## Mining Multidimensional Association Rules

Multidimensional Association Rule Mining extends traditional association rule mining by considering multiple attributes (dimensions) in a dataset rather than just items in transactions. This allows for more complex and meaningful insights, particularly when analyzing datasets with categorical, numerical, or hierarchical attributes.

Traditional association rules typically focus on relationships between items in a transaction, such as:

📌 **Example of a traditional association rule:**
**{Milk, Bread} → {Butter}**
(Support = 30%, Confidence = 80%)

However, real-world datasets often contain multiple attributes such as **age, gender, location, time, and product category**, which can provide deeper insights. **Multidimensional association rule mining** incorporates these attributes, making it useful for various domains.

📌 **Example of a multidimensional association rule:**
**(Age = 20-30, Gender = Male, Location = Urban) → (Buys = Sports Shoes)**
(Support = 15%, Confidence = 70%)

This rule indicates that young urban males are more likely to buy sports shoes.

### Types of Multidimensional Association Rules

Multidimensional rules can be classified into three types:

1. **Inter-Dimension Association Rules**
    - These rules involve relationships between **different dimensions** in the dataset.
    - **Example:** (Education = Graduate, Income = High) → (Buys = Luxury Car)
        - This shows that highly educated, high-income individuals are likely to buy luxury cars.
2. **Hybrid-Dimension Association Rules**
    - These rules mix both **categorical attributes and item-based transactions**.
    - **Example:** (Age = 25-40, Location = USA) → {Buys = Smartphone, Laptop}
        - This rule suggests that people aged 25-40 in the USA frequently buy smartphones and laptops together.
3. **Intra-Dimension Association Rules**
    - These rules capture patterns within the **same dimension** (e.g., product categories).
    - **Example:** (Product Category = Electronics) → (Buys = Laptop, Smartphone)

- This shows a strong relationship between products within the same category.

**Techniques for Mining Multidimensional Association Rules**

(a) Generalized Apriori Algorithm

The Apriori algorithm is a classic method for mining frequent itemsets. When extended to multiple dimensions:

- It treats categorical attributes as items.
- It prunes infrequent attribute combinations.
- The minimum support and confidence thresholds guide rule selection.

📌 Example:
If we have a dataset with attributes like Age, Income, Product Bought, the algorithm finds frequent patterns such as:
(Age = 30-40, Income = Medium) → (Buys = Smartwatch)
(Support = 12%, Confidence = 60%)

(b) FP-Growth (Frequent Pattern Growth)

FP-Growth is an optimized alternative to Apriori that constructs an FP-Tree for efficient mining.

- It eliminates candidate generation, making it faster for large datasets.
- It finds frequent patterns by recursively dividing the dataset.

📌 Example:
Using FP-Growth, we may discover:
(Region = Asia, Payment Method = Credit Card) → (Buys = Smart TV)
(Support = 10%, Confidence = 75%)

(c) Constraint-Based Mining

- In many real-world applications, users are interested in only specific types of rules.
- Constraints like min/max support, confidence thresholds, or specific dimensions can be applied.

📌 Example:
A retailer may set constraints such as:

- Only mine rules for customers aged 18-35.
- Only analyze purchases above $500.
  This helps extract relevant insights while reducing noise.

(d) OLAP-Based Data Cube Approach

- Online Analytical Processing (OLAP) cubes allow efficient multidimensional analysis.
- Data is stored in cubes where dimensions can be drilled down or rolled up for detailed insights.

📌 Example:
A supermarket database stores sales data as a cube with dimensions:

- Time (Daily, Weekly, Monthly)
- Product (Electronics, Clothing, Groceries)
- Location (USA, Europe, Asia)

By drilling down, we might find:
(Month = December, Location = USA) → (Buys = Winter Jackets)
(Support = 20%, Confidence = 85%)

**Example: Mining Multidimensional Association Rules in Retail**

Consider a **grocery store database** with the following attributes:

| CustomerID | Age | Gender | Income | Location | Product Bought |
|---|---|---|---|---|---|
| C001 | 25 | Male | Medium | Urban | Bread, Milk |
| C002 | 32 | Female | High | Suburban | Wine, Cheese |
| C003 | 40 | Male | High | Urban | Steak, Wine |
| C004 | 28 | Female | Medium | Rural | Bread, Butter |
| C005 | 30 | Male | Medium | Urban | Milk, Eggs |

**Step 1: Data Preprocessing**

Convert categorical attributes into **binary form** (one-hot encoding) or use hierarchical categorization.

**Step 2: Apply Apriori or FP-Growth Algorithm**

Find frequent patterns such as:
**(Age = 25-35, Income = Medium, Location = Urban) → (Buys = Milk, Bread)**
(Support = 18%, Confidence = 70%)

**Step 3: Interpret Results**

- **Business Insight:** Young urban customers with medium income frequently buy milk and bread together.

- **Actionable Strategy:** Offer discounts or bundle deals on these products for this demographic.

**Correlation Analysis**

**Correlation analysis** in data mining is used to measure the relationship between variables in large datasets. It helps in understanding dependencies, identifying trends, and improving decision-making. It plays a key role in **feature selection, association rule mining, anomaly detection, and market analysis**.

📌 **Example Use Cases:**

- **Retail:** Understanding the correlation between product sales (e.g., "Does an increase in soft drink sales correlate with an increase in snack sales?").
- **Healthcare:** Finding relationships between **lifestyle factors** (e.g., smoking, diet) and diseases.
- **Finance:** Identifying correlations between **stock prices and interest rates**.

**Types of Correlation**

1. **Positive Correlation (Direct Relationship)**
    - When one variable **increases**, the other variable **also increases**.
    - **Example:** The more time spent studying, the higher the test score.
2. **Negative Correlation (Inverse Relationship)**
    - When one variable **increases**, the other variable **decreases**.
    - **Example:** As the speed of a car increases, the time taken to reach a destination decreases.
3. **No Correlation (No Relationship)**
    - No meaningful relationship between the two variables.
    - **Example:** A person's height and their phone number.

**Measures of Correlation**

Several statistical methods are used to measure correlation:

**(a) Pearson Correlation Coefficient (r)**

- Measures **linear correlation** between two continuous variables.
- Values range from **-1 to +1**:
    - **+1** → Perfect positive correlation
    - **0** → No correlation

- **-1** → Perfect negative correlation

📌 **Formula:**

$$r = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2 \sum(Y_i - \bar{Y})^2}}$$

**Spearman's Rank Correlation (ρ)**

- Measures **monotonic relationships** (not necessarily linear).
- Used when data is **ordinal** (ranked data).

📌 **Formula:**

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

where **di** is the difference between the ranks of corresponding values.

**Chi-Square Test for Categorical Data**

- Used to find relationships between **categorical** variables.
- Helps answer: **"Are two variables independent?"**

📌 **Formula:**

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

/here:

- $O_i$ = observed frequency
- $E_i$ = expected frequency

**Kendall's Tau (τ)**

- Measures the strength of **ordinal associations** between two variables.
- Similar to Spearman's rank but more robust for **small datasets with tied ranks**.

📌 **Example:**

- Relationship between **employee performance rankings** and **salary increase rankings**.

**Applications of Correlation Analysis**

🔹 **Feature Selection:** Helps identify the most relevant attributes in machine learning models.
🔹 **Market Basket Analysis:** Finds correlations between product purchases (e.g., "People who buy laptops often buy mouse and keyboards").
🔹 **Fraud Detection:** Detects unusual patterns in financial transactions.
🔹 **Healthcare & Epidemiology:** Identifies correlations between **symptoms and diseases**.
🔹 **Stock Market Analysis:** Finds relationships between **stock prices, inflation, and interest rates**.

**Constraint Based Association Mining**

Constraint-Based Association Mining is an advanced approach in association rule mining where users define constraints to focus on specific patterns instead of mining all possible association rules. This method enhances efficiency, improves relevance, and reduces computational complexity by filtering out uninteresting or irrelevant rules.

📌 Example: Instead of mining all possible purchase patterns, a store manager may only want to find relationships between expensive electronic items (like laptops and smartphones).

Standard association rule mining (e.g., Apriori, FP-Growth) may generate a large number of rules, many of which are useless or irrelevant. Constraints help by:

1. Focusing on relevant rules that match business needs.
2. Reducing computational complexity by filtering out unimportant itemsets.
3. Handling large datasets efficiently.
4. Providing user control over the mining process.

**Types of Constraints in Association Rule Mining**

Constraints can be applied to different aspects of the mining process:

1. Knowledge-Based Constraints

- Users specify prior knowledge to focus on specific relationships.
- 📌 Example:
    - "Find associations only between electronic items."
    - "Ignore rules involving low-cost products."

2. Data Constraints

- Based on attributes like price, quantity, or category.

- 📌 Example:
    - "Find associations only for transactions above $100."
    - "Consider purchases only from weekends."

3. Aggregate Constraints

- Based on aggregated values like sum, count, or average.
- 📌 Example:
    - "Find product sets where the total price exceeds $200."
    - "Consider items that appear in at least 500 transactions."

4. Length Constraints

- Controls the size of itemsets in the rules.
- 📌 Example:
    - "Find association rules with exactly 3 items."
    - "Exclude rules with more than 4 items."

5. Hierarchical Constraints

- Allows mining at different levels of abstraction in category hierarchies.
- 📌 Example:
    - "Find associations between broad product categories (e.g., dairy, beverages) instead of individual items."
    - "Analyze patterns at brand level instead of product level."

**Example: Constraint-Based Association Rule Mining**

Scenario:

A retail store wants to find frequent purchase patterns, but only for electronics that cost $100 or more.

Step 1: Sample Transactions

| Transaction ID | Items Purchased |
|---|---|
| 1 | Laptop, Mouse, Keyboard |
| 2 | Smartphone, Headphones |
| 3 | Laptop, Smartphone, TV |

| Transaction ID | Items Purchased |
|---|---|
| 4 | TV, Sound System |
| 5 | Smartphone, Charger |

Step 2: Apply Constraints

- Category constraint: Only consider electronics.
- Price constraint: Ignore items below $100.

Step 3: Generate Frequent Itemsets

- {Laptop, Smartphone} (support: 40%) ✅
- {TV, Sound System} (support: 30%) ✅
- {Smartphone, Headphones} (support: 20%) ❌ (below threshold)

Step 4: Generate Association Rules

- Rule 1: If a customer buys a Laptop, they are likely to buy a Smartphone (confidence: 75%).
- Rule 2: If a customer buys a TV, they are likely to buy a Sound System (confidence: 80%).

**Real-World Applications of Constraint-Based Mining**

**1.** Retail & Market Basket Analysis

- 🔹 Find purchase patterns only in high-value transactions.
- 🔹 Identify seasonal purchase trends (e.g., summer vs. winter).

2. Healthcare

- 🔹 Analyze disease patterns only for specific age groups.
- 🔹 Find correlations between lifestyle and diseases (e.g., smoking & lung disease).

3. Fraud Detection

- 🔹 Detect fraudulent transactions above a specific amount.
- 🔹 Find patterns of suspicious activities in banking.

4. Web & Social Media Mining

- 🔹 Analyze trends only in a specific geographic region.
- 🔹 Identify hashtags that appear at least 1000 times in tweets.