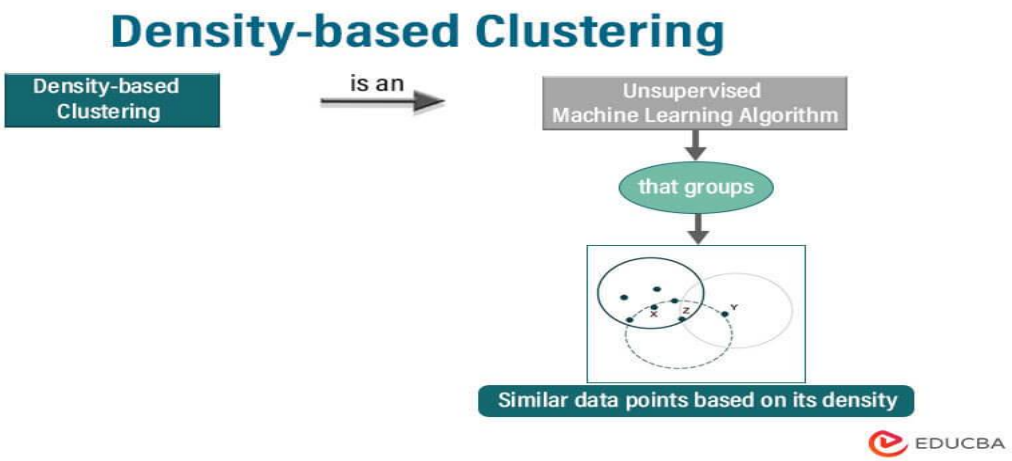
- **Divisive Approach:** The divisive approach is also known as the top-down approach. In this approach, we would start with the data objects that are in the same cluster. The group of individual clusters is divided into small clusters by continuous iteration. The iteration continues until the condition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

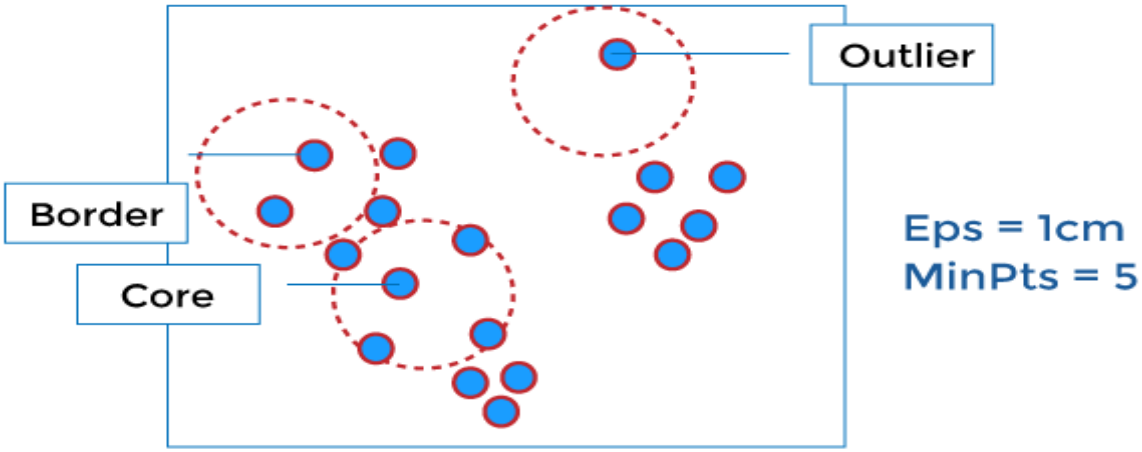
Density-Based Method: The density-based method mainly focuses on density. In this method, the given cluster will keep on growing continuously as long as the density in the neighbourhood exceeds some threshold, i.e, for each data point within a given cluster. The radius of a given cluster has to contain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



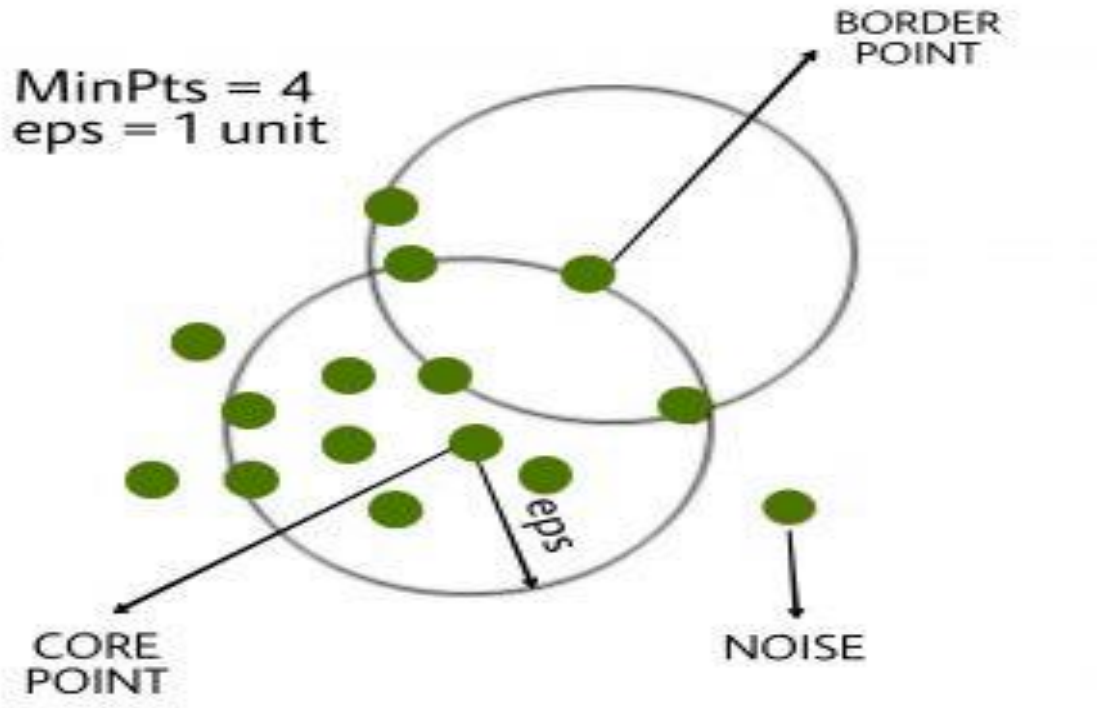
Parameters Required For DBSCAN Algorithm

1. **eps**: It defines the neighborhood around a data point i.e. if the distance between two points is lower or equal to 'eps' then they are considered neighbors. If the eps value is chosen too small then a large part of the data will be considered as an outlier. If it is chosen very large then the clusters will merge and the majority of the data points will be in the same clusters. One way to find the eps value is based on the ***k-distance graph***.
2. **MinPts**: Minimum number of neighbors (data points) within eps radius. The larger the dataset, the larger value of MinPts must be chosen. As a general rule, the minimum MinPts can be derived from the number of dimensions D in the dataset as, MinPts >= D+1. The minimum value of MinPts must be chosen at least 3.

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.



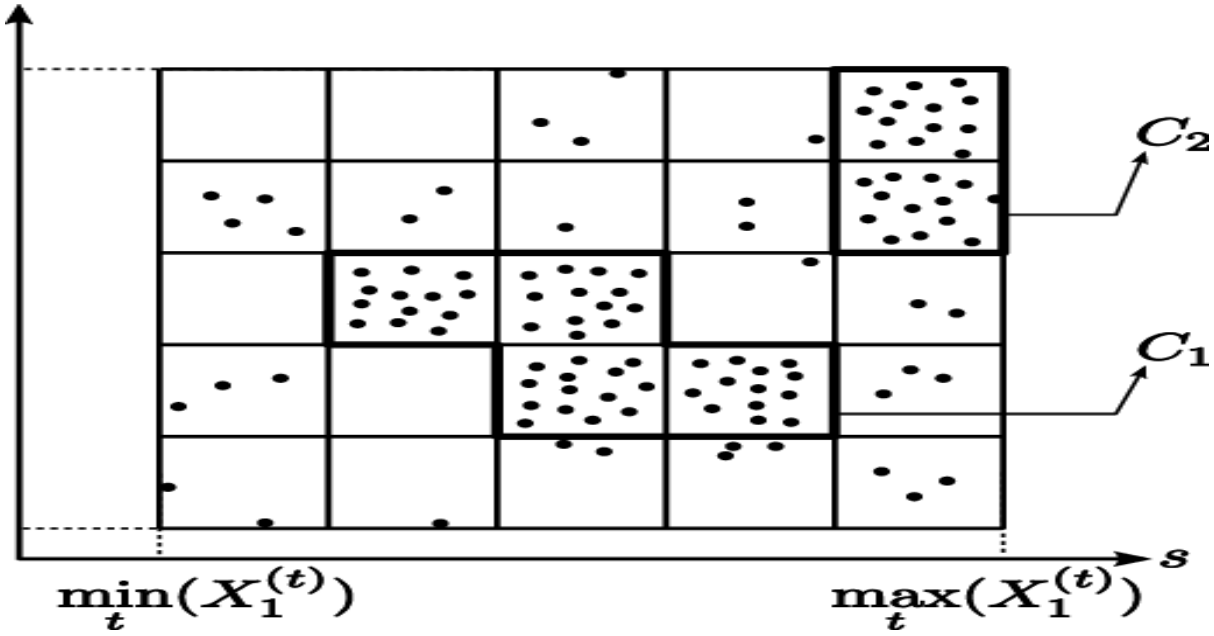
In this algorithm, we have 3 types of data points.
Core Point*: A point is a core point if it has more than MinPts points within eps.*
Border Point*: A point which has fewer than MinPts within eps but it is in the neighborhood of a core point.*
Noise or outlier*: A point which is not a core point or border point.*



Steps Used In DBSCAN Algorithm

1. Find all the neighbor points within eps and identify the core points or visited with more than MinPts neighbors.
2. For each core point if it is not already assigned to a cluster, create a new cluster.
3. Find recursively all its density-connected points and assign them to the same cluster as the core point.
   A point *a* and *b* are said to be density connected if there exists a point *c* which has a sufficient number of points in its neighbors and both points *a* and *b* are within the *eps distance*. This is a chaining process. So, if *b* is a neighbor of *c*, *c* is a neighbor of *d*, and *d* is a neighbor of *e*, which in turn is neighbor of *a* implying that *b* is a neighbor of *a*.
4. Iterate through the remaining unvisited points in the dataset. Those points that do not belong to any cluster are noise.

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.



Basic Grid-based Algorithm

1. Define a set of grid-cells
2. Assign objects to the appropriate grid cell and compute the density of each cell.
3. Eliminate cells, whose density is below a certain threshold t.
4. Form clusters from contiguous (adjacent) groups of dense cells (usually minimizing a given objective function).

Advantages of Grid-based Clustering Algorithms:

- fast:
  - No distance computations
  - Clustering is performed on summaries and not individual objects; complexity is usually O(#-populated-grid-cells) and not O(#objects)
  - Easy to determine which clusters are neighboring
- Shapes are limited to union of rectangular grid-cells

Model-Based Method:

Model-based clustering is a statistical approach to data clustering. The observed (multivariate) data is considered to have been created from a finite combination of component models. Each component model is a probability distribution, generally a parametric multivariate distribution.

For instance, in a multivariate Gaussian mixture model, each component is a multivariate Gaussian distribution. The component responsible for generating a particular observation determines the cluster to which the observation belongs.

Model-based clustering is a try to advance the fit between the given data and some mathematical model and is based on the assumption that data are created by a combination of a basic probability distribution.

There are the following types of model-based clustering are as follows −

Statistical approach − Expectation maximization is a popular iterative refinement algorithm. An extension to k-means −

- It can assign each object to a cluster according to weight (probability distribution).
- New means are computed based on weight measures.

The basic idea is as follows −

- It can start with an initial estimate of the parameter vector.
- It can be used to iteratively rescore the designs against the mixture density made by the parameter vector.
- It is used to rescored patterns are used to update the parameter estimates.
- It can be used to pattern belonging to the same cluster if they are placed by their scores in a particular component.

Algorithm

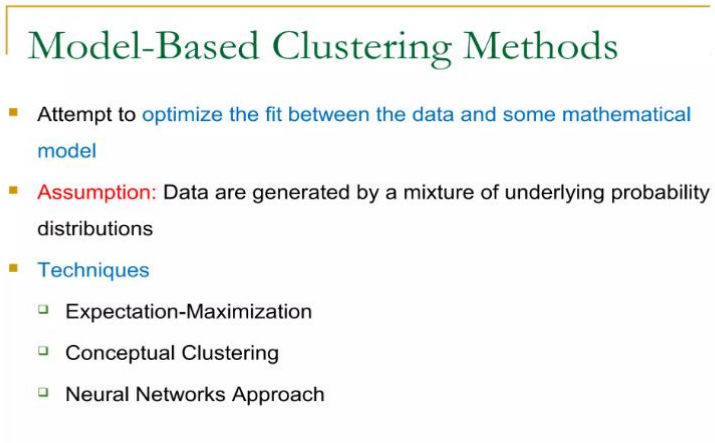
- Initially, assign k cluster centers randomly.
- It can be iteratively refined the clusters based on two steps are as follows −

Expectation step − It can assign each data point X_i to cluster C_i with the following probability

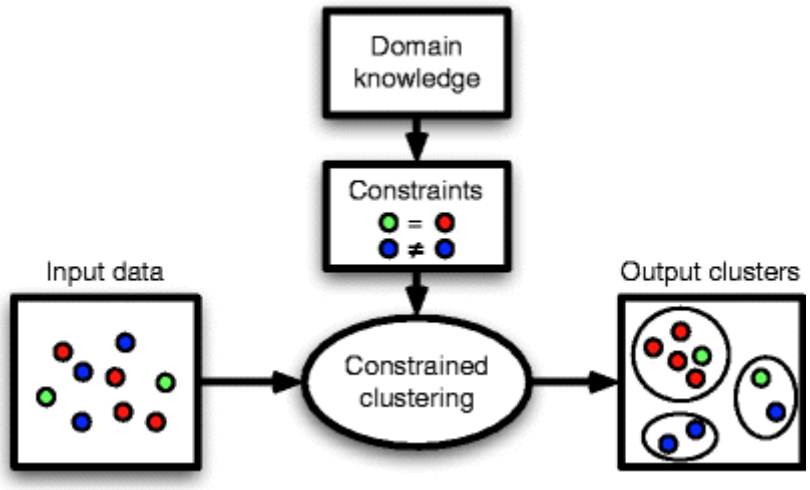
$$P(Xi∈Ck)=P(Ck□Xi)=P(Ck)P(Xi□Ck)P(Xi)P(Xi∈Ck)=P(Ck□Xi)=P(Ck)P(Xi□Ck)P(Xi)$$

Maximization step − It can be used to estimate of model parameter

$$mk=1N∑i=1NXiP(Xi∈Ck)XjP(Xi)∈Cjmk=1N∑i=1NXiP(Xi∈Ck)XjP(Xi)∈Cj$$



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.



Constraint-based clustering finds clusters that satisfy user-stated preferences or constraints. It is based on the nature of the constraints, constraint-based clustering can adopt instead of different approaches. There are several categories of constraints which are as follows −

- **Constraints on individual objects** − It can define constraints on the objects to be clustered. In a real estate application, for instance, one can like to spatially cluster only those luxury mansions worth over a million dollars. This constraint confines the collection of objects to be clustered. It can simply be managed by preprocessing (e.g., implementing selection using an SQL query), after which the problem decreases to an example of unconstrained clustering.
- **Constraints on the selection of clustering parameters** − A user can like to set a desired area for each clustering parameter. Clustering parameters are generally quite specific to the given clustering algorithm. Examples of parameters contain k, the desired number of clusters in a k-means algorithm; or ε (the radius) and MinPts (the minimum number of points) in the DBSCAN algorithm.
  Although such user-stated parameters can strongly hold the clustering results, they are generally confined to the algorithm itself. Therefore, their fine-tuning and processing are generally not treated as a form of constraint-based clustering.
- **Constraints on distance or similarity functions** − It can define several distances or similarity functions for definite attributes of the objects to be clustered, or different distance measures for limited pairs of objects. When clustering sportsmen, for instance, it can use several weighting schemes for height, body weight, age, and skill level.

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.

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

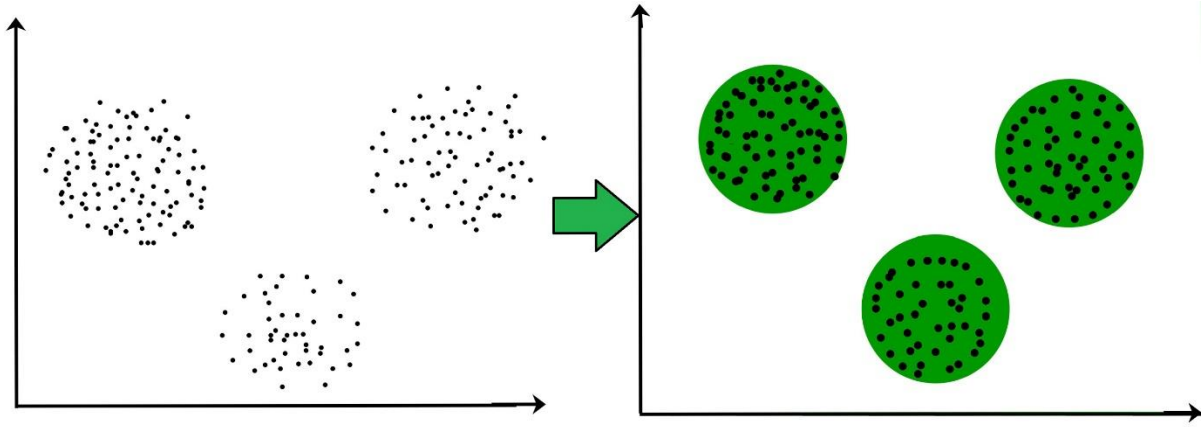
The main idea of cluster analysis is that it would arrange all the data points by forming clusters like cars cluster which contains all the cars, bikes clusters which contains all the bikes, etc.

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

Clustering High Dimensional Data

Clustering is basically a type of unsupervised learning method. An unsupervised learning method is a method in which we draw references from datasets consisting of input data without labeled responses.

Clustering is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group and dissimilar to the data points in other groups.



Challenges of Clustering High-Dimensional Data:

Clustering of the High-Dimensional Data return the group of objects which are clusters.

It is required to group similar types of objects together to perform the cluster analysis of high-dimensional data, But the High-Dimensional data space is huge and it has complex data types and attributes.

A major challenge is that we need to find out the set of attributes that are present in each cluster.

A cluster is defined and characterized based on the attributes present in the cluster. Clustering High-Dimensional Data we need to search for clusters and find out the space for the existing clusters.

The High-Dimensional data is reduced to low-dimension data to make the clustering and search for clusters simple. some applications need the appropriate models of clusters, especially the high-dimensional data. clusters in the high-dimensional data are significantly small. the conventional distance measures can be ineffective. Instead, To find the hidden clusters in high-dimensional data we need to apply sophisticated techniques that can model correlations among the objects in subspaces.

Outlier Analysis

Outlier is a data object that deviates significantly from the rest of the data objects and behaves in a different manner. An outlier is an object that deviates significantly from the rest of the objects. They can be caused by measurement or execution errors. The analysis of outlier data is referred to as outlier analysis or outlier mining.

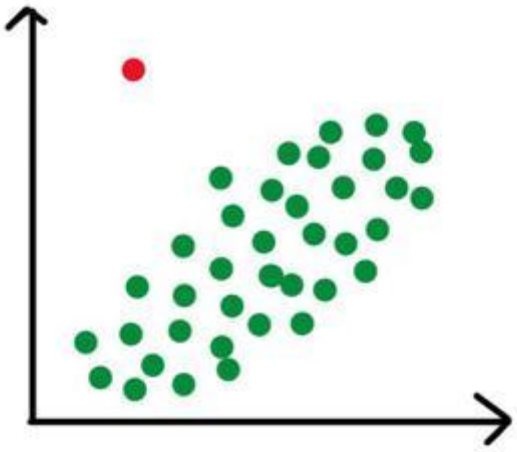
An outlier cannot be termed as a noise or error. Instead, they are suspected of not being generated by the same method as the rest of the data objects.

Outliers are of three types, namely –

1. Global (or Point) Outliers
2. Collective Outliers
3. Contextual (or Conditional) Outliers

1. Global Outliers

1. Definition: Global outliers are data points that deviate significantly from the overall distribution of a dataset.
2. Causes: Errors in data collection, measurement errors, or truly unusual events can result in global outliers.
3. Impact: Global outliers can distort data analysis results and affect machine learning model performance.
4. Detection: Techniques include statistical methods (e.g., z-score, Mahalanobis distance), machine learning algorithms (e.g., isolation forest, one-class SVM), and data visualization techniques.
5. Handling: Options may include removing or correcting outliers, transforming data, or using robust methods.
6. Considerations: Carefully considering the impact of global outliers is crucial for accurate data analysis and machine learning model outcomes.

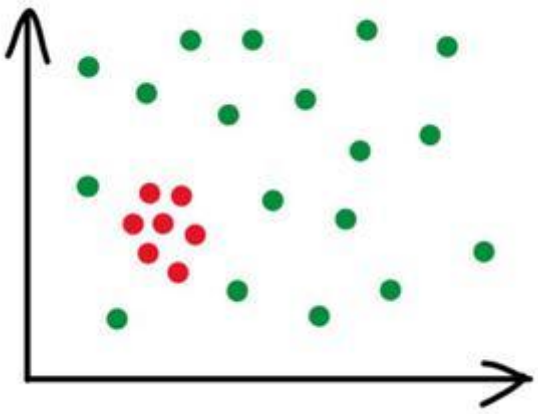


The red data point is a global outlier.

2. Collective Outliers

1. Definition: Collective outliers are groups of data points that collectively deviate significantly from the overall distribution of a dataset.
2. Characteristics: Collective outliers may not be outliers when considered individually, but as a group, they exhibit unusual behavior.
3. Detection: Techniques for detecting collective outliers include clustering algorithms, density-based methods, and subspace-based approaches.
4 Impact: Collective outliers can represent interesting patterns or anomalies in data that may require special attention or further investigation.
5. Handling: Handling collective outliers depends on the specific use case and may involve further analysis of the group behavior, identification of contributing factors, or considering contextual information.
6. Considerations: Detecting and interpreting collective outliers can be more complex than

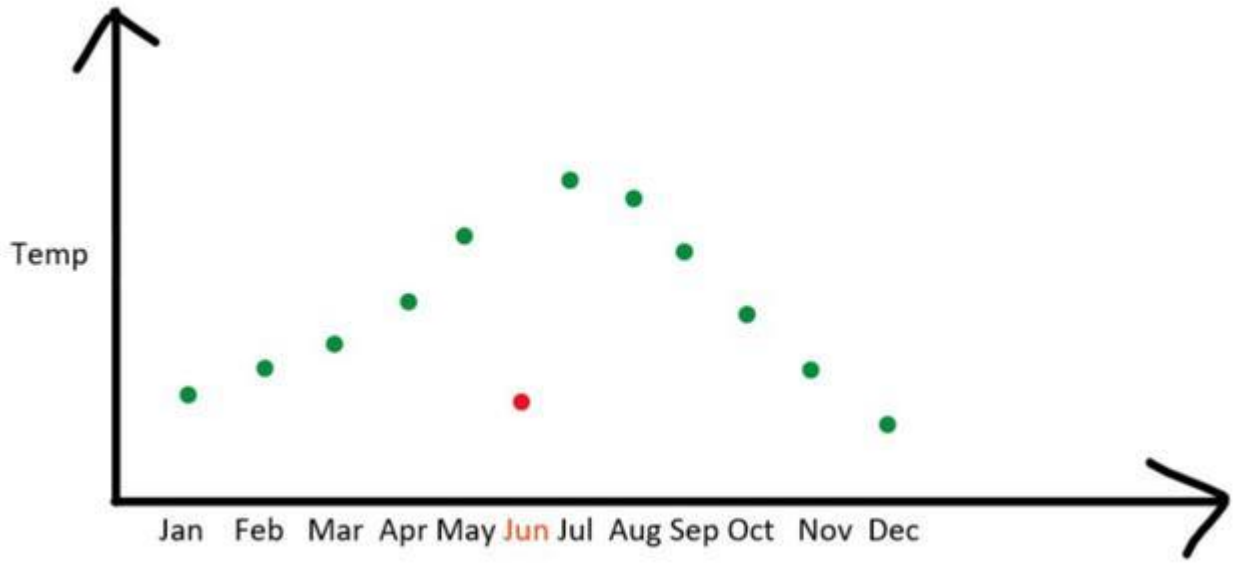
individual outliers, as the focus is on group behavior rather than individual data points. Proper understanding of the data context and domain knowledge is crucial for effective handling of collective outliers.



The red data points as a whole are collective outliers.

3. Contextual Outliers

1. Definition: Contextual outliers are data points that deviate significantly from the expected behavior within a specific context or subgroup.
2. Characteristics: Contextual outliers may not be outliers when considered in the entire dataset, but they exhibit unusual behavior within a specific context or subgroup.
3. Detection: Techniques for detecting contextual outliers include contextual clustering, contextual anomaly detection, and context-aware machine learning approaches.
4. Contextual Information: Contextual information such as time, location, or other relevant factors are crucial in identifying contextual outliers.
5. Impact: Contextual outliers can represent unusual or anomalous behavior within a specific context, which may require further investigation or attention.
6. Handling: Handling contextual outliers may involve considering the contextual information, contextual normalization or transformation of data, or using context-specific models or algorithms.
7. Considerations: Proper understanding of the context and domain-specific knowledge is crucial for accurate detection and interpretation of contextual outliers, as they may vary based on the specific context or subgroup being considered.



A low temperature value in June is a contextual outlier because the same value in December is not an outlier.

Data Mining Applications

A list of data mining applications across various domains:

1. **Marketing and Customer Relationship Management (CRM):**
   - Customer segmentation for targeted marketing.
   - Market basket analysis for product recommendations.
   - Churn prediction to retain customers.
   - Campaign response prediction.
2. **Finance:**
   - Credit scoring and risk assessment.
   - Fraud detection in financial transactions.
   - Stock market prediction.
   - Portfolio management and optimization.
3. **Healthcare:**
   - Disease diagnosis and prediction.
   - Drug discovery and development.
   - Healthcare fraud detection.
   - Patient outcome analysis.
4. **Retail and E-commerce:**
   - Inventory management and demand forecasting.
   - Price optimization.
   - Recommender systems for product recommendations.
   - Customer sentiment analysis.
5. **Manufacturing and Supply Chain:**
   - Quality control and defect detection.
   - Demand forecasting for production planning.
   - Supplier evaluation and selection.
   - Supply chain optimization.

6. **Social Media and Web Analysis:**
    - Sentiment analysis for brand monitoring.
    - Clickstream analysis for website optimization.
    - Content recommendation.
    - Social network analysis.
7. **Environmental Science:**
    - Climate modeling and prediction.
    - Species habitat modeling for conservation.
    - Air and water quality monitoring.
    - Natural disaster prediction.
8. **Government and Public Policy:**
    - Crime pattern analysis for law enforcement.
    - Traffic analysis for urban planning.
    - Healthcare data analysis for public health policies.
    - Education data for educational planning and resource allocation.
9. **Education:**
    - Student performance analysis and early intervention.
    - Adaptive learning and personalized education.
    - Course recommendation.
    - Educational data mining for research.
10. **Telecommunications:**
    - Network performance optimization.
    - Customer churn prediction.
    - Fraud detection in telecom services.
    - Network security analysis.
11. **Sports Analytics:**
    - Player performance analysis.
    - Injury prediction and prevention.
    - Fan engagement and marketing.
    - Game strategy optimization.
12. **Human Resources:**
    - Employee attrition prediction.
    - Talent acquisition and recruitment.
    - Employee performance evaluation.
    - Workforce planning and optimization.
13. **Energy Management:**
    - Energy consumption analysis and optimization.
    - Predictive maintenance for energy equipment.
    - Renewable energy forecasting.
    - Demand response management.
14. **Agriculture:**
    - Crop yield prediction.
    - Pest and disease monitoring.
    - Precision agriculture for resource optimization.
    - Soil quality analysis.
15. **Transportation and Logistics:**
    - Route optimization and fleet management.
    - Demand forecasting for transportation services.
    - Traffic congestion prediction.
    - Vehicle health monitoring.