AKTU B.Tech 3rd Year

CS & CS Allied

Data Analytics & Visualization

Unit-4

in ONE SHOT



AKTU PYQs Covered

By Shubhangi Ma'am

Gateway Classes

GATEWAY CLASSES

GATEWAY CLASSES

SYLLABUS

UNIT - IV

Frequent Itemsets and Clustering: Mining frequent itemsets, market based modelling, Apriori algorithm, handling large data sets in main memory, limited pass algorithm, counting frequent itemsets in a stream, clustering techniques: hierarchical, K-means, clustering high dimensional data, CLIQUE and ProCLUS, frequent pattern based clustering methods, clustering in non-Euclidean space, clustering for streams and parallelism.

Gateway Classes

PREVIOUS YEAR IMPORTANT QUESTIONS

Q.1 : What is frequent itemset mining ? **AKTU(2025-26)**

Q.2 : Explain Apriori algorithm and its role in frequent itemset mining . **AKTU(2025-26)**

Q.3 : Explain clustering techniques used in data mining. **AKTU(2025-26)**

Q.4 : Analyze the working of K – means clustering with an example . **AKTU(2025-26)**

Q.5 : How do methods like CLIQUE & ProCLUS overcome traditional clustering limitations. **AKTU(2024-25)**

Q.6 : Compare & Contrast hierarchical clustering & K – means clustering. What are the advantages and disadvantages of each ? **AKTU(2024-25)**

Q.7 : What is parallelism in clustering algorithm? How does it improve performance in handling large datasets ? **AKTU(2024-25)**

Q.8 : How is clustering applied to streaming data ? Discuss the modifications needed in clustering algorithms for stream analytics. **AKTU(2024-25)**

Gateway Classes

GATEWAY CLASSES

GATEWAY CLASSES

Q.9 : List various methods of clustering. **AKTU(2024-24)**

Q.10 : Illustrate with examples CLIQUE & ProCLUS . **AKTU(2023-24)**

Q.11 : explain the following : **AKTU(2023-24)**

i. Apriori Akgorithm

ii. Market based modeling

Q.12 : Discuss k – means clustering ? Explain its working. Write k – means algorithm for partitioning . **AKTU(2023-24)**

Q.13 : Explain hierarchical method of clustering. **AKTU(2022-23)**

Q.14 : Explain market based analysis with an example . **AKTU(2022-23)**

Q.15 : What do you mean by k – means algorithm? How does the k means algorithm work? Write k means algorithm for partitioning ? **AKTU(2022-23)**

Gateway Classes

GW GATEWAY CLASSES

Q.16 : Explain the importance of frequent itemset in market based analysis. **AKTU(2025-26)**

Q.17 : Write two limitations of K – mean. **AKTU(2024-25)**

Q.18 : How do stream – based algorithms differ from batch processing algorithm in the context of frequent itemset mining. **AKTU(2023-24)**

Q.19 : What are the challenges associated with implementing Apriori in scenarios with limited available memory ? **AKTU(2023-24)**

Q.20 : Compare the strengths and weakness of hierarchical clustering & k – means clustering. Under what circumstances would one techniques be preferred over the other , & why ? **AKTU(2023-24)**

Q.21 : Provide an in- depth Comparison between the CLIQUE & ProCLUS clustering algorithms. How do these methods handle challenges such as noise, outliers and varying cluster shapes? **AKTU(2022-23)**

Gateway Classes

ITEM

An item is a single object or product present in a transaction.

Example :

If a customer buys Milk, Bread and Eggs, then :

Milk → Item

Bread → Item

Eggs → Item

ITEMSET

An itemset is a collection of one or more items.

Example :

{Milk} → 1-itemset

{Milk, Bread} → 2-itemset

{Milk, Bread, Eggs} → 3-itemset

Note: The order of items does not matter.

{Milk, Bread} = {Bread, Milk}

Gateway Classes

TRANSACTION

- A transaction is a single record that contains the items purchased or present together.
- Each transaction is usually given a unique Transaction ID (TID).

Example :

TID	ITEMS
T1	Milk, Bread
T2	Bread, Eggs
T3	Milk, Bread, Eggs

Here, T1, T2 and T3 are transactions.

Gateway Classes

SUPPORT

Support tells us how frequently an itemset occurs in the complete set of transactions.

$$Support(X) = \frac{\text{Number of transactions containing X}}{\text{Total number of transactions}}$$

Example : Suppose there are **5 transactions**, and {Milk, Bread} appears in **3 transactions**.

$$Support(\{\text{Milk, Bread}\}) = \frac{3}{5} = 0.6 = 60\%$$

So, the support of {Milk, Bread} is **60%**.

MINIMUM SUPPORT

- Minimum Support (Min-Support) is the minimum support value required for an itemset to be considered frequent.
- It is selected as a threshold before mining frequent itemsets.

Example:

If minimum support = **50%**:

- Itemset with support **60%** → Frequent
- Itemset with support **40%** → Infrequent

Gateway Classes

FREQUENT ITEMSET

An itemset whose support is greater than or equal to the minimum support is called a Frequent Itemset.

$$Support(X) \geq Minimum\ Support$$

Example: Minimum support = 50%

If {Milk, Bread} has support = 60%, then:

{Milk, Bread} → Frequent Itemset

INFREQUENT ITEMSET

An itemset whose support is less than the minimum support is called an Infrequent Itemset.

$$Support(X) < Minimum\ Support$$

Example: Minimum support = 50%

If {Milk, Eggs} has support = 30%, then:

{Milk, Eggs} → Infrequent Itemset

GATEWAY CLASSES

MINING FREQUENT ITEMSETS

GATEWAY CLASSES

FREQUENT ITEMSET MINING

Frequent Itemset Mining is the process of finding groups of items that occur together frequently in a transaction dataset.

It uses the support of an itemset to decide whether it is frequent or not.

Example : Suppose a store has these transactions:

TRANSACTION	ITEMS
T1	Milk, Bread
T2	Milk, Bread, Eggs
T3	Bread, Eggs
T4	Milk, Bread

Here, {Milk, Bread} appears in 3 out of 4 transactions.

So,

Support = 3/4 = 75%

If the minimum support is 50%, then {Milk, Bread} is a frequent itemset.

Gateway Classes

GATEWAY CLASSES

MAIN STEPS

1. Collect the transaction data.
2. Find different itemsets in the data.
3. Calculate their support.
4. Compare support with minimum support.
5. Keep the itemsets that satisfy the minimum support.

These frequent itemsets can then be used to discover useful relationships between items.

APPLICATIONS

1. Market Basket Analysis : Find products that customers often buy together.

2. Product Recommendation : Suggest related products to customers based on items they have already selected.

3. Store Arrangement : Products that are frequently bought together can be placed near each other.

4. Sales and Promotion : Frequently related products can be offered together in combo offers or discounts.

5. Customer Behaviour Analysis : It helps businesses understand buying patterns and preferences.

Gateway Classes

MARKET-BASED MODELLING

Market-Based Modelling is the use of transaction data to understand which products are commonly purchased together.

It is mainly used in market basket analysis.

Example : Suppose transaction data shows:

- Customers frequently buy Laptop + Laptop Bag
- Customers frequently buy Bread + Butter
- Customers frequently buy Phone + Earphones

The business can use these patterns to:

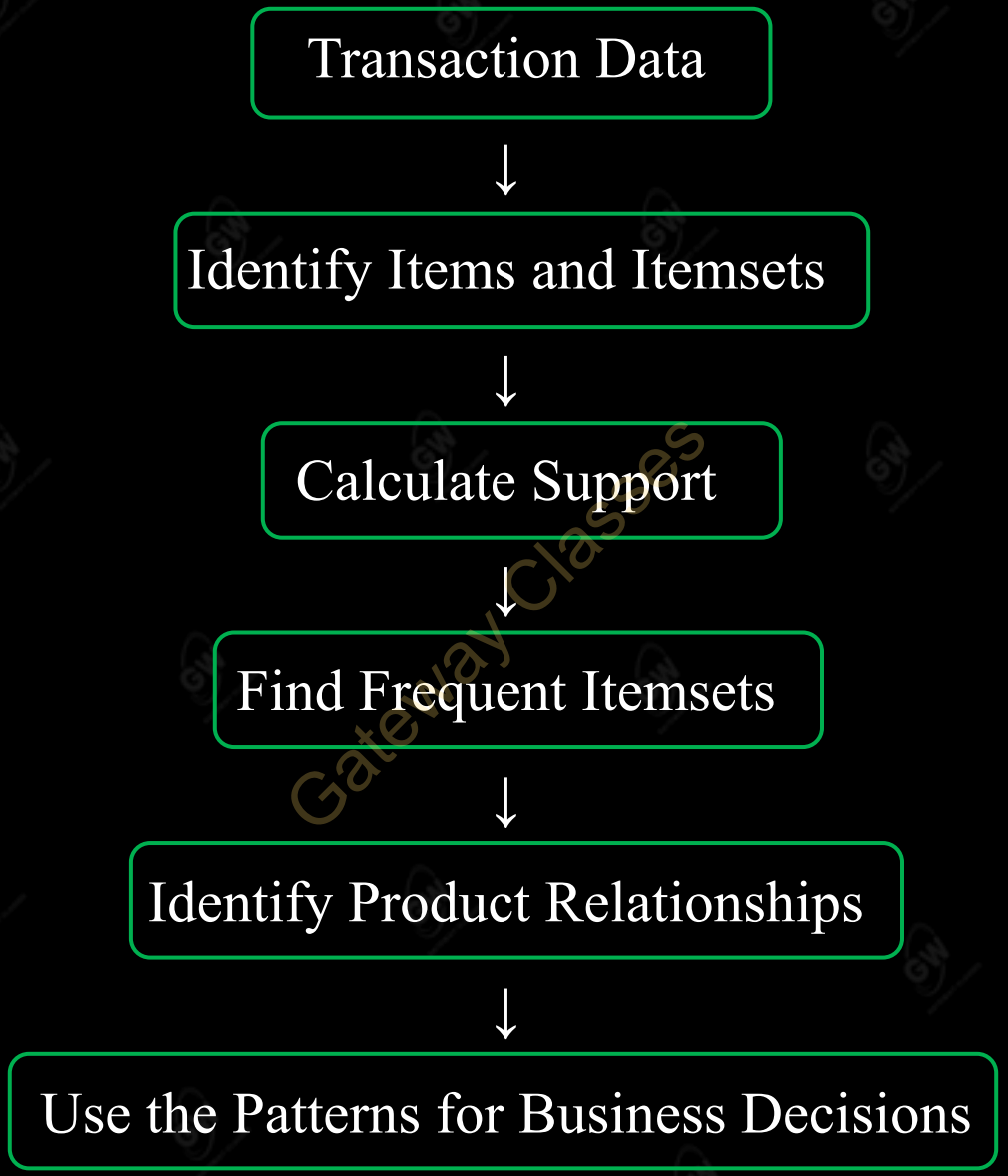
- recommend related products,
- create product bundles,
- plan promotions,
- arrange products effectively.

Advantages

- Improves product recommendations.
- Supports better store/product arrangement.
- Helps in planning promotions and combo offers.
- Helps businesses understand customer behaviour.

Gateway Classes

HOW MARKET-BASED MODELLING WORKS



GATEWAY CLASSES
Gateway Classes

1. COLLECT TRANSACTION DATA

The transaction records of customers are collected.

Example:

TRANSACTION	ITEMS
T1	Milk, Bread
T2	Milk, Bread, Eggs
T3	Bread, Eggs
T4	Milk, Bread

This data represents the products purchased by different customers.

Gateway Classes

2. FIND FREQUENTLY PURCHASED ITEMS

The data is analyzed to find which **items or groups of items occur frequently**.

For example, if Bread and Butter occur together in many transactions, **{Bread, Butter}** is a frequent itemset.

3. MEASURE THE FREQUENCY

The frequency of an itemset is measured using **support**.

$$Support(X) = \frac{\text{Number of transactions containing X}}{\text{Total number of transactions}}$$

A **minimum support** value is used to select important frequent itemsets.

4. DISCOVER PRODUCT RELATIONSHIPS

After finding frequent itemsets, relationships between products can be studied.

For example :

Bread → Butter

This means customers who purchase Bread may also purchase Butter.

Gateway Classes

APPLICATIONS OF MARKET-BASED MODELLING

1. Product Recommendation : Businesses can recommend related products to customers.

Example : If a customer buys a camera, the system may recommend a memory card.

2. Product Placement : Frequently purchased products can be placed close to each other in a store.

3. Combo Offers : Products that are commonly bought together can be offered as a **bundle or combo**.

4. Sales and Promotions : Businesses can create targeted discounts and promotional offers based on buying patterns.

5. Customer Behaviour Analysis : It helps businesses understand **customer preferences and purchasing habits**

Gateway Classes

GATEWAY CLASSES

GATEWAY CLASSES

Importance of Frequent Itemsets in Market-Based Analysis

- Frequent itemsets help businesses identify products that are frequently purchased together.
- They reveal common customer buying patterns.
- They help in product recommendations.
- They support product placement in stores.
- They help create combo offers and promotions.
- They help businesses understand customer purchasing behaviour.
- They can support better sales and marketing decisions.

Gateway Classes

Example :

If Bread and Butter frequently occur together, a store can place them closer or recommend one when the other is purchased.

APRIORI ALGORITHM

The **Apriori Algorithm** is a popular algorithm used for **mining frequent itemsets** from a transaction database.

It finds frequent itemsets by:

1. Finding frequent 1-itemsets.
2. Generating larger candidate itemsets.
3. Removing candidates that cannot be frequent.
4. Counting their support.
5. Repeating the process until no more frequent itemsets can be found.

Gateway Classes

APRIORI PRINCIPLE / PROPERTY

The Apriori Principle states : If an itemset is frequent, then all of its non-empty subsets must also be frequent.

Similarly , If an itemset is infrequent, all of its supersets will also be infrequent.

Example : Suppose : {Milk, Bread, Eggs} is frequent. Then its subsets must also be frequent:

- {Milk, Bread}
- {Milk, Eggs}
- {Bread, Eggs}
- {Milk}
- {Bread}
- {Eggs}

Why is it useful?

It helps to **reduce the number of candidates** that need to be checked.

Instead of checking every possible itemset, Apriori removes impossible candidates early.

If {Milk, Bread} is infrequent, then : {Milk, Bread, Eggs} cannot be frequent.

Gateway Classes

CANDIDATE GENERATION

Candidate Generation means creating possible larger itemsets from the frequent itemsets of the previous level. These possible itemsets are called candidate itemsets.

Example : Suppose frequent 1-itemsets are:

L_1 = {A}, {B}, {C}

We generate candidate 2-itemsets : **C_2 = {A,B}, {A,C}, {B,C}**

Now suppose :

L_2 = {A,B}, {A,C}

To generate 3-itemsets, we try to combine them:

C_3 = {A,B,C}

So , **Frequent k-itemsets → Candidate (k+1)-itemsets**

Gateway Classes

CANDIDATE PRUNING

After generating candidates, Apriori removes the candidates that cannot be frequent. This process is called Candidate Pruning.

Pruning Rule : If any (k−1)-subset of a candidate k-itemset is infrequent, the candidate is removed.

Example : Suppose : **C_3 = {A, B, C}** Its 2-item subsets are :

- {A,B}
- {A,C}
- {B,C}

Suppose :

- {A,B} → Frequent
- {A,C} → Frequent
- {B,C} → **Infrequent**

⟶ Then : **{A,B,C} is pruned ,**because one of its subsets, **{B,C}**, is infrequent.

Gateway Classes

FREQUENT ITEMSETS IN LARGE DATA

When the transaction database is very large, finding frequent itemsets becomes difficult because the **complete dataset may not fit into main memory** and repeated scanning can take more time.

To handle this problem, special techniques are used :

1. Handling Large Data Sets in Main Memory
2. Limited-Pass Algorithms
3. Counting Frequent Itemsets in a Stream

Gateway Classes

1. HANDLING LARGE DATA SETS IN MAIN MEMORY

Main Memory (RAM) has limited space. If the transaction database is very large, it may not be possible to store the entire database in RAM.

Therefore, frequent itemset mining algorithms try to **use memory efficiently and reduce unnecessary data processing**.

Main Problems

- Large datasets may not fit completely in main memory.
- Storing all candidate itemsets requires a lot of memory.
- Repeatedly scanning a large database increases execution time.
- The number of possible itemsets can become very large.

Gateway Classes

GATEWAY CLASSES

METHODS TO HANDLE LARGE DATA

1. Store Only Required Data : Only the information needed for finding frequent itemsets is kept in memory. This reduces memory usage.

2. Reduce Candidate Itemsets : The **Apriori property** is used to remove itemsets that cannot be frequent. This prevents unnecessary candidates from being stored and counted.

3. Process Data in Parts : A large database can be divided into smaller parts and processed one part at a time.

4. Use Efficient Data Structures : Structures such as **hash trees** can be used to store and count candidate itemsets efficiently.

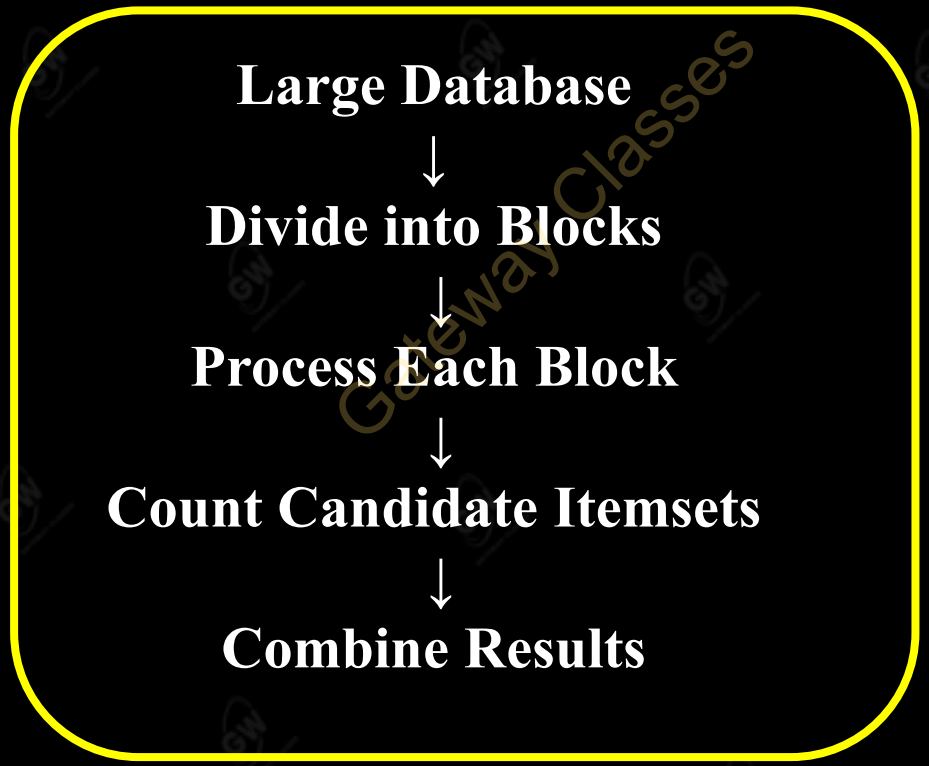
Gateway Classes

GATEWAY CLASSES

GATEWAY CLASSES

Example : Suppose a database contains, 1,000,000 **transactions** and only **100,000 transactions** can be processed in memory at one time.

Instead of trying to load everything together, the data can be processed in **blocks/chunks**.



Gateway Classes

2. LIMITED-PASS ALGORITHMS

- A **Limited-Pass Algorithm** is an algorithm that requires only a **small number of scans (passes) over the database**.
- A **pass** means reading the transaction database from beginning to end once.

Why Limited Passes are Needed?

For a very large database, scanning the complete database many times is expensive.

For example

- 1 pass → database is read once
- 2 passes → database is read twice
- 5 passes → database is read five times

Therefore, reducing the number of passes can **save time and I/O cost**.

Gateway Classes

GATEWAY CLASSES

Two-Pass Approach : A common idea is to find frequent itemsets using **two database scans**.

Pass 1

- Read the database.
- Find information required to identify candidate/frequent itemsets.
- Generate a manageable set of candidates.

Pass 2

- Scan the database again.
- Count the occurrences of the candidates.
- Calculate their support.
- Keep the itemsets satisfying minimum support.

Gateway Classes

3. COUNTING FREQUENT ITEMSETS IN A STREAM

A **data stream** is a continuous flow of data that arrives over time.

Examples :

- Online shopping transactions
- Credit card transactions
- Sensor data
- Website clicks
- Social media activity

Unlike a static database, stream data may be **continuous and very large**, so storing everything is often not practical.

Therefore, frequent itemsets must be counted **while the data is arriving**.

Incoming Transactions

↓

Read Stream Data

↓

Maintain Counts of Itemsets

↓

Update Counts as New Data Arrives

↓

Compare with Minimum Support

↓

Identify Frequent Itemsets

Gateway Classes

CHALLENGES IN STREAM DATA

1. Continuous Data : New transactions keep arriving continuously.

2. Large Volume : The amount of data can become extremely large.

3. Limited Memory : It is usually not possible to store the complete stream.

4. Changing Patterns : Customer behaviour and data patterns may change over time.

For example, during winter, **jackets** may become more frequently purchased than **T-shirts**.

GATEWAY CLASSES
Gateway Classes

CLUSTERING

Clustering is a data mining technique used to group similar data objects into the same group, called a cluster. Objects in the same cluster are more similar to each other, while objects in different clusters are less similar.

Example : Suppose we have customer data based on their shopping habits

- Customers who buy similar products → **Cluster 1**
- Customers with different shopping habits → **Cluster 2**
- Another group with similar habits → **Cluster 3**

Data Objects → Measure Similarity → Form Groups → Similar Objects in Same Cluster

Applications of Clustering

- Customer segmentation
- Image classification
- Market analysis
- Pattern recognition
- Social network analysis

GatewayClasses

TYPES / TECHNIQUES OF CLUSTERING

1. Hierarchical Clustering : It creates a **hierarchy of clusters**.

It can work in two ways :

- **Agglomerative :** Starts with individual objects and combines them into larger clusters.
- **Divisive :** Starts with one large cluster and divides it into smaller clusters.

Output : Usually represented using a **dendrogram**.

2. Partitioning Clustering : It divides the data into a **fixed number of clusters**. The most common example is **K-Means Clustering**.

Example : If K = 3, the algorithm divides the data into **3 clusters**.

3. Density-Based Clustering : It forms clusters based on **dense regions of data points**.

Dense areas become clusters, while points in sparse areas may be treated as noise or outliers.

Gateway Classes

GATEWAY CLASSES
GATEWAY CLASSES

4. Grid-Based Clustering : It divides the data space into a **finite number of grid cells** and performs clustering using these cells.

It can be useful for **large datasets** because it works with grid cells instead of processing every data point individually.

5. Model-Based Clustering : It assumes that the data is generated according to a **mathematical or statistical model**.

The algorithm tries to find the model that best represents the data.

Gateway Classes

HIERARCHICAL CLUSTERING

Hierarchical Clustering is a clustering technique that groups similar data objects into a **hierarchy of clusters**.

It creates clusters step by step based on the **similarity or distance** between data objects.

The result is represented using a **dendrogram**, which shows how clusters are formed or divided.

Types of Hierarchical Clustering : There are two main approaches --

1. Agglomerative Hierarchical Clustering

It follows a **bottom-up approach**.

- Initially, each data object is treated as a separate cluster.
- The closest clusters are merged.
- This process continues until all objects form one cluster or the required number of clusters is obtained.

Gateway Classes

GATEWAY CLASSES

Example : Initially : A B C D After merging : {A,B} {C} {D}

Then : {A,B} {C,D} Finally : {A,B,C,D}

2. Divisive Hierarchical Clustering : It follows a **top-down approach**.

- Initially, all data objects are placed in one cluster.
- The cluster is divided into smaller clusters.
- The process continues until the required clusters are obtained.

Example : Initially : {A,B,C,D}

After division : {A,B} {C,D} Further division : {A} {B} {C} {D}

Gateway Classes

Dendrogram

A **dendrogram** is a tree-like diagram that represents the hierarchy of clusters.

It shows :

- Which objects are grouped together.
- The order in which clusters are merged or divided.
- The similarity or distance at which clusters are joined.

WORKING OF AGGLOMERATIVE HIERARCHICAL CLUSTERING

1. Start with each data object as a separate cluster.
2. Calculate the distance between clusters.
3. Find the two most similar/closest clusters.
4. Merge these clusters.
5. Recalculate the distances between the new cluster and other clusters.
6. Repeat the process until the required number of clusters is obtained.

Example : For four objects **A, B, C, D**:

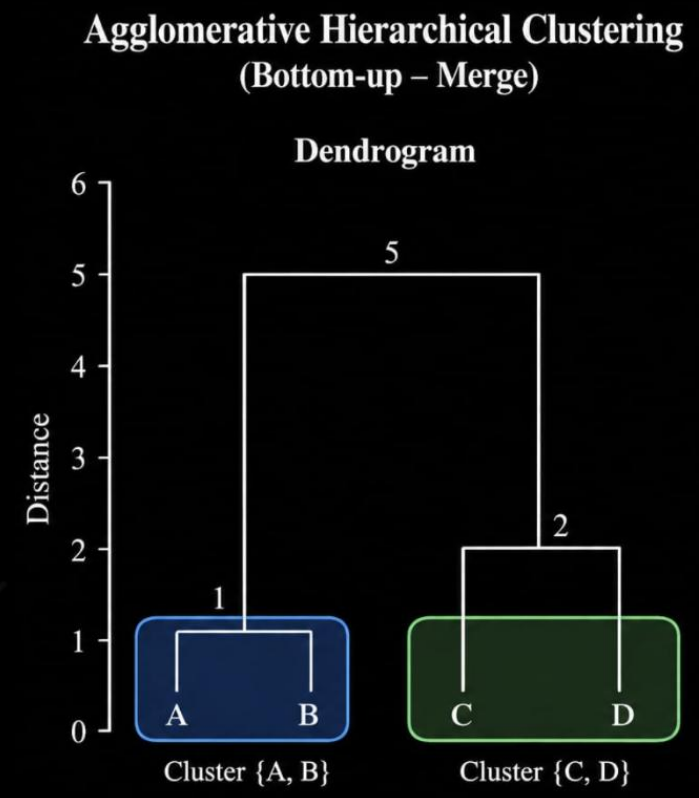
Step 1 : {A}, {B}, {C}, {D}

Step 2 : If A and B are closest → {A,B}, {C}, {D}

Step 3 : If C and D are closest → {A,B}, {C,D}

Step 4 : Merge both → {A,B,C,D}

The complete process is represented through a dendrogram.



Gateway Classes

WORKING OF DIVISIVE HIEARCHICAL CLUSTERING

1. Start with one cluster containing all data points.
2. Select the cluster that needs to be divided.
3. Find the points that are most dissimilar or farthest from each other.
4. Split the selected cluster into smaller clusters.
5. Repeat the splitting process for the resulting clusters.
6. Stop when the required number of clusters is obtained.

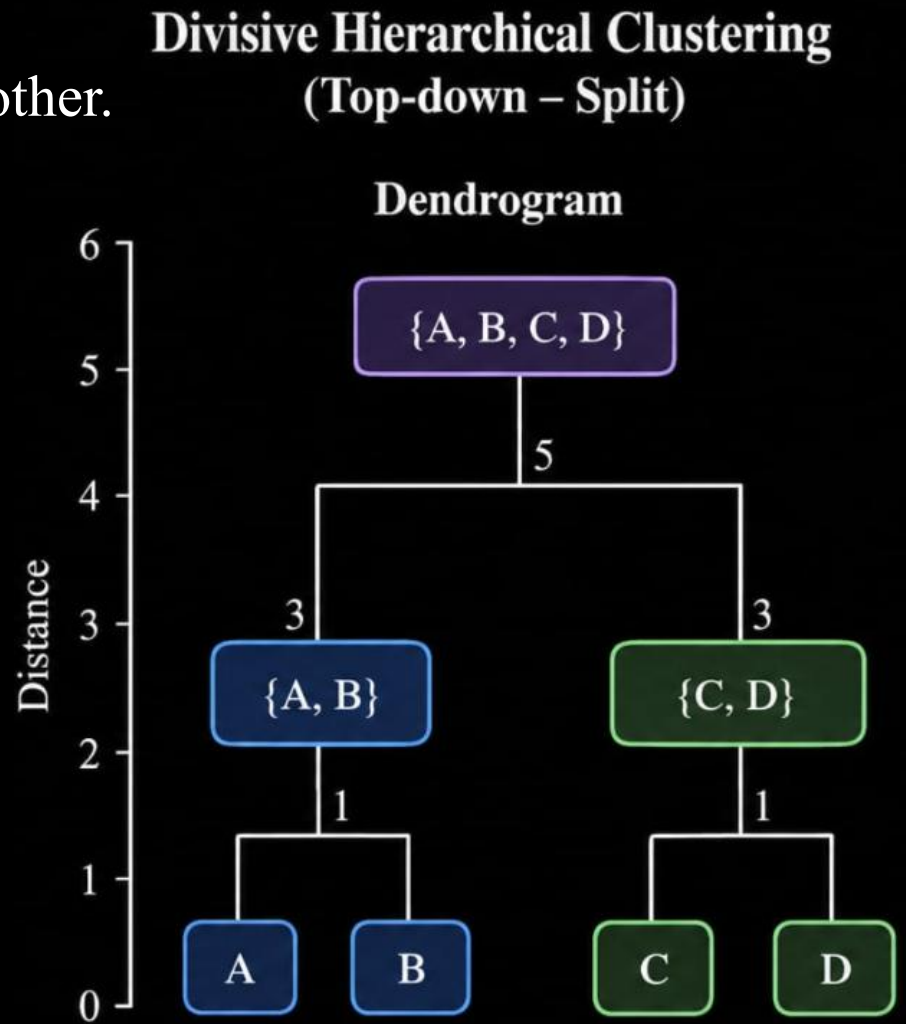
Example : Suppose the data points are A, B, C, D:

Step 1 : {A, B, C, D}

Step 2: Split into two groups : {A, B} {C, D}

Step 3: Further split if required : {A} {B} {C} {D}

Gateway Classes



GATEWAY CLASSES

Advantages of hierarchical clustering

- Easy to understand and visualize.
- Produces a hierarchy of clusters.
- A dendrogram gives information at different levels of clustering.
- The number of clusters does not have to be fixed at the beginning.

Limitations of hierarchical clustering

- Can be slow for large datasets.
- Once two clusters are merged, the decision is difficult to reverse.
- Results depend on the distance measure and clustering method used.

Gateway Classes

K-MEANS CLUSTERING

K-Means Clustering is a **partitioning clustering technique** that divides a dataset into **K clusters**.

Each cluster is represented by a **centroid**, which is the mean of the data points belonging to that cluster.

The main aim is to place similar data points in the same cluster and different data points in different clusters.

IMPORTANT TERMS

K : **K** represents the number of clusters to be formed.

For example, if : $K = 3$, then the data will be divided into **3 clusters**.

Centroid : A **centroid** is the mean position of all data points in a cluster.

For 2-dimensional data : Centroid=($\sum x/ n$, $\sum y / n$)

Distance : Distance is used to find how close a data point is to a centroid.

The commonly used distance is **Euclidean Distance** : $d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$

Gateway Classes

K-MEANS ALGORITHM FOR PARTITIONING

Step 1 : Select K : Choose the number of clusters **K**.

Step 2 : Select Initial Centroids : Select **K initial centroids** from the data.

Step 3 : Calculate Distance : Calculate the distance of every data point from each centroid.

Step 4 : Assign Data Points : Assign each data point to the cluster whose centroid is **nearest**.

Step 5 : Calculate New Centroids : For each cluster, calculate the mean of its data points.

Step 6 : Repeat : Again calculate distances and reassign the points using the new centroids.

Step 7 : Stop : Repeat until the cluster assignments or centroids **do not change significantly**.

Gateway Classes

GATEWAY CLASSES

Advantages of K-Means

- Simple and easy to understand.
- Easy to implement.
- Fast for large datasets.
- Works well when clusters are clearly separated.
- Gives a clear partition of the data.

Limitations of K-Means

- The value of **K must be decided in advance**.
- Results may depend on the initial centroids.
- Sensitive to outliers.
- Not suitable for clusters with very irregular shapes.
- Different initial centroids may produce different results.

Gateway Classes

DIFFERENCE BETWEEN HIEARCHICAL & K – MEANS CLUSTERING

BASIS	HIERARCHICAL	K – MEANS
TYPE	Hierarchical	Partitioning
BASIC IDEA	Builds a hierarchy of clusters	Divides data into k Clusters
NUMBER OF CLUSTERS	Can be selected later	Must be specified
MAIN CONCEPT	Merging or splitting	Centroid & distance
OUTPUT	Dendrogram	Final K clusters
CENTROID	Not required	Required
LARGE DATASETS	Generally more expensive	Generally Faster

GATEWAY CLASSES

HIGH-DIMENSIONAL DATA

High-dimensional data is data that contains a **large number of attributes or dimensions** for each data object.

A dimension represents a feature or attribute of the data.

Example : In a customer dataset, features may include :

Age , Income , Location ,Spending ,Education , Purchase frequency ,Number of products purchased

If a dataset contains **hundreds or thousands of features**, it is called **high-dimensional data**.

Examples : Text documents, gene-expression data, image data and customer datasets can contain many dimensions

Gateway Classes

PROBLEMS / CHALLENGES

1. Curse of Dimensionality : As the number of dimensions increases, the distance between data points becomes less meaningful. It becomes difficult to clearly identify which points are **similar or different**.

2. Irrelevant Features : Not every dimension is useful for clustering. Some features may contain irrelevant information and can reduce the quality of clusters.

3. High Computational Cost : A large number of dimensions requires more **time and memory** for distance calculation and clustering.

4. Sparse Data : In high-dimensional datasets, data points may be spread across a large space, making the data **sparse**.

5. Difficulty in Visualization : It is easy to visualize data in 2 or 3 dimensions, but high-dimensional data cannot be directly visualized easily.

6. Poor Clustering Quality : Traditional clustering methods may fail to find meaningful groups because of irrelevant dimensions and unreliable distance measurements.

Gateway Classes

NEED FOR SPECIALIZED CLUSTERING METHODS

Traditional clustering methods such as **K-Means** may not work effectively on high-dimensional data.

Therefore, specialized methods are needed to :

- Select **important dimensions or features**.
- Reduce the effect of irrelevant attributes.
- Find meaningful clusters in a large number of dimensions.
- Handle sparse high-dimensional data efficiently.
- Reduce computational cost.
- Improve the quality and accuracy of clusters.

Examples of Specialized Methods : Some methods designed for high-dimensional data are:

(1) CLIQUE (2) ProCLUS

Gateway Classes

CLIQUE

CLIQUE (Clustering In QUEst) is a **grid-based clustering algorithm** designed for **high-dimensional data**.

It finds clusters by identifying **dense regions** in different subsets of dimensions.

The main idea is that a cluster may exist only in **some dimensions**, rather than in all dimensions.

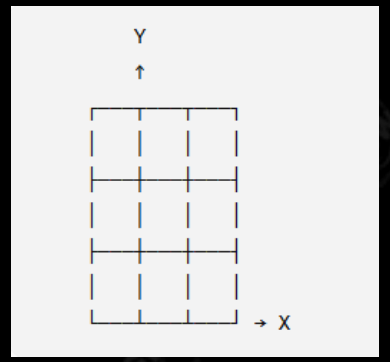
Key Points

- It is designed for **high-dimensional datasets**.
- It divides the data space into **small rectangular units called units/cells**.
- It identifies **dense units** containing enough data points.
- Connected dense units are combined to form clusters.
- It can find clusters having **different shapes and dimensions**.

Gateway Classes

WORKING OF CLIQUE

Step 1 : Divide the Data Space : Each dimension is divided into a number of **intervals**. This creates a grid of small units.



Example : If two dimensions X and Y are divided into 3 intervals each:

Step 2 : Find Dense Units : The number of data points in each unit is counted.

A unit is called **dense** if the number of points in it satisfies the given **density threshold**.

Units that do not satisfy the threshold are treated as **non-dense**.

Step 3 : Identify Relevant Dimensions : CLIQUE searches for dimensions or combinations of dimensions where **dense regions exist**.

This is important because in high-dimensional data, meaningful clusters may appear only in a subset of dimensions.

Gateway Classes

GATEWAY CLASSES

GATEWAY CLASSES

Step 4 : Find Connected Dense Units : Neighbouring dense units are connected.

Connected dense units are considered part of the **same cluster**.

Step 5 : Generate Clusters : The connected dense regions are combined to form the final clusters.

Thus, CLIQUE identifies clusters in different **subspaces** of the high-dimensional data.

NOISE & OUTLIERS :

Density-based processing can leave low-density regions outside the identified dense clusters, so isolated points may not belong to dense clusters.

VARYING CLUSTER SHAPES :

Can identify clusters of different shapes when they appear as connected dense grid regions.

Gateway Classes

CLUSTERING IN HIGH-DIMENSIONAL SPACE

CLIQUE (Clustering In QUEst) is a **grid-based clustering algorithm** designed for **high-dimensional data**.

In high-dimensional data, using **all dimensions together** may not produce meaningful clusters.

CLIQUE solves this by performing **subspace clustering**.

Subspace Clustering : A **subspace** is a subset of the dimensions of the complete dataset.

For example, suppose data has : **X_1, X_2, X_3, X_4**

A cluster may be meaningful only in : **X_1 and X_3**

CLIQUE can identify this cluster without requiring all four dimensions.

Example : Suppose a dataset has dimensions : **Age, Income, Spending, Location**

A group of customers may show a clear pattern based on : **Income + Spending**

but not based on Location.

CLIQUE can identify the dense region in the **Income–Spending subspace**.

Gateway Classes

GATEWAY CLASSES

Advantages of CLIQUE

- Designed specifically for **high-dimensional data**.
- Finds clusters in **subspaces**.
- Can discover clusters of different shapes.
- Reduces the effect of irrelevant dimensions.
- Uses grid-based processing, making dense regions easier to identify.

Limitations of CLIQUE

- The result can depend on the choice of **grid size and density threshold**.
- A large number of dimensions can produce many possible subspaces.
- Selecting suitable parameters may be difficult.

Gateway Classes

ProCLUS

ProCLUS (PROjective CLUStering) is a **projective clustering algorithm** designed for **high-dimensional data**.

It finds clusters by selecting **different relevant dimensions for different clusters**.

The main idea is that a cluster may be meaningful only in a **subset of dimensions**, while the other dimensions may be irrelevant for that cluster.

Example : Suppose customer data has : **Age, Income, Spending, Location, Education**

For one cluster, the important dimensions may be : **Income + Spending**

while for another cluster, the important dimensions may be : **Age + Education**

ProCLUS can identify these clusters using their **relevant dimensions**.

Gateway Classes

WORKING OF ProCLUS

Step 1: Select Initial Medoids : A small set of representative data points called **medoids** is selected.

A medoid is an actual data point that represents a cluster.

Step 2: Select Relevant Dimensions : For each cluster, ProCLUS identifies the dimensions in which the data points are most closely related. Thus, **different clusters can use different dimensions**.

Step 3: Assign Data Points : Each data point is assigned to the **nearest medoid** using only the relevant dimensions of that cluster.

Step 4: Calculate Cluster Quality : The quality of the clusters is evaluated using the distance of points from their corresponding medoids.

Step 5: Improve the Clusters : The algorithm tries different medoids and dimensions and selects a combination that produces better clusters.

Step 6 : Final Clusters : The process continues until a good set of **clusters and relevant dimensions** is obtained.

Gateway Classes

GATEWAY CLASSES

NOISE & OUTLIERS :

Uses medoids and relevant dimensions, which can reduce the influence of irrelevant dimensions; however, outliers can still affect clustering and need suitable handling.

VARYING CLUSTER SHAPES :

Focuses more on clusters in relevant dimensions rather than relying only on a single global geometric shape.

GATEWAY CLASSES

Advantages of ProCLUS

- Designed for high-dimensional data.
- Finds clusters in relevant subspaces.
- Different clusters can have different relevant dimensions.
- Reduces the effect of irrelevant dimensions.
- Uses medoids, which are actual data points.

Limitations of ProCLUS

- The number of clusters generally needs to be specified.
- Choosing suitable parameters can affect the result.
- More complex than simple clustering methods such as K-Means.

Gateway Classes

GATEWAY CLASSES
GATEWAY CLASSES

How CLIQUE and ProCLUS Overcome Traditional Clustering Limitations ?

- Traditional clustering methods such as K-Means may face problems with high-dimensional data, irrelevant dimensions and complex cluster structures.
- CLIQUE performs subspace clustering by finding dense regions in selected subsets of dimensions. Thus, irrelevant dimensions can be avoided.
- ProCLUS performs projective clustering by selecting relevant dimensions separately for different clusters.
- Both methods are suitable for high-dimensional data where meaningful clusters may exist only in some dimensions.
- Therefore, they overcome the limitation of applying clustering to all dimensions together.

Gateway Classes

FREQUENT PATTERN-BASED CLUSTERING

Frequent Pattern-Based Clustering is a clustering method that groups data objects based on frequent patterns or features they share.

A frequent pattern is a pattern that occurs repeatedly in the dataset.

Objects having many common frequent patterns are considered similar and are placed in the same cluster.

It is useful when traditional distance measures are not suitable, especially for transactional and high-dimensional data.

Example : If customers frequently buy Milk, Bread and Butter, customers sharing these patterns can be grouped into the same cluster.

Gateway Classes

GATEWAY CLASSES

- Frequent patterns are used as similarity information between objects.
- First, frequent patterns are discovered from the data.
- Common patterns shared by objects are identified.
- Objects sharing more frequent patterns are considered more similar.
- These similarities are then used to form clusters.
- **Example:**

CUSTOMER	PURCHASED ITEMS
C1	Milk, Bread, Butter
C2	Milk, Bread
C3	Rice, Oil
C4	Milk, Bread, Butter

Here, {Milk, Bread} and {Milk, Bread, Butter} are common patterns.C1, C2 and C4 can therefore form one cluster.

Gateway Classes

GATEWAY CLASSES

BASIC WORKING

GATEWAY CLASSES

- Collect the dataset containing objects and their features/items.
- Find frequent patterns using a minimum support threshold.
- Represent objects using these patterns.
- Measure similarity between objects based on common frequent patterns.
- Group similar objects into the same cluster.
- Form final clusters containing objects with similar frequent patterns.

Gateway Classes

NON-EUCLIDEAN SPACE

Non-Euclidean space is a data space where the similarity or distance between objects cannot be properly represented using ordinary Euclidean distance.

Euclidean distance works well for normal numerical data such as points in 2D or 3D.

Some data has more complex relationships, where other distance or similarity measures are required.

Examples: text documents, graphs, sequences, categorical data, and biological data.

Example : Two documents may be similar because they contain many common words, but their similarity cannot be effectively measured using ordinary Euclidean distance.

GATEWAY CLASSES
Gateway Classes

GATEWAY CLASSES

Need for Non-Euclidean Clustering

Traditional clustering methods such as K-Means mainly use Euclidean distance, which may not work well for complex data.

Non-Euclidean clustering is needed to :

- Handle complex and non-numerical data.
- Use suitable similarity or distance measures.
- Find meaningful groups when Euclidean distance is not appropriate.
- Handle data such as text, graphs, sequences and categorical objects.
- Improve clustering quality for complex relationships.

Gateway Classes

GATEWAY CLASSES

Basic Clustering Approach

- Represent the data in a suitable form.
- Select an appropriate similarity or distance measure for the data.
- Calculate the similarity/distance between objects.
- Group objects having high similarity or small distance.
- Repeat the grouping process until stable clusters are obtained.

STREAM DATA

- Stream data is data that arrives continuously and rapidly over time.
- Data is generated continuously rather than stored all at once.
- The complete dataset may be very large or unlimited.

Examples : sensor data, online transactions, network traffic, social media data and website clicks.

Challenges in Stream Clustering

Continuous arrival : New data keeps coming.

High volume : A large amount of data may arrive quickly.

Limited memory : It is not practical to store the complete stream.

Changing patterns : Data patterns may change over time.

Fast processing : Clusters need to be updated quickly.

Gateway Classes

CLUSTERING CONTINUOUSLY ARRIVING DATA

The clustering process must update clusters as new data arrives.

Basic Working :

- New data objects arrive continuously.
- Each object is processed immediately.
- It is assigned to an existing cluster or used to create/update a cluster.
- Cluster information is updated continuously.
- Old or less relevant information may be removed to save memory.
- The process continues as new data arrives.

Example : In network monitoring, new network activities continuously arrive. Clusters are updated to identify changing patterns of network traffic.

Gateway Classes

GATEWAY CLASSES

GATEWAY CLASSES

Modifications Needed for Stream Clustering ??

Traditional clustering algorithms need modification because stream data arrives continuously and cannot usually be stored completely.

- Incremental processing : Process new data as it arrives instead of processing the complete dataset again.
- Limited memory : Store only important summary information rather than the complete stream.
- Continuous cluster update : Update existing clusters whenever new data arrives.
- Time-based relevance : Give more importance to recent data when data patterns change.
- Fast processing : Use efficient methods because data may arrive at high speed.
- Adaptation to changes : Update or modify clusters when the underlying data pattern changes.

Gateway Classes

PARALLELISM

Parallelism is a technique in which a large task is divided into smaller tasks, and these tasks are processed simultaneously using multiple processors, cores, or machines.

It is mainly used to reduce processing time and handle large amounts of data efficiently.

Example : Suppose a dataset contains 1 million records. Instead of processing all records using one processor

Parallelism in Data Mining

In data mining, parallelism is useful when datasets are very large. The dataset or computation can be divided among multiple processors or machines. Each processor performs its part of the computation, and the partial results are combined to produce the final result.

Gateway Classes

TYPES OF PARALLELISM

1. Data Parallelism : The same operation is performed on different parts of the dataset simultaneously.

Example : A large dataset is divided into four parts, and four processors perform the same calculation on each part.

2. Task Parallelism : Different tasks are performed simultaneously.

Example : One processor performs data cleaning, another performs data transformation, and another performs analysis.

WORKING OF PARALLELISM

1. **Divide the task :** A large task or dataset is divided into smaller independent tasks.
2. **Distribute the tasks :** These tasks are assigned to different processors, cores, or machines.
3. **Process simultaneously :** Multiple processors execute their assigned tasks at the same time.
4. **Collect the results :** The results produced by different processors are combined to obtain the final result.

GATEWAY CLASSES
Gateway Classes

GATEWAY CLASSES

Advantages of Parallelism

- Reduces processing time by performing tasks simultaneously.
- Handles large datasets efficiently.
- Improves system performance.
- Makes better use of multiple processors/cores.
- Supports scalable data processing.
- Useful in Big Data, data mining, machine learning, and scientific computing.

Limitations

- Requires multiple processors or computing resources.
- Some tasks cannot be divided easily into independent parts.
- Communication and synchronization between processors may add overhead.
- Managing parallel tasks can be more complex than sequential processing.

Gateway Classes

Gateway Classes

@GATEWAY CLASSES

Empowering Learners, Transforming Futures

THANK YOU

Gateway Classes

GET IT ON Google Play

Download on the App Store

Helpline No- 7819 0058 53 , 9259833200

AKTU B.Tech 3rd Year

CS & CS Allied

Data Analytics & Visualization

Unit-4

in ONE SHOT

Part-II

AKTU PYQs Covered



By Shubhangi Ma'am

GATEWAY CLASSES

Gateway Classes

SYLLABUS

UNIT - IV

Frequent Itemsets and Clustering

Mining frequent itemsets, market based modelling, Apriori algorithm, handling large data sets in main memory, limited pass algorithm, counting frequent itemsets in a stream, clustering techniques: hierarchical, K-means, clustering high dimensional data, CLIQUE and ProCLUS, frequent pattern based clustering methods, clustering in non-Euclidean space, clustering for streams and parallelism.

Gateway Classes

Q.1 : For the given data, find the association rules using Apriori Algorithm. AKTU(2022-23).

TID	Itemsets
T1	A, B
T2	B, D
T3	B, C
T4	A, B, D
T5	A, C
T6	B, C
T7	A, C
T8	A, B, C, E
T9	A, B, C

Given :

- **Minimum Support = 2**
- **Minimum Confidence = 50%**

Gateway Classes

Solution : Total transactions = **9**

Step 1 : Generate C_1 and L_1

Count each item : Minimum support = 2.

Therefore, **E is pruned** because its support count is 1. **L_1,** L1={A,B,C,D}

Item	Support Count
A	6
B	7
C	6
D	2
E	1

Gateway Classes

Step 2 : Generate C_2

Possible 2-itemsets : Apply minimum support = 2.

- AB ✓
- AC ✓
- AD ✗
- BC ✓
- BD ✓
- CD ✗

L_2 , L2={AB,AC,BC,BD}

Itemset	Support Count
AB	4
AC	4
AD	1
BC	4
BD	2
CD	0

Gateway Classes

Step 3 : Generate C_3 and Pruning

Possible 3-itemsets : **ABC**

Its 2-item subsets are:

- AB ✓
- AC ✓
- BC ✓

So **ABC is not pruned**.

Support of ABC : It occurs in:

- T8
- T9

Therefore :

$Support(ABC) = 2$ So, ABC is frequent.

ABD

Its subsets include : $AB,\ AD,\ BD$

But **AD is infrequent**, therefore : $\boxed{ABD \text{ is pruned}}$

Other 3-itemsets also contain an infrequent 2-itemset, so they are pruned. $\mathbf{L_3}$, $L_3 = \{ABC\}$

No further frequent itemset can be generated.

Frequent Itemsets

Therefore, all frequent itemsets are :

1-itemsets : $\boxed{A, B, C, D}$

2-itemsets : $\boxed{AB, AC, BC, BD}$

3-itemsets : ABC

Gateway Classes

Step 4 : Generate Association Rules

Minimum confidence = **50%**

Formula : $Confidence(X \rightarrow Y) = \frac{Support(X \cup Y)}{Support(X)} \times 100$

From AB : A → B $= \frac{4}{6} \times 100 = 66.67\%$

✓ Strong

B → A : $\frac{4}{7} \times 100 = 57.14\%$

✓ Strong

From AC : A → C : $\frac{4}{6} \times 100 = 66.67\%$

✓ Strong

C → A : $\frac{4}{6} \times 100 = 66.67\%$

✓ Strong

From BC : B → C : $\frac{4}{7} \times 100 = 57.14\%$

✓ Strong

C → B : $\frac{4}{6} \times 100 = 66.67\%$

✓ Strong

Gateway Classes

From BD

$\mathbf{B \rightarrow D : \frac{2}{7} \times 100 = 28.57\%}$

✗ Not strong

$\mathbf{D \rightarrow B : \frac{2}{2} \times 100 = 100\%}$

✓ Strong

From ABC

Possible rules :

$\mathbf{AB \rightarrow C : \frac{2}{4} \times 100 = 50\%}$

✓ Strong

$\mathbf{AC \rightarrow B : \frac{2}{4} \times 100 = 50\%}$

✓ Strong

$\mathbf{BC \rightarrow A :}$ 42×100=50%

Gateway Classes

Q.2 : The database has 6 transactions. Assume Support threshold = 50%, Confidence = 60%.

(i) Use Apriori algorithm to find all frequent itemsets.

(ii) Show all the strong association rules (with support and confidence). AKTU(2024-25)

TID	Items Bought
10	Beer, Nuts, Diaper
20	Beer, Coffee, Diaper
30	Beer, Diaper, Eggs
40	Nuts, Eggs, Milk
50	Nuts, Coffee, Diaper, Eggs, Milk
60	Beer, Nuts, Diaper

Gateway Classes

1. Total Transactions = 6
Minimum Support = 50%
Minimum confidence = 60%

Support count Threshold $= \frac{50}{100} \times 6 = 3$

So, minimum support count = 3

Step-1 C_1 - 1-itemsets

item	count
Beer (B)	4
Nuts (N)	4
Diaper (D)	5
Coffee (C)	2
Eggs (E)	3
Milk (M)	2

$L1 = \{Beer, Nuts, Diaper, Eggs\}$

Coffee & Milk - discard

Gateway Classes

Step-2 C_2 - 2 itemsets

2-itemset	Count
Beer, nuts	2
Beer, Diaper	4
Beer, Eggs	1
Nuts, Diaper	3
Nuts, Eggs	2
Diaper, Eggs	2

Therefore,

$L_2 = \{\{Beer, Nuts\}, \{Nuts, Diaper\}\}$

Step 3 - C_3 - 3 itemsets

Beer, Nuts, Diaper

No - 3 itemset

Gateway Classes

(2) Strong Association Rules

$$\text{Confidence}(X \rightarrow Y) = \frac{\text{Support}(X \cup Y)}{\text{Support}(X)} \times 100$$

(1) from {Beer, Diaper}

Support count = 4

Beer → Diaper

$\text{Confidence} = \frac{4}{4} \times 100 = 100\%$

$\text{Support} = \frac{4}{6} \times 100 = 66.67\%$

100% > 60% Strong

Diaper → Beer

$\text{Confidence} = \frac{4}{5} \times 100 = 80\%$

$\text{Support} = \frac{4}{6} \times 100 = 66.67\%$

80% > 66.67%

Strong

(2) from {Nuts, Diaper}

Support count = 3

Nuts → Diaper

$\text{Confidence} = \frac{3}{5} \times 100 = 60\%$

$\text{Support} = \frac{3}{6} \times 100 = 50\%$

Strong (60% > 50%)

Diaper → Nuts

$\text{Confidence} = \frac{3}{5} \times 100 = 60\%$

$\text{Support} = \frac{3}{6} \times 100 = 50\%$

Strong (60% > 50%)

Gateway Classes

Q.3 : Given transactions :

Minimum Support = 50%

Find all frequent itemsets using Apriori algorithm and show candidate generation and pruning.

TID	Items
T1	A, B, C
T2	A, B
T3	A, C
T4	B, C
T5	A, B, C
T6	A, B

Solution : total transactions = 6

Minimum support count : $6 \times 50\% = 3$

C1

Item	Count
A	5
B	5
C	4

All ≥ 3. L1 = {A , B , C}

Gateway Classes

C2 : Possible 2-itemsets : $AB,\ AC,\ BC$

Counts

- AB = 4
- AC = 3
- BC = 3

All are frequent. L2 = {AB , AC , BC}

Final Answer : $\boxed{L_1 = \{A, B, C\}}$ $\boxed{L_2 = \{A, B, AC,\ BC\}}$

No frequent 3-itemset.

Gateway Classes

Q.4 : Given,

Minimum Support = 40%

Minimum Confidence = 70%

Find frequent itemsets and generate strong association rules.

TID	Items
T1	A, B, C
T2	A, B
T3	A, C
T4	B, C
T5	A, B, C
T6	A, B, C
T7	B, C

Solution : Total transactions = 7

Minimum support count : $7 \times 40\% = 2.8$

So minimum count = **3**.

1-itemsets

- A = 5
- B = 6
- C = 6 , Therefore : L1={A,B,C}

Gateway Classes

2-itemsets :

- AB = 4
- AC = 4
- BC = 5

All ≥ 3 , $L_2 = \{A, B, ACBC\}$

3-itemset : ABC occurs in T1, T5, T6 : $ABC = 3$

So , L3={ABC}

Association Rules

For $AB : A \rightarrow B$

$$Confidence = \frac{Support(AB)}{Support(A)} = \frac{4}{5} = 80\%$$

Strong ✓ B→A = 64 = 66.67%

For $AC : A \rightarrow C = \frac{4}{5} = 80\%$

Strong ✓

C → A= 64 = 66.67%

For $BC : B \rightarrow C = \frac{5}{6} = 83.33\%$

Strong ✓

C→B=65=83.33%

Strong Rules : C→B

$\boxed{A \rightarrow B}$

$\boxed{A \rightarrow C}$

$\boxed{B \rightarrow C}$

Gateway Classes

Q.5 : A database contains 12 transactions.

Itemset {Milk, Bread} occurs in 7 transactions.

Minimum Support = 60%.

Determine:

1. **Support count**
2. **Support percentage**
3. **Whether the itemset is frequent.**

Solution : Support count : $\boxed{7}$

Support : $\frac{7}{12} \times 100 = 58.33\%$

Compare: $58.33\% < 60\%$

Therefore : Infrequent Itemset

Gateway Classes

Q.6 : Consider four points:

- **A(3,2)**
- **B(2,1)**
- **C(7,8)**
- **D(8,9)**

Take: $K = 2$

Initial centroids: $C_1 = (1,2), C_2 = (7,8)$

Use Euclidean distance. Find the final clusters.

Solution : Formula : $d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$

① Point(A)

$A \rightarrow C_1(1,2)$

$d = \sqrt{(3-1)^2 + (2-2)^2} = 2$

$A \rightarrow C_2$

$d = \sqrt{(3-7)^2 + (2-8)^2} = 7.21$

So, $A \rightarrow$ Cluster 1

② point(B)

$B \rightarrow C1$

$d = \sqrt{(2-1)^2 + (1-2)^2} = 1.41$

So $B \rightarrow$ cluster 2

$B \rightarrow C2$ $d = \sqrt{(2-7)^2 + (1-8)^2} = 8.60$

Gateway Classes

(3)

Point C (7, 8)

$C \rightarrow c1$

$$d = \sqrt{(7-1)^2 + (8-2)^2} = 8.49$$

$C \rightarrow C_2$

$$d = \sqrt{(7-7)^2 + (8-8)^2} = 0$$

So, C → cluster 2

(4) point D (8, 9)

$D \rightarrow c1$

$$d = \sqrt{(8-1)^2 + (9-2)^2} = 9.90$$

$D \rightarrow C_2$

$$d = \sqrt{(8-7)^2 + (9-8)^2} = 1.41$$

So D → cluster 2

Gateway Classes

Calculate New Centroids

Cluster 1 contains $A=(3,2)$
$B(2,1)$

X-coordinate: $\frac{3+2}{2} = 2.5$

Y-coordinate $= \frac{2+1}{2} = 1.5$

$C_1 = (2.5, 1.5)$

Cluster 2 contains $C=(7,8)$
$D=(8,9)$

X-coordinate $= \frac{7+8}{2} = 7.5$

Y-coordinate $= \frac{8+9}{2} = 8.5$

$C_2 = (7.5, 8.5)$

Gateway Classes

Q.9 : Given transactions : AKTU(2025-26)

- T1 = {A, B, C}
- T2 = {A, B}
- T3 = {A, C}
- T4 = {B, C}
- T5 = {A, B, C}

If **minimum support = 60%**, find all frequent itemsets and show pruning.

Gateway Classes

Solution

Total transactions = **5**

Minimum Support Count : $5 \times 60\% = 3$

So, minimum support count = **3**.

Step 1 : $C_1 \rightarrow L_1$

Count each item:

All have count ≥ 3.

Therefore:

L1={A,B,C}

Item	Support Count
A	4
B	4
C	4

Step 2 : $C_2 \rightarrow L_2$

Generate 2-itemsets:

All have count ≥ 3.

Therefore:

L2 ={AB,AC,BC}

Itemset	Support Count
AB	3
AC	3
BC	3

Gateway Classes

Step 3 : C_3 and Pruning

Generate : $C_3 = \{ABC\}$

Its 2-item subsets are :

- AB ✓
- AC ✓
- BC ✓

Since **all subsets are frequent**, ABC is **not pruned**.

Now calculate support : ABC occurs in:

- T1
- T5

So : $Support(ABC) = 2$

But minimum support count = 3.

$2 < 3$

Therefore : ABC is pruned

Final Frequent Itemsets

1-itemsets : $\boxed{\{A\}, \{B\}, \{C\}}$

2-itemsets : $\boxed{\{A, B\}, \{A, C\}, \{B, C\}}$

3-itemsets : None

Gateway Classes

PREVIOUS YEAR IMPORTANT QUESTIONS

Q.1 : Given transactions : AKTU(2025-26)

- T1 = {A, B, C}
- T2 = {A, B}
- T3 = {A, C}
- T4 = {B, C}
- T5 = {A, B, C}

If **minimum support = 60%**, find all frequent itemsets and show pruning.

Gateway Classes

Q.2 : For the given data, find the association rules using Apriori Algorithm. AKTU(2022-23).

TID	Itemsets
T1	A, B
T2	B, D
T3	B, C
T4	A, B, D
T5	A, C
T6	B, C
T7	A, C
T8	A, B, C, E
T9	A, B, C

Given :

- **Minimum Support = 2**
- **Minimum Confidence = 50%**

Gateway Classes

Q.3 : The database has 6 transactions. Assume Support threshold = 50%, Confidence = 60%.

(i) Use Apriori algorithm to find all frequent itemsets.

(ii) Show all the strong association rules (with support and confidence). AKTU(2024-25)

TID	Items Bought
10	Beer, Nuts, Diaper
20	Beer, Coffee, Diaper
30	Beer, Diaper, Eggs
40	Nuts, Eggs, Milk
50	Nuts, Coffee, Diaper, Eggs, Milk
60	Beer, Nuts, Diaper

Gateway Classes

Q.4 : 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)
- A7(1, 2)

The distance function between two points

a = (x_1, y_1) and b = (x_2, y_2) is : $P(a, b) = | x_2 - x_1 | + | y_2 - y_1 |$

Use K-Means Algorithm to find the three cluster centers after assigning all eight points.

AKTU(2024-25)

Gateway Classes

Gateway Classes

Empowering Learners, Transforming Futures

THANK YOU

Gateway Classes

GET IT ON Google Play

Download on the App Store

Helpline No- 7819 0058 53 , 9259833200