SNS COLLEGE OF TECHNOLOGY

Coimbatore-35
An Autonomous Institution

Accredited by NBA – AICTE and Accredited by NAAC – UGC with 'A+' Grade
Approved by AICTE, New Delhi & Affiliated to Anna University, Chennai

DEPARTMENT OF INFORMATION TECHNOLOGY

Data Mining and Warehousing

COURSE NAME: Data Mining and Warehousing

COURSE CODE: 19ITT301

SEMESTER: 5

CONTENTS:

- Association Analysis to Correlation Analysis
- Explore Weka and run Apriori algorithm with different support and confidence values (Supermarket dataset)

Association Analysis to Correlation Analysis: Introduction

- **Definition**:
  - **Association Analysis**: Focuses on discovering interesting relationships (or association rules) between variables in large datasets.
  - **Correlation Analysis**: Measures the strength and direction of a linear relationship between two quantitative variables.
- **Transition**: Association analysis typically deals with finding patterns in categorical data, whereas correlation analysis focuses on quantitative data.
- **Purpose**: Understanding the relationship between products or variables can help businesses or researchers predict behavior or trends.

Key Differences Between Association and Correlation Analysis

- **Association Analysis**:
  - Deals with discovering hidden relationships between item sets in a dataset.
  - Uses metrics like **support, confidence**, and **lift** to measure associations.
  - Typically used in retail, e-commerce, and marketing (e.g., Market Basket Analysis).
- **Correlation Analysis**:
  - Measures how strongly two continuous variables are related.
  - Uses metrics like the **correlation coefficient (Pearson's r)**.
  - Typically used in statistical analysis and regression modeling to predict linear relationships.

When to Use Association vs. Correlation Analysis

- **Use Association Analysis**:
  - When dealing with **categorical data** like transactions in a supermarket or survey responses.
  - To find patterns and relationships between different product sets.
  - Example: If customers buy bread, they are likely to buy butter.
- **Use Correlation Analysis**:
  - When dealing with **continuous data** to measure the strength of relationships.
  - To assess **linear relationships** between two variables (e.g., income and expenditure).
  - Example: How strongly are age and income related?

Understanding Correlation in Association Rules

Lift in Association Analysis:

Measures how much more likely two items are to be bought together compared to being purchased independently.

Fo**rmula: Lift(A → B) = Confidence(A → B) / Support(B).**

Interpretation:

If Lift > 1, A and B are positively correlated.

If Lift = 1, A and B are independent.

If Lift < 1, A and B are negatively correlated.

Correlation Coefficient:

Pearson's r: Measures the strength and direction of a linear relationship between two variables.

Values range from -1 to 1:

1: Perfect positive correlation.

0: No correlation.

-1: Perfect negative correlation.

Introduction to Weka

- **What is Weka?**:
  - Weka (Waikato Environment for Knowledge Analysis) is a popular open-source software that provides machine learning algorithms for data mining tasks.
  - It includes tools for data pre-processing, classification, regression, clustering, and association rule mining.
- **Why Use Weka for Apriori?**:
  - Provides a simple graphical interface for running data mining algorithms.
  - Allows experimenting with different datasets and algorithm parameters like support and confidence.

Exploring Weka: Loading the Supermarket Dataset

- **Steps**:
  - **Download Dataset**: Use the **Supermarket dataset** from the Weka sample datasets.
  - **Open Weka Explorer**: Navigate to the "Explorer" tab in the Weka GUI.
  - **Load Dataset**: Click on "Open File" and load the **Supermarket.arff** dataset.
  - **View Dataset**: Observe the transactions, where each row represents items bought by a customer.
- **Dataset Overview**: The supermarket dataset contains transactions from a retail store, where each transaction consists of multiple items.

Running the Apriori Algorithm in Weka

Steps to Run Apriori:

- **Select Apriori Algorithm:** Go to the "Associate" tab and choose the Apriori algorithm.
- **Set Parameters:** Set different values for support and confidence to control the frequency and strength of the generated rules.
- **Support:** Set as a percentage (e.g., 10%, 20%).
- **Confidence:** Set between 0.5 and 1 (50% to 100%).
- **Run the Algorithm:** Click "Start" to generate association rules.
- **Understanding the Output:** Weka displays frequent item sets and association rules with their respective support, confidence, and lift values.

Experimenting with Different Support and Confidence Values

- **Experiment 1**: Support = 20%, Confidence = 80%.
  - Higher support identifies only the most frequently purchased items.
  - Output: Rules generated for commonly bought items like "bread" and "butter."
- **Experiment 2**: Support = 10%, Confidence = 60%.
  - Lower support identifies less frequent but still important patterns.
  - Output: More rules are generated, including less frequent combinations.
- **Observation**: Decreasing support generates more rules, but some rules may be less actionable. Higher confidence ensures stronger associations.

Analysis of Results from Weka

- **Support and Confidence Trade-off**:
  - **Higher Support**: Results in fewer, more reliable rules, but may miss less frequent associations.
  - **Lower Support**: Uncovers more item combinations but may include noise or less important rules.
- **Best Practice**: Choose a **balance** between support and confidence to avoid generating too few or too many rules.
- **Application**: The rules can be used for recommendation engines, inventory optimization, and product placement strategies.

Advantages of Using Weka for Apriori

- **User-Friendly Interface**: Provides a GUI, making it easier to visualize and experiment with different parameters.
- **Visualization**: Weka allows users to visualize the dataset and the rules generated by Apriori.
- **Customizable**: Allows fine-tuning of algorithm parameters like minimum support, confidence, and maximum number of rules.
- **Efficient for Beginners**: A great tool for students and data analysts to explore association rule mining without requiring programming knowledge.

Conclusion

- **Association Analysis** helps identify relationships between items in large datasets, while **Correlation Analysis** measures the strength of linear relationships between variables.
- **Weka** offers a powerful platform for experimenting with the Apriori algorithm, allowing users to adjust support and confidence to uncover actionable patterns.
- With the **Supermarket dataset**, Weka helps uncover important product associations that can inform retail strategies.
- Finding the right balance between **support** and **confidence** is crucial to generating useful and reliable association rules.