Al-Mustaqbal University
College of Sciences
Intelligent Medical System Department

جامعة المستقبل
AL MUSTAQBAL UNIVERSITY

كلية العلوم
قسم الانظمة الطبية الذكية

Lecture: (4)

Classification of Data Mining Systems & KDD

Subject: Clinical Data Mining

Level: Four

Lecturer: Dr. Maytham Nabeel Meqdad

Classification of Data Mining Systems & KDD

What are KDD and Data Mining?

1. KDD (Knowledge Discovery in Databases)

KDD is the **overall process** used to extract useful knowledge from large volumes of data.
- It consists of a series of sequential steps.

2. Data Mining

Data Mining is the **main step within KDD**, where algorithms are applied to extract patterns and relationships from data.

- **In brief:**

- KDD = the complete process
- Data Mining = a step within this process

KDD Process Steps:

1. **Data Cleaning** → Removing errors and handling missing values.
2. **Data Integration** → Combining data from multiple sources.
3. **Data Selection** → Choosing data relevant to the problem at hand.
4. **Data Transformation** → Transforming data into a suitable format for analysis (e.g., normalization or summarization).
5. **Data Mining** → Applying algorithms to extract patterns and knowledge.
6. **Pattern Evaluation** → Evaluating extracted patterns to select useful knowledge.
7. **Knowledge Presentation** → Presenting results using visualization techniques.

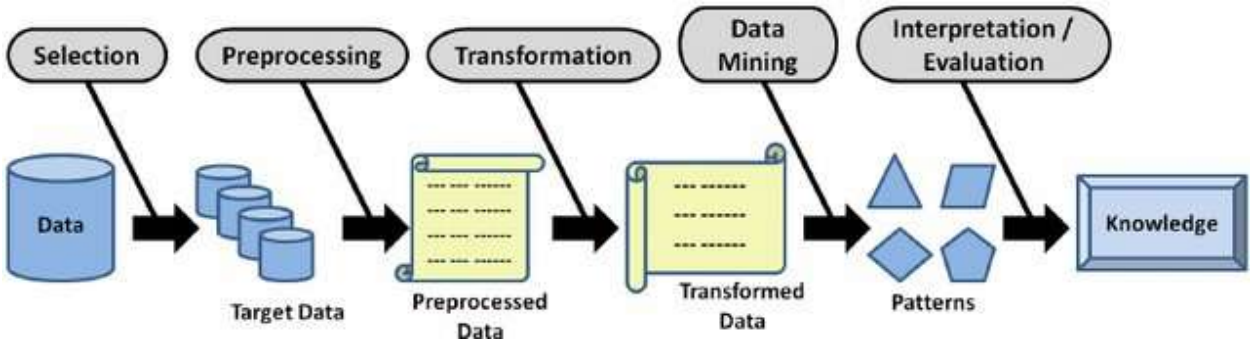


Fig. 1. The Knowledge Discovery in Databases (KDD) process (DOI: 10.1016/j.phpro.2015.02.005)

Classification of Data Mining Systems

Data mining systems can be classified in several ways depending on the perspective:

1. Classification according to the kinds of databases mined

- Relational Databases (e.g., SQL)
- Data Warehouses
- Transactional Databases
- Object-Oriented Databases
- Spatial / Temporal Databases
- Multimedia Databases (audio, images, video)
- Text Databases
- Web Databases

2. Classification according to the kinds of knowledge mined

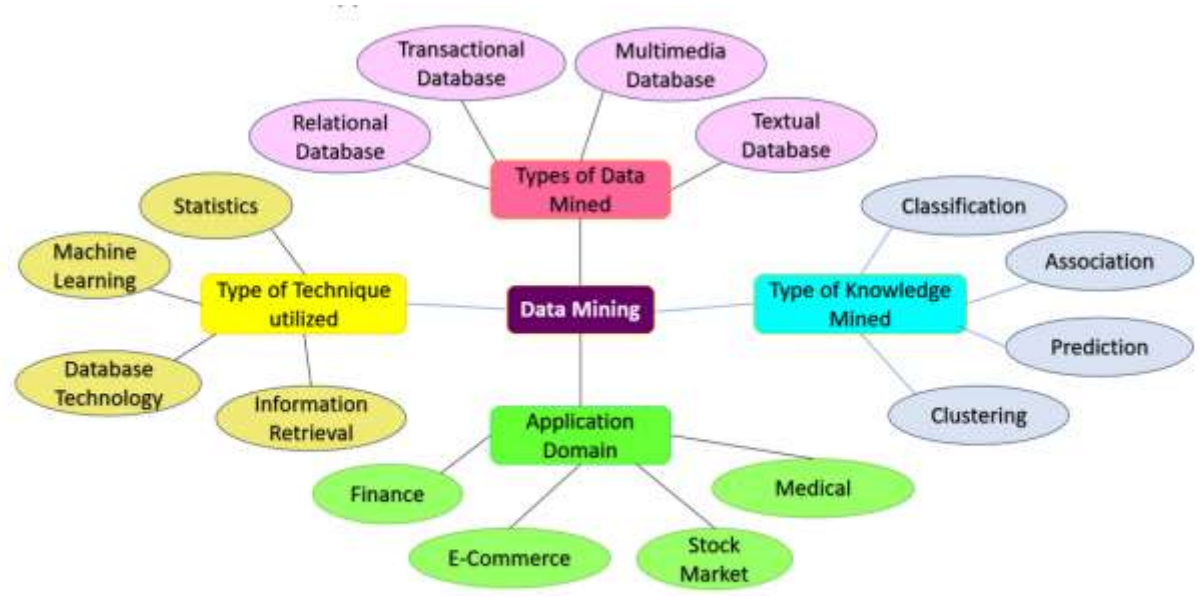
- **Association Rules** → Discovering relationships between items (e.g., market basket analysis).
- **Classification** → Categorizing data into labels or classes.
- **Clustering** → Grouping similar data without predefined classes.
- **Prediction** → Forecasting future values.
- **Outlier Analysis** → Detecting abnormal or rare data.
- **Evolution Analysis** → Analyzing changes over time.

3. Classification according to the kinds of techniques used

- **Machine Learning** (Decision Trees, Neural Networks, SVM)
- **Statistical Methods** (Regression, Probabilistic models)
- **Database-oriented techniques** (SQL queries)
- **Visualization** (Visual representation of patterns)

4. Classification according to the applications adapted

- **Business Analysis**
- **Marketing**
- **Medical Diagnosis**
- **Banking & Insurance**
- **Manufacturing**
- **Government & Security**



Classification based on Types of Data Mined

==========

Summary of the difference between KDD and Data Mining

Aspect	KDD	Data Mining
Definition	A comprehensive process to discover knowledge	A step within KDD to extract patterns
Steps	Cleaning, Integration, Selection, Transformation, Mining, Evaluation, Presentation	Applying algorithms only
Focus	The complete process	Data analysis

Case Study: Medical Data Analysis Using KDD at Al-Mustaqbal University Hospital

1. Background

Al-Mustaqbal University Hospital collects extensive medical data daily from a large number of patients, including:

- Demographic information (age, gender, residence)
- Vital signs (blood pressure, blood sugar, heart rate, etc.)
- Laboratory test results (Blood Tests, Cholesterol, Uric acid, etc.)
- Previous medical diagnoses
- Patient visit patterns and family disease history

Objectives:

- Early detection of chronic diseases (e.g., diabetes, heart disease, gout, hypertension)
- Support medical decision-making
- Reduce healthcare costs and improve quality of care

2. KDD Process Steps Applied

2.1 Data Cleaning

- Remove duplicate or incomplete records.
- Handle missing values using mean imputation or previous patient records.
- Standardize units (e.g., convert all blood sugar readings to mg/dL).

2.2 Data Integration

- Merge patient data from multiple departments (emergency, clinics, lab) into a unified data warehouse.

2.3 Data Selection

- Select data for a specific age group (e.g., 30–60 years).
- Focus on lab tests relevant to chronic diseases (blood sugar, cholesterol, blood pressure, BMI).

2.4 Data Transformation

- Convert numerical values into categories (e.g., blood pressure → normal / high / low).
- Create new features (e.g., BMI calculated from height and weight).

- Normalize values for compatibility with data mining algorithms.

3. Data Mining Techniques

3.1 Classification

- Apply Decision Tree or Random Forest to classify patients into:
  - Healthy
  - At Risk
  - Diagnosed with a chronic disease
- Purpose: Early prediction of at-risk patients.

3.2 Clustering

- Use K-Means to group patients based on health patterns:
  - Cluster 1: Obese + high blood sugar + high blood pressure → high risk
  - Cluster 2: Mostly healthy patients → low risk
  - Cluster 3: Elderly patients with intermediate values → medium risk

3.3 Association Rule Mining

- Apply Apriori algorithm to discover hidden relationships among lab tests.
- Example: Patients with high BMI + high cholesterol have a 70% chance of high blood pressure.
- Supports preventive decision-making by doctors.

4. Pattern Evaluation & Presentation

- Evaluate classification performance using Accuracy, Precision, and Recall.
- Select rules with high support and confidence.
- Present results via dashboards and visual charts for easy interpretation by medical staff.

5. Expected Results

- Reduce diagnosis time by 20–30%.
- Early detection of hundreds of cases before disease progression.
- Support intelligent medical decision-making.
- Develop a prototype predictive system linked with Electronic Health Records (EHR).

Conclusion

This case study demonstrates how students in Intelligent Medical Systems at Al-Mustaqbal University can apply KDD and Data Mining techniques to analyze large-scale medical data, enhance decision-making, and improve overall healthcare quality.

```
# Import libraries
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.tree import DecisionTreeClassifier
from sklearn.metrics import classification_report, confusion_matrix
from sklearn.cluster import KMeans
from mlxtend.frequent_patterns import apriori, association_rules

# ------------------------------
# 1. Load Sample Data (Simulated Medical Records)
# ------------------------------
data = {
    'Age': [25, 45, 50, 35, 60, 40, 30, 55],
    'BMI': [22, 30, 28, 25, 33, 27, 24, 31],
    'BloodSugar': [90, 150, 140, 100, 160, 130, 95, 155],
    'Cholesterol': [180, 240, 230, 190, 250, 220, 185, 245],
    'BloodPressure': [120, 150, 145, 125, 155, 135, 118, 150],
    'Disease': ['Healthy', 'Diabetes', 'Diabetes', 'Healthy', 'Hypertension', 'Pre-Diabetes', 'Healthy', 'Diabetes']
}

df = pd.DataFrame(data)

# ------------------------------
# 2. Data Cleaning (Handling missing values)
# ------------------------------
df.fillna(df.mean(numeric_only=True), inplace=True)

# ------------------------------
# 3. Data Transformation (Scaling features)
# ------------------------------
features = ['Age', 'BMI', 'BloodSugar', 'Cholesterol', 'BloodPressure']
X = df[features]
y = df['Disease']

scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)

# ------------------------------
# 4. Classification (Decision Tree)
# ------------------------------
X_train, X_test, y_train, y_test = train_test_split(X_scaled, y, test_size=0.3, random_state=42)

clf = DecisionTreeClassifier()
clf.fit(X_train, y_train)
y_pred = clf.predict(X_test)

print("Classification Report:")
print(classification_report(y_test, y_pred))
print("Confusion Matrix:")
print(confusion_matrix(y_test, y_pred))

# ------------------------------
# 5. Clustering (K-Means)
# ------------------------------
kmeans = KMeans(n_clusters=3, random_state=42)
clusters = kmeans.fit_predict(X_scaled)
df['Cluster'] = clusters
print("\nPatient Clusters:")
print(df[['Age', 'BMI', 'BloodSugar', 'Cluster']])

# ------------------------------
# 6. Association Rule Mining (Apriori)
# ------------------------------
```

```
# Simulate categorical data for Apriori
df_apriori = df.copy()
df_apriori['HighBMI'] = df_apriori['BMI'] > 28
df_apriori['HighSugar'] = df_apriori['BloodSugar'] > 140
df_apriori['HighCholesterol'] = df_apriori['Cholesterol'] > 220

apriori_df = df_apriori[['HighBMI','HighSugar','HighCholesterol']]
apriori_df = apriori_df.astype(int)

frequent_itemsets = apriori(apriori_df, min_support=0.3, use_colnames=True)
rules = association_rules(frequent_itemsets, metric="confidence", min_threshold=0.7)

print("\nAssociation Rules:")
print(rules)
```

Case Study: Customer Purchase Behavior Analysis Using KDD in a Retail Company

1. Background

A large retail company collects daily transactional data from thousands of customers, including:

- Customer Demographics: Age, Gender, Location
- Purchase History: Items bought, quantity, price
- Payment Method: Cash, Credit Card, Online Payment
- Visit Patterns: Time and day of purchase, frequency of visits

Objectives:

- Understand customer buying behavior.
- Identify high-value customers for loyalty programs.
- Optimize marketing campaigns and promotions.
- Increase sales and customer retention.

2. KDD Process Steps Applied

2.1 Data Cleaning

- Remove duplicate transactions and invalid records.
- Handle missing values in customer profiles or transaction details.

2.2 Data Integration

- Merge data from multiple store branches and online sales platforms into a centralized data warehouse.

2.3 Data Selection

- Select relevant features for analysis: Customer Age, Gender, Total Purchase Amount, Purchase Frequency, Product Categories.

2.4 Data Transformation

- Convert numerical purchase amounts into categories (Low, Medium, High).
- Encode categorical features (Payment Method, Product Category) for analysis.

- Aggregate data by customer to calculate total spending and frequency.

3. Data Mining Techniques

3.1 Classification

- Use Decision Tree or Random Forest to classify customers into segments:
  - High-Value Customers
  - Medium-Value Customers
  - Low-Value Customers

3.2 Clustering

- Apply K-Means to group customers based on buying behavior:
  - Cluster 1: Frequent buyers of high-value items → High Loyalty
  - Cluster 2: Occasional buyers with low spending → Low Loyalty
  - Cluster 3: Medium frequency and spending → Medium Loyalty

3.3 Association Rule Mining

- Use Apriori to discover relationships between purchased items:
  - Example: Customers who buy bread often also buy milk (Market Basket Analysis).
- Supports targeted promotions and cross-selling strategies.

4. Pattern Evaluation & Presentation

- Evaluate classification performance with Accuracy, Precision, Recall.
- Filter association rules with high support and confidence.
- Present insights via dashboards and visualizations for marketing teams.

5. Expected Outcomes

- Increased sales through personalized promotions.
- Better understanding of customer preferences and buying patterns.
- Improved customer retention and loyalty program effectiveness.
- Data-driven marketing and inventory planning.

Conclusion

This case demonstrates how KDD and Data Mining techniques can help businesses analyze transactional data, segment customers, uncover hidden patterns, and optimize marketing strategies for better profitability and customer satisfaction.

References

[1] Han and M. Kamber, " Data Mining Tools and Technique s", Morgan Kaufmann Publishers.

[2] .M.H. Dunham, " Data Mining Introductory and Adv anced Topics", Pear son Education