Chapter 1: Introduction to Data Mining

CSE 435:Data Mining

GREEN UNIVERSITY OF BANGLADESH

Md. Atikuzzaman
Lecturer
Department of Computer Science & Engineering
Green University of Bangladesh
atik@cse.green.edu.bd

Table of Contents

1. What Is Data Mining?
2. The Knowledge Discovery Process
3. The Knowledge Discovery Process
4. Diversity of Data Types
5. Mining Various Kinds of Knowledge
6. Confluence of Disciplines
7. Applications
8. Data Mining and Society
9. Summary

What Is Data Mining?

The Modern Data Landscape

We live in a world where vast amounts of data are generated constantly and rapidly.

Definition (Data Mining)

Data mining is the process of discovering interesting patterns and knowledge in large data sets.

A More Accurate Term

The term ”Data mining” is often considered a misnomer. A more fitting name would be **”knowledge mining from data”**. Other terms include KDD (Knowledge Discovery from Data) and data analytics.

Example: Turning Data into Knowledge

Google Flu Trends

- Google found a close relationship between the number of people who search for flu-related information and the number of people who actually have flu symptoms.
- This allowed them to estimate flu activity up to two weeks faster than traditional systems.

Data mining is a young, dynamic, and promising field.

Data Mining: A Step in Knowledge Discovery

Data mining is an essential step in the **Knowledge Discovery from Data (KDD)** process.

1. Data cleaning: Remove noise and inconsistencies.
2. Data integration: Combine multiple data sources.
3. Data selection: Retrieve relevant data.
4. Data transformation: Convert data into a suitable format.
5. **Data mining**: Apply intelligent methods to extract patterns.
6. Pattern evaluation: Identify truly interesting patterns.
7. Knowledge presentation: Visualize and present the knowledge.

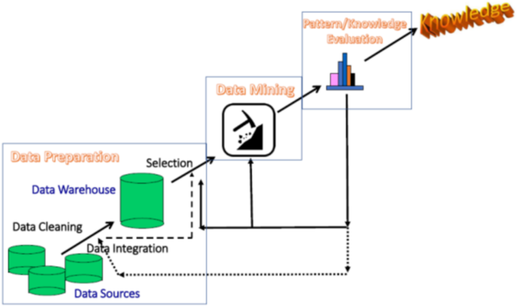


Figure: The Knowledge Discovery Process

Data Types for Mining (Slide 1 of 2)

Structured Data

Has a uniform, table-like structure with a fixed set of attributes. Often found in relational databases and data warehouses.

Semi-structured Data

Allows for more flexible or nested structures. Examples include transactional data, sequence data (like time-series or weblogs), and graph or network data.

Data Types for Mining (Slide 2 of 2)

Unstructured Data

Includes text data and multimedia content such as audio, images, and video.

Real-World Data

Data in the real world is often a complex mixture of these different types.

Streaming Data

Data can also arrive as a continuous, dynamic stream (e.g., from video surveillance), which poses challenges for real-time analysis.

Multidimensional Data Summarization

This involves generalizing, summarizing, and contrasting data characteristics, often using data cube technology.

- Utilizes Online Analytical Processing (OLAP).
- Requires scalable methods for computing multidimensional aggregates efficiently.

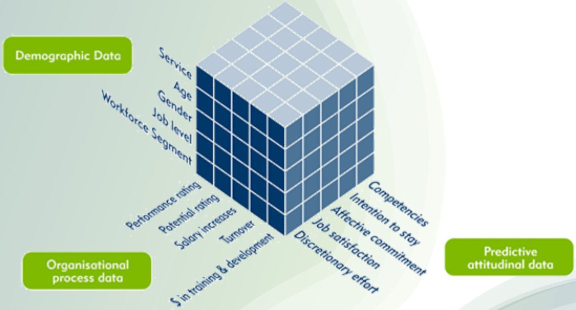


Figure: A data cube for multidimensional analysis.

Frequent Patterns and Associations

This task finds items that frequently co-occur in your data.

Typical Association Rule

An example rule from market basket analysis could be:

$$\text{Diaper} \rightarrow \text{Beer}\ [\text{support} = 0.5\%, \text{confidence} = 75\%]$$

This suggests that customers who buy diapers often buy beer as well.

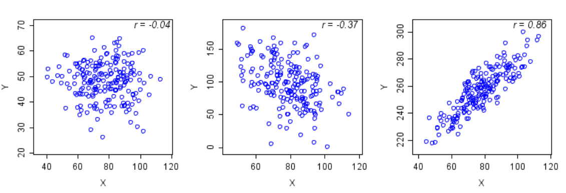


Figure: Scatter plots showing different correlations.

Classification and Regression (Slide 1 of 2)

Definition (Predictive Analysis)

Classification and regression are forms of predictive analysis. They build models from training data to predict unknown class labels or continuous values.

Common Methods

- Decision trees
- Naïve Bayesian classification
- Support vector machines (SVM)
- Neural networks
- Logistic regression

Classification and Regression (Slide 2 of 2)

Typical Applications

- Credit card fraud detection
- Direct marketing campaign targeting
- Classifying medical diseases
- Spam email filtering

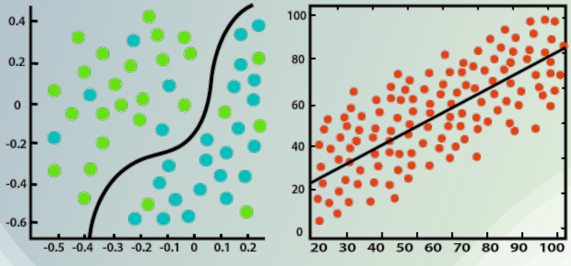


Figure: Regression vs Classification

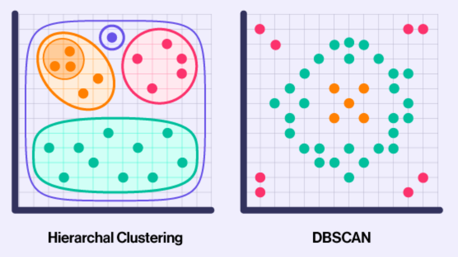
Cluster Analysis

Definition (Unsupervised Learning)

Cluster analysis groups data to form new categories (clusters) without any prior knowledge of class labels.

Core Principle

Maximize the similarity of objects within the same class and minimize the similarity between objects in different classes.



Deep Learning

A rapidly expanding field that uses various neural network architectures (CNNs, RNNs, Transformers) for tasks like classification, clustering, and outlier detection.

It has broad applications in:

- Computer vision
- Natural language processing
- Social network analysis
- Bioinformatics

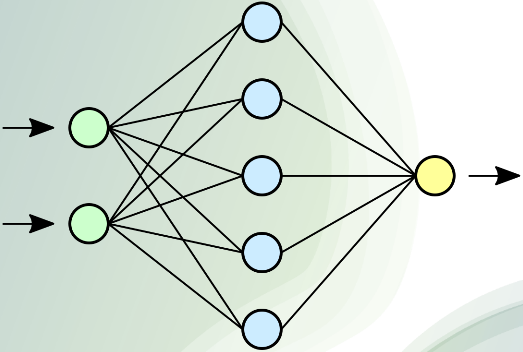


Figure: A neural network architecture.

Outlier Analysis

Definition (Outlier)

An outlier is a data object that deviates significantly from the general behavior of the data. It can be noise or a valuable discovery.

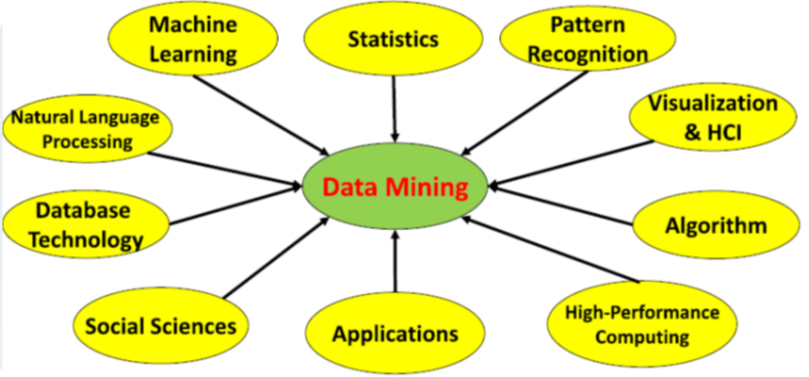
Applications

Outlier analysis is highly useful in fraud detection, network intrusion detection, and the analysis of rare events.

One person's noise could be another's treasure.

Data Mining: An Interdisciplinary Field

Data mining is a confluence of multiple disciplines, drawing from statistics, machine learning, database systems, and visualization.



This interdisciplinary nature is necessary to handle the scale, high-dimensionality, and complexity of modern data.

Data Mining Applications

- Web page analysis, ranking, and recommender systems.
- Market basket analysis for targeted marketing.
- Biological and medical data analysis.
- Software engineering and text analysis.
- Social and information network analysis.

Major Tools and Systems

SAS, Microsoft SQL Server Analysis Manager, Oracle Data Mining Tools.

Data Mining and Society

Benefits

Data mining can help scientific discovery, improve business management, and enhance security (e.g., cyberattack discovery).

Risks and Concerns

We must guard against the misuse of data mining, as it poses risks of unintentionally disclosing confidential business or personal information.

The Path Forward

The goal is to preserve data security and privacy while still performing successful data mining. Research in **privacy-preserving data mining** is a crucial, ongoing theme.

Chapter Summary

- **Data mining** is the process of discovering interesting knowledge from massive amounts of data.
- It is a crucial step in the **KDD process**, which also includes data preparation and knowledge presentation.
- Key **functionalities** include summarization, classification, clustering, deep learning, and outlier analysis.
- It is a **confluence of multiple disciplines** and has broad applications across many industries.
- It is important to promote **secure and ethical data mining** practices to benefit society while protecting privacy.

References

[1] Jiawei Han, Micheline Kamber, & Jian Pei, *Data Mining: Concepts and Techniques*, 4th Edition, Morgan Kaufmann, 2012.

[2] David J. Hand, Heikki Mannila, & Padhraic Smyth, *Principles of Data Mining*, First Edition, A Bradford Book, 2001.

[3] Richard O. Duda, Peter E. Hart, & David G. Stork, *Pattern Classification*, 2nd Edition, Wiley, 2001.