Unit-2: Association Rule Mining

Introduction:

Association Rule Mining is a part of the data mining process which is performed to find the interesting association rules from a large sample of datasets.

An association rule is a pattern/relationship which portrays the frequent association of two itemsets.

It is represented as: A => B(where, A, B are itemsets (one/more))

which means that if a customer bought the items of (A) set, then the customer is likely to also buy the items of (B) set.

This Mining is performed on the Transactional databases of large no.of E.com platforms to increase the sales and maximize profits.

Ex: if a customer buys milk, then he is more likely to buy bread or biscuits.

I.e., {milk} => {bread, biscuits}

Ex: if a customer buys a toothbrush and toothpaste, then he is more likely to also buy the mouth-washer.

I.e, {toothbrush, toothpaste} => {mouth-washer}

Hence, the process of analyzing the large transactional datasets and finding the interesting association rules, is known as Association Rule mining.

A transactional dataset is a dataset containing information about the previously performed transactions (items bought) by the customers.

It only contains the transaction_ID column and itemsets column.

Mining Frequent Patterns:

Frequent Patterns are nothing but the interesting patterns or associations which occur frequently/redundantly in the given dataset.

That is, the patterns that appear frequently in the given large sample of dataset, are known as frequent patterns.

A frequent pattern can be a frequent dataitem, frequent sequence (sequence of data items) or frequent substructures (such as, graph, tree, etc).

Ex: 'Milk' dataitem occurs frequently.

Ex: {Milk, Bread, Biscuits} occurs frequently.

Market Basket Analysis (MBA) is a technique which aims for frequent patterns mining. It analyses the customers' buying habits by finding associations between different items that customers place in their shopping baskets.

Frequent patterns are represented by association rules.

Association Rule is represented as: A => B(where, A,B are items sets)

Here, set(A) (LHS) is called Antecedent, and set(B) (RHS) is called Consequent.

The metrics used in association mining for measuring interestingness of patterns are: support and confidence.

-> Support: The proportion of transactions that contain both the antecedent(A) and the consequent(B).

I.e., How frequently the determined association rule is applicable/appears in the given dataset.

Support(A=>B) =

For a single dataitem or an itemset:

Support(A) =

-> Confidence: It measures the strength/reliability of the association rule. The frequent occurrence of (B) itemset containing the items of (A). (or) The Support of the rule itself (A=>B) (antecedent and consequent (AUB) over the support of antecedent (A).

Confidence(A=>B) =

Mining Methods:

The are two methods (algorithms) for mining the interesting association rules from the given large sample of transaction dataset, viz.:

(1) Apriori Algorithm

(2) FP Growth Algorithm

The aim of both these algorithms is determining the interesting frequent itemsets by finding strong association rules.

In these algorithms we use minimum support (min_sup) and minimum confidence (min_con) to eliminate weak association rules and have only strong association rules (known as threshold values).

I.e., if only some of the customers bought a set of things, doesn't mean that we must consider it as a new pattern. We only consider those patterns which occur most frequently (I.e., the majority of the customers must buy the set of things). ence, to eliminate those weak patterns (non-frequent), we set a minimum-limit for the value of support and confidence.

Hence, the strong association rules are those rules whose support and confidence values are greater than or equal to the minimum support and confidence values, respectively.

(1) Apriori Algorithm:

The two primary drawbacks of the Apriori Algorithm are:

At each step, candidate sets have to be built.

To build the candidate sets, the algorithm has to repeatedly scan the database.

These two properties inevitably make the algorithm slower.

To overcome these redundant steps, a new association-rule mining algorithm was developed named Frequent Pattern Growth Algorithm.

It overcomes the disadvantages of the Apriori algorithm by storing all the transactions in a Trie Data Structure.

(2) FP (Frequent Pattern) Growth Algorithm:

The FP-Growth algorithm is a frequent pattern mining algorithm that efficiently discovers frequent itemsets in a dataset.

It avoids the need for generating candidate itemsets like the Apriori algorithm by using a data structure called the FP-Tree.

Mining Various Kinds of Association Rules:

Mining various kinds of association rules refers to the process of discovering different types of relationships and patterns within a dataset.

Association rule mining aims to find associations, correlations, or dependencies among items or attributes in the data.

The different kinds of association rules that can be mined include:

(1) Mining Multi-Level Association Rules:

- Multi-level association rules involve relationships between items at various levels in the antecedent and/or consequent of the rule.

- This means considering combinations of items that occur at different hierarchical levels.

- This approach provides a broader view and deeper understanding of the connections between different levels of items.

Example: Consider a retail scenario. A multi-level association rule could be:

{Category: Electronics, Subcategory: Laptops} -> {Brand: XYZ, Rating: High}

This rule suggests that when customers purchase laptops in the Electronics category, particularly the XYZ brand, they tend to rate the product as "High."

(2) Mining Multi-Dimensional Association Rules from Databases or Data Warehouses:

- Multi-dimensional association rule mining analyzes data along multiple dimensions (e.g., time, location, customer segment).

- It discovers associations specific to different combinations of dimension values.

- For instance, it may uncover relationships between specific products sold in a particular region during a specific period.

Example: In a sales database, a multi-dimensional association rule might be:

{**Region: East, Product: Electronics} -> {Customer Segment: Corporate, Promotion: Yes}**

This rule implies that in the Eastern region, when selling electronics, the corporate customer segment responds positively when a promotion is applied.

(3) Mining Multi-Dimensional Association Rules from Static Discretization of Quantitative Attributes:

- This involves discovering association rules when dealing with datasets containing quantitative (numeric) attributes.

- Static discretization refers to converting continuous numerical values into discrete intervals or categories before applying association rule mining.

- By discretizing the attributes, associations can be discovered based on the defined intervals.

Example: Assume a dataset with a "Salary" attribute. A multi-dimensional association rule could be:

{Age: 25-35, Salary: Moderate} -> {Education: College, Job Type: Full-Time}

This rule suggests that individuals aged 25-35 with a moderate salary, often have a college education and work in full-time positions.

(4) Mining Quantitative Association Rules:

- Mining quantitative association rules involve discovering patterns where both the antecedent and consequent of the rule contain quantitative (numeric) attributes.

- It identifies associations based on values or numerical relationships rather than discrete items.

- This type of mining helps understands numerical dependencies and correlations between attributes in the dataset.

Example: Consider a financial dataset. A quantitative association rule might be:

{Income: >$100,000, Savings: >$20,000} -> {Investment: High Risk}

This rule suggests that individuals with high income and substantial savings are more likely to opt for high-risk investment strategies.

Constraint Based Association Mining

Constraint-based association mining refers to the process of discovering interesting patterns, associations, or rules in large datasets while incorporating specific constraints/rules to guide the mining process.

These constraints help in focusing the analysis on patterns that meet certain criteria.

The constraint types applied are:

(1) Knowledge Type Constraint

(2) Data Constraint

(3) 3D Level Constraint

(4) Interestingness Constraints

(5) Rule Constraints

(1) Knowledge Type Constraint:

These constraints are related to the type of knowledge or patterns that the association mining process aims to discover.

Example: If you are interested in discovering frequent itemsets related to customer purchasing behavior, your knowledge type constraint could specify that you are only interested in association rules related to certain types of products or specific customer segments.

(2) Data Constraint:

Data constraints involve restrictions based on the characteristics of the dataset being analyzed.]

Example: You might set a data constraint to consider only transactions that occurred within a specific time period or involve a certain subset of items.

(3) 3D Level Constraint:

3D level constraints involve applying constraints to associations in a three-dimensional space, considering multiple dimensions or attributes simultaneously.

Example: In a retail dataset, we can apply a 3D level constraint to find associations between products, customer demographics, and geographic locations. This helps identify specific patterns for different customer segments and regions

(4) Interestingness Constraints:

Interestingness constraints are used to filter out uninteresting or trivial rules. They help focus on rules that are more likely to be valuable or insightful.

Example: We can set an interestingness constraint to consider only associations with a minimum support of 5% and a minimum confidence of 80%. This ensures that the discovered associations are statistically significant and reliable.

(5) Rule Constraints:

Rule constraints specify additional conditions or requirements that association rules must satisfy.

Example: We can set a rule constraint that an association rule should have a certain item in the antecedent (left-hand side) or consequent (right-hand side). For example, we might require an association rule to include the item "milk" in the consequent

Graph Pattern Mining: