Association & Correlation in Data Mining

Aditya Raj
RA2311003030555
CSE(Core)-I

What is Association Analysis?

Content:

- Retail product placement
- Recommendation systems (Amazon, Netflix)
- Medical diagnosis patterns
- Fraud detection
- Web usage mining

Market Basket Analysis

Content:

- A technique used to find associations between products purchased together.
- Helps retailers:
  - Improve product placement
  - Design combo offers
  - Increase sales

Example:

Milk → Bread → Butter

Key Terms in Association Rules

Content:

- **Itemset:** Collection of one or more items.
- **Support:** Frequency of itemset in dataset.
- **Confidence:** Likelihood that item B is purchased when item A is purchased.
- **Lift:** Measures strength of association rule.

Formula:

- Support(A) = Transactions containing A / Total transactions
- Confidence(A→B) = Support(A∪B) / Support(A)

Example of Association Rule

Content:
Rule: {Milk} → {Bread}

- Support = 40%
- Confidence = 70%
- Lift > 1 → Strong relationship

Meaning:
Customers buying milk are likely to buy bread.

Apriori Algorithm (Overview)

Content:

- Used to find frequent itemsets.
- Based on the **Apriori Principle**:
  If an itemset is frequent, all its subsets must also be frequent.

Steps:

1. Find frequent 1-itemsets.
2. Generate candidate itemsets.
3. Prune infrequent itemsets.
4. Repeat until no new itemsets.

Apriori Algorithm Example

Content:

Transactions:

- T1: Milk, Bread
- T2: Milk, Butter
- T3: Milk, Bread, Butter

Frequent Itemsets:

- Milk ✔
- Bread ✔
- Milk & Bread ✔

Mining Without Candidate Generation (FP-Growth)

Content:

- FP-Growth improves efficiency over Apriori.
- Uses **FP-Tree** structure.
- Avoids generating candidate itemsets.

Advantages:

- Faster
- Less memory usage
- Suitable for large datasets

Vertical Data Format Mining

Content:

- Stores data in vertical format (item → transaction IDs).
- Makes support counting faster.
- Used in algorithms like **ECLAT**.

Example:

Milk → T1, T2, T3

Bread → T1, T3

Closed Frequent Itemsets

Content:

- A frequent itemset with no superset having same support.
- Reduces redundancy.
- Helps compress frequent patterns.

Benefit:

Smaller result set, same information.

Multilevel Association Rules

Content:

- Finds associations at different abstraction levels.

Example:

- Level 1: Dairy → Bakery
- Level 2: Milk → Bread

Use: Retail product hierarchy.

Multidimensional Association Rules

Content:

- Rules involving multiple attributes.

Example:

Age = 20–30 AND Buys = Laptop → Buys = Headphones

Used in:

- Customer profiling
- Targeted marketing

Correlation Analysis

Content:

- Measures strength of relationship between variables.
- Helps identify meaningful patterns.

Types:

- Positive correlation
- Negative correlation
- No correlation

Example:

Temperature ↑ → Ice cream sales ↑

Constraint-Based Association Mining

Content:

- Applies constraints to focus on useful rules.
- Reduces number of irrelevant patterns.

Types of Constraints:

- Item constraints
- Rule constraints
- Aggregate constraints

Benefit:

Faster and more relevant results.

Applications of Association Mining

Content:

- Retail product placement
- Recommendation systems (Amazon, Netflix)
- Medical diagnosis patterns
- Fraud detection
- Web usage mining

Thank You !