Knowledge Discovery and Data Mining Basics

Knowledge Discovery

Knowledge Discovery is the process of identifying implicit, potentially useful, or previously unknown information from data. It involves multiple steps to transform raw data into actionable knowledge.

Knowledge Discovery Process

The process of knowledge discovery can be broken down into several steps:

1. **Data Integration:**
   - Data is collected from various sources.
   - It is integrated into a single data store, often referred to as the **target data**.
2. **Data Preprocessing:**
   - The integrated data is cleaned, filtered, and transformed into a standard format suitable for analysis.
3. **Data Mining:**
   - Specialized algorithms are applied to the processed data.
   - These algorithms identify patterns, relationships, or rules hidden within the data.
4. **Interpretation:**
   - The patterns and rules discovered are interpreted.
   - This interpretation transforms raw patterns into **useful knowledge** or **actionable information**.

Goal of Knowledge Discovery

The primary objective of the knowledge discovery process is to:

- Identify hidden patterns or trends within large datasets.
- Extract valuable information that can lead to informed decision-making.

Role of Data Mining in Knowledge Discovery

Data mining is the **core step** in the knowledge discovery process. It involves applying advanced algorithms and techniques to:

- Discover meaningful patterns and correlations.
- Provide insights that can be interpreted and utilized effectively.

What is Data Mining?

Data Mining is a process that transforms raw data into useful information by analyzing large datasets to discover patterns, trends, and insights. It enables businesses and industries to make data-driven decisions and improve operational efficiency.

Key Definitions of Data Mining

1. **Business Perspective:**

   Data mining involves the use of software to identify patterns in large data batches, helping businesses understand their customers, create better marketing strategies, boost sales, and reduce costs.

2. **Analytical Perspective:**

   It is the practice of exploring massive datasets to find new and hidden information that enhances business performance and decision-making.

3. **Technical Perspective:**

   Data mining relies on data collection, data warehousing, and computational techniques, utilizing advanced algorithms to uncover patterns and trends beyond simple analysis.

4. **Knowledge Discovery Perspective:**

   Data mining is often referred to as **Knowledge Discovery in Data (KDD)** because it focuses on identifying actionable insights from data.

Key Properties of Data Mining

- **Automatic Pattern Discovery:**
  Algorithms uncover hidden patterns without manual intervention.
- **Prediction of Likely Outcomes:**
  Helps forecast future trends and events based on historical data.
- **Actionable Information Creation:**
  Converts data insights into strategies and solutions for business improvement.
- **Large Dataset Focus:**
  Designed to handle and analyze massive data collections efficiently.
- **Answers Complex Questions:**
  Solves queries that traditional reporting and querying cannot address.

Data Mining Architecture

Data mining architecture defines how a data mining system is designed and interacts with other systems like databases or data warehouses. There are four primary architectures:

1. No-Coupling Architecture

- **Description:**
  The data mining system works independently, without utilizing the functionalities of databases or data warehouses.

- **How it Works:**
  - Data is retrieved from specific sources (e.g., file systems).
  - Processes data using algorithms.
  - Stores results in a file system.
- **Drawbacks:**
  - Does not benefit from the efficiency of databases or data warehouses.
  - Poor architecture for complex tasks.
- **Use Case:**
  Simple and standalone data mining tasks.

2. Loose Coupling Architecture

- **Description:**
  The data mining system uses a database or data warehouse for data retrieval but operates independently for processing.
- **How it Works:**
  - Data is retrieved from databases or warehouses.
  - Data mining algorithms process the data.
  - Results are stored back in the database or warehouse.
- **Features:**
  - Suitable for memory-based systems.
  - Limited scalability and performance.
- **Use Case:**
  Systems with moderate complexity where full integration is not necessary.

3. Semi-Tight Coupling Architecture

- **Description:**
  The system links more closely to databases or data warehouses, utilizing their advanced features.
- **How it Works:**
  - Leverages database/warehouse functionalities like sorting, indexing, and aggregation.
  - Intermediate results may be stored in the database for better performance.
- **Features:**
  - Enhanced efficiency due to partial integration.
  - Better performance than loose coupling.
- **Use Case:**
  Systems requiring moderate integration for improved performance.

4. Tight Coupling Architecture

- **Description:**
  Fully integrates the database or data warehouse with the data mining system for complete functionality.
- **How it Works:**
  - Database or warehouse is used for both information retrieval and data mining tasks.
  - All database features (e.g., scalability, performance) are utilized.
- **Features:**
  - High system scalability and performance.
  - Comprehensive integration ensures seamless operation.
- **Use Case:**
  Complex systems requiring robust, high-performance data mining.

Comparison Table

Feature	No-Coupling	Loose Coupling	Semi-Tight Coupling	Tight Coupling
Database Utilization	None	Partial	Moderate	Full
Efficiency	Low	Medium	High	Very High
Integration	None	Low	Medium	Full
Scalability	Low	Moderate	High	Very High
Use Case	Simple Tasks	Memory-Based Tasks	Intermediate Tasks	Complex Systems

Data Mining Applications

1. Sales/Marketing

- **Improving Campaigns:** Helps businesses analyze past purchase patterns to design cost-effective marketing campaigns.
- **Market Basket Analysis:** Identifies product combinations, purchase sequences, and timing to promote related or profitable products.
- **Customer Behavior Analysis:** Retailers analyze buying patterns to better understand customer preferences.

2. Banking/Finance

- **Fraud Detection:** Uses techniques like distributed data mining to detect credit card fraud.
- **Customer Loyalty:** Analyzes purchasing frequency, monetary value, and last purchase dates to measure loyalty.
- **Retention Strategies:** Predicts customers likely to leave and helps plan special offers to retain them.

- **Financial Analysis:** Identifies spending patterns and correlations between financial indicators.
- **Stock Market Insights:** Develops trading rules from historical data.

3. Health Care and Insurance

- **Claims Analysis:** Identifies trends in claimed medical procedures.
- **Forecasting Purchases:** Predicts which customers might buy new policies.
- **Risk Detection:** Discovers risky behavior patterns among customers.

4. Medicine

- **Patient Monitoring:** Characterizes patient activities to forecast office visits.
- **Therapy Patterns:** Identifies successful treatment methods for various illnesses.

Advantages of Data Mining

1. Marketing/Retail

- Predicts customer responses to campaigns using historical data models.

2. Finance/Banking

- Evaluates loan risks and detects fraudulent credit card transactions.

3. Manufacturing

- Identifies faulty equipment and optimizes control parameters.

4. Government

- Analyzes financial records to detect money laundering or criminal activity.

Disadvantages of Data Mining

1. Privacy Issues

- Personal data can be collected and misused, raising ethical concerns.

2. Security Issues

- Sensitive information (e.g., social security numbers, payroll) may be vulnerable to hacking, leading to identity theft or fraud.

3. Misuse of Information

- Data meant for marketing can be exploited to discriminate or harm vulnerable groups.

Data Mining Process (CRISP-DM)

The **Cross-Industry Standard Process for Data Mining (CRISP-DM)** is a reliable, repeatable framework developed for effective data mining across industries. It consists of **six cyclical phases**:

1. Business Understanding

- **Define Objectives:** Identify clear business goals and understand client expectations.
- **Assess Resources:** Evaluate the current resources, constraints, and assumptions.
- **Set Goals:** Align data mining goals with business objectives.
- **Create a Plan:** Develop a detailed strategy to achieve these objectives.

2. Data Understanding

- **Data Collection:** Gather data from various sources.
- **Data Exploration:** Examine surface-level data properties through querying, reporting, and visualization.
- **Data Quality Check:** Ensure completeness and identify missing or inconsistent data.

3. Data Preparation

- **Data Selection:** Choose relevant data sources.
- **Data Cleaning:** Handle missing values, outliers, and inconsistencies.
- **Data Transformation:** Format and construct data for analysis.

4. Modeling

- **Technique Selection:** Choose appropriate modeling techniques.
- **Model Creation:** Train models using the prepared dataset.
- **Model Validation:** Test models to ensure quality and relevance to business goals.
- **Stakeholder Involvement:** Collaborate with stakeholders for model assessment.

5. Evaluation

- **Result Analysis:** Assess model results against business objectives.
- **Raise New Requirements:** Identify any additional needs based on discovered patterns or business factors.

6. Deployment

- **Present Insights:** Deliver findings in actionable formats, such as reports or dashboards.
- **Integration:** Implement repeatable processes for ongoing use across the organization.

Advantages of CRISP-DM

1. **Uniform Framework:** Provides consistent documentation and guidelines.
2. **Flexibility:** Applicable across various industries and data types.

Data Mining Techniques

Several **data mining techniques** are widely used to analyze and interpret data in projects. These techniques help extract meaningful patterns and insights. Below are the four major techniques:

1. Association

- **Definition:** Identifies patterns by analyzing the relationship between items in a transaction.
- **Example Use Case:**
  - In **market basket analysis**, businesses discover which products are frequently bought together (e.g., "bread and butter").
  - Insight: Use this data to design targeted marketing strategies, such as product bundling or discounts.

2. Classification

- **Definition:** Classifies items in a dataset into predefined groups or categories.
- **Methods Used:** Decision trees, neural networks, linear programming, and statistical methods.
- **Example Use Case:**
  - Predict whether employees will "stay" or "leave" the company based on historical data.
  - Insight: Helps in workforce management and planning by identifying patterns in employee behavior.

3. Clustering

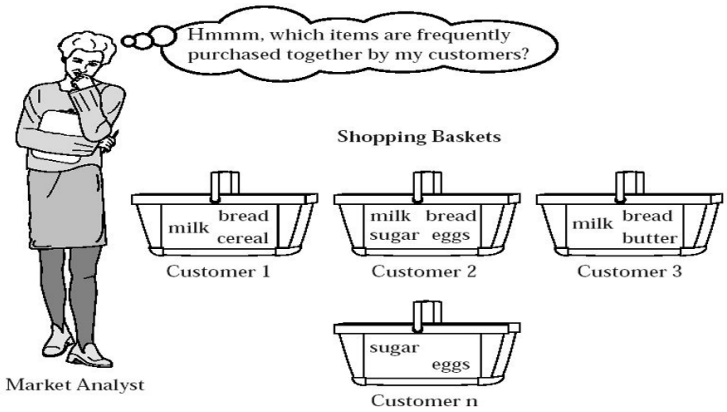
- **Definition:** Groups objects into clusters based on shared characteristics without predefined categories.
- **Difference from Classification:** Clustering defines classes automatically, whereas classification uses predefined groups.
- **Example Use Case:**
  - Organizing books in a library into clusters by topic (e.g., Science, Fiction) to improve accessibility for readers.
  - Insight: Simplifies navigation by grouping similar items together.

4. Prediction

- **Definition:** Identifies relationships between dependent and independent variables to predict future outcomes.
- **Example Use Case:**
  - Forecasting profits based on sales data, where sales are independent variables and profit is the dependent variable.
  - Insight: Enables strategic decision-making by predicting future trends.

Association Rule Mining

Association rule mining is a data mining technique used to discover interesting relationships, patterns, or associations between items in large datasets. It is commonly used in transactional data to find frequent itemsets and generate rules that explain the relationships.



Key Concepts in Association Rule Mining

1. **Support:**
   - Measures the frequency of an itemset in the dataset.
   - Formula:

$$Support(A) = \frac{\text{Transactions containing A}}{\text{Total transactions}}$$

   - Example: If "bread" appears in 30 out of 100 transactions, the support is 30%.
2. **Confidence:**
   - Indicates the likelihood that an item B is purchased when item A is purchased.
   - Formula:

$$Confidence(A \rightarrow B) = \frac{\text{Support}(\text{A} \cup \text{B})}{\text{Support}(\text{A})}$$

   - Example: If 20 out of 30 transactions containing "bread" also include "butter," the confidence is 20 / 30 = 66.7%

Process of Association Rule Mining

1. **Frequent Itemset Generation:**
   - Identify item combinations that meet the minimum support threshold.

- Common algorithm: **Apriori**.

2. **Rule Generation:**
   - Generate association rules from the frequent itemsets that satisfy minimum confidence thresholds.

Example:

Q. A database has four transactions.

TID	Items-Bought
T100	{A, B, D, K}
T200	{A, B, C, D, E}
T300	{A, B, C, E}
T400	{A, B, D}

Find all frequent itemsets using Apriori algorithm with min_sup=3, i.e., any itemset occurring in less than 3 transactions is considered to be infrequent. Also list all the strong association rules with min_sup=3 and min_conf=80%.

TID	Items-Bought
T100	{A, B, D, K}
T200	{A, B, C, D, E}
T300	{A, B, C, E}
T400	{A, B, D}

First Scan (1-itemsets):

ItemSet	Sup Count
A	4
B	4
C	2
D	3
E	2
K	1

Items count ≥ 3:

{A, B, D}

Second Scan (2-itemsets)

Itemset	Sup Count
A, B	4
A, D	3
B, D	3

Items count ≥ 3

{A, B, A, D, B, D}

Third Scan (3-itemsets)

Item Set	Sup Count
A, B, D	3

Items count ≥ 3

{A, B, D}

Derive Strong Association Rules:

1. $A \rightarrow B$:

Support(A, B): $\frac{4}{4} = 1 = 100\%$

Confidence: $\frac{\text{Support}(A,B)}{\text{Support}(A)} = \frac{4}{4} = 100\%$

2. $B \rightarrow A$:

Support: $\frac{4}{4} = 1 = 100\%$

Confidence: $\frac{\text{Support}(B,A)}{\text{Support}(B)} = \frac{4}{4} = 100\%$

3. $A \rightarrow D$:

Support: $\frac{3}{4} = 75\%$

Confidence: $\frac{3}{4} = 75\%$

4. $D \rightarrow A$

Support: $\frac{3}{4} = 75\%$

Confidence: $\frac{3}{3} = 100\%$

5. $B \rightarrow D$

Support: $\frac{3}{4} = 75\%$

confidence: $\frac{3}{4} = 75\%$

6. $D \rightarrow B$

Support: $\frac{3}{4} = 75\%$

Confidence: $\frac{3}{3} = 100\%$

7. $AB \rightarrow D$:

Support: $\frac{3}{4} = 75\%$

Confidence: $\frac{3}{4} = 75\%$

8. $D \rightarrow AB$:

Support: $\frac{3}{4} = 75\%$

Confidence: $\frac{3}{3} = 100\%$

9. $AD \rightarrow B$:

Support: $\frac{3}{4} = 75\%$

Confidence: $\frac{3}{3} = 100\%$

10. $B \rightarrow AD$:

Support: 75%

Confidence: $\frac{3}{4} = 75\%$

11. $BD \rightarrow A$:

Support: $\frac{3}{4} = 75\%$

Confidence: $\frac{3}{3} = 100\%$

12. $A \rightarrow BD$

Support: 75%

Confidence: $\frac{3}{4} = 75\%$

As, min_conf = 80%

So, Strong Rules are:

$F = \{A \rightarrow B, B \rightarrow A, D \rightarrow A, D \rightarrow B, D \rightarrow AB, AD \rightarrow B, BD \rightarrow A\}$

b) Suppose you have market basket data consisting of 100 transactions and 20 items. If the support for item a is 25%, the support for item b is 90% and the support for itemset {a, b} is 20%. Let the support and confidence thresholds be 10% and 60%, respectively. Compute the confidence of the association rule {a} → {b}. Is the rule interesting according to the confidence measure?

Ans: Confidence= support of {a,b}/support of {a} = 20%/25% = 80%
Rule is also interesting because confidence is greater than 60%.