INTERNATIONAL UNIVERSITY - HCMC
School of Computer Science & Engineering

Lecture 1: Introduction to Data Mining

Lecturer: Dr. Nguyen, Thi Thanh Sang
(nttsang@hcmiu.edu.vn)

References:

[1] Chapter 1 in *Data Mining: Concepts and Techniques* (4th Edition), by Jiawei Han, et.al.
[2] Chapter 1 in *Data Mining: Practical Machine Learning Tools and Techniques* (4th Edition), by Ian H.Witten, et.al.

1

Introduction

- What is data mining?
- Data Mining Goals
- Stages of the Data Mining Process
- Data Mining Techniques
- Knowledge Representation Methods
- Applications
- Example: weather data

2

What is data mining?

- Example 1: *Web usage mining*
  - Given: click streams
  - Problem: prediction of user behaviour
  - Data: historical records of embryos and outcome
- Example 2: cow culling
  - Given: cows described by 700 features
  - Problem: selection of cows that should be culled
  - Data: historical records and farmers' decisions

3

What is data mining?

- Extracting
  - implicit,
  - previously unknown,
  - potentially useful

  information from data
- Needed: programs that detect patterns and regularities in the data
- Strong patterns → good predictions
  - Problem 1: most patterns are not interesting
  - Problem 2: patterns may be inexact (or spurious)
  - Problem 3: data may be garbled or missing

4

What Is Data Mining?

- Data mining (knowledge discovery from data)
  - Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data
  - Data mining: a misnomer?
- Alternative names
  - Knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence, etc.
- Watch out: Is everything "data mining"?
  - Simple search and query processing
  - (Deductive) expert systems

5

What is data mining?

Definitions:

- DM: The practice of examining large databases in order to generate new information.
- DM: The process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cut costs, or both.

6

What is data mining?

Data mining is defined as the process of discovering patterns in data.

- The process must be automatic or (more usually) semiautomatic.
- The patterns discovered must be meaningful in that they lead to some advantage, usually an economic one.
- The data is invariably presented in substantial quantities.

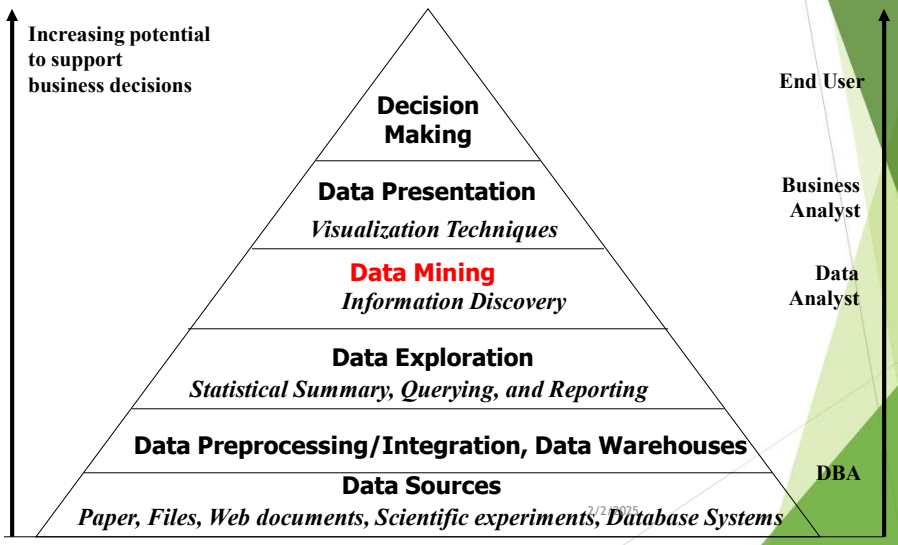
7

Introduction

- What is data mining?
- Data Mining Goals
- Stages of the Data Mining Process
- Data Mining Techniques
- Knowledge Representation Methods
- Applications
- Example: weather data

8

Data Mining in Business Intelligence



Data Mining Goals

- Extract information from a data set
- Transform it into an understandable structure for further use
- The ultimate **goal** of **data mining** is prediction

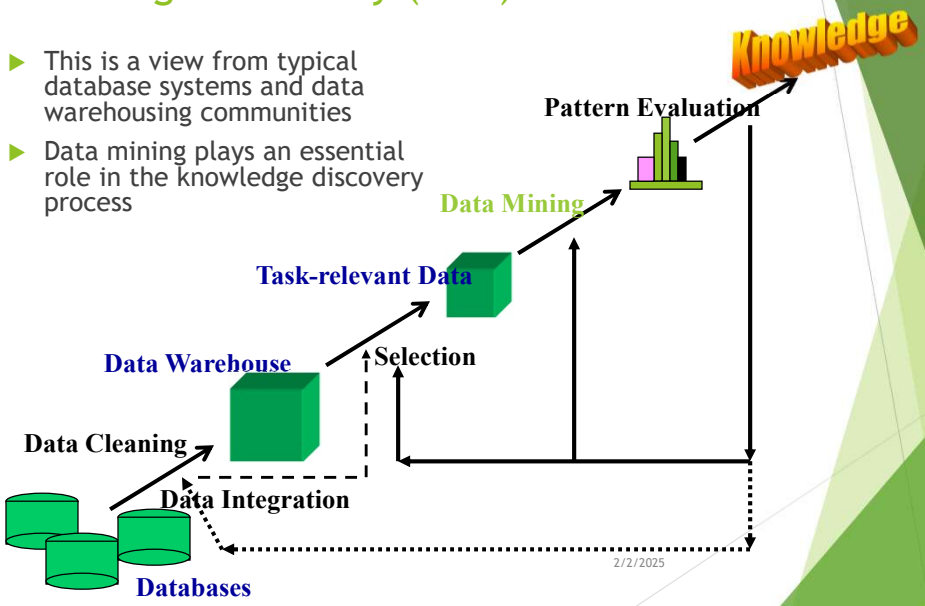
Introduction

- What is data mining?
- Data Mining Goals
- Stages of the Data Mining Process
- Data Mining Techniques
- Knowledge Representation Methods
- Applications
- Example: weather data

11

Knowledge Discovery (KDD) Process

- This is a view from typical database systems and data warehousing communities
- Data mining plays an essential role in the knowledge discovery process

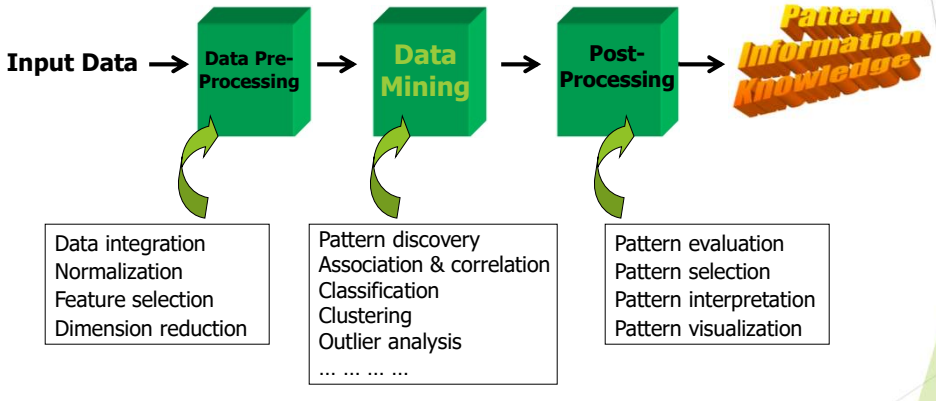


12

Example: A Web Mining Framework

- Web mining usually involves
  - Data cleaning
  - Data integration from multiple sources
  - Warehousing the data
  - Data cube construction
  - Data selection for data mining
  - Data mining
  - Presentation of the mining results
  - Patterns and knowledge to be used or stored into knowledge-base

KDD Process: A Typical View from ML and Statistics



- This is a view from typical machine learning and statistics communities

Which View Do You Prefer?

- Which view do you prefer?
  - KDD vs. ML/Stat. vs. Business Intelligence
  - Depending on the data, applications, and your focus
- Data Mining vs. Data Exploration
  - Business intelligence view
    - Warehouse, data cube, reporting but not much mining
  - Business objects vs. data mining tools
  - Supply chain example: mining vs. OLAP vs. presentation tools
  - Data presentation vs. data exploration

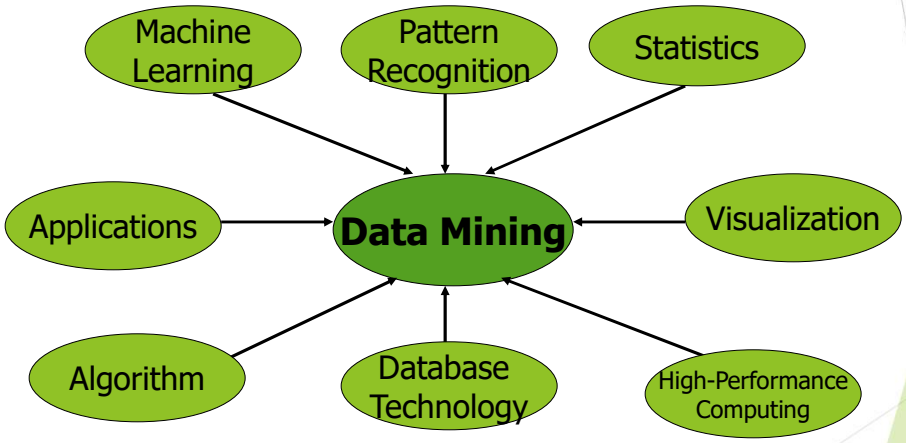
15

Introduction

- What is data mining?
- Data Mining Goals
- Stages of the Data Mining Process
- Data Mining Techniques
- Knowledge Representation Methods
- Applications
- Example: weather data

16

Data Mining: Confluence of Multiple Disciplines



Why Confluence of Multiple Disciplines?

- Tremendous amount of data
  - Algorithms must be scalable to handle big data
- High-dimensionality of data
  - Micro-array may have tens of thousands of dimensions
- High complexity of data
  - Data streams and sensor data
  - Time-series data, temporal data, sequence data
  - Structure data, graphs, social and information networks
  - Spatial, spatiotemporal, multimedia, text and Web data
  - Software programs, scientific simulations
- New and sophisticated applications

Machine learning techniques

- *Algorithms for acquiring structural descriptions from examples*
- Structural descriptions represent patterns explicitly
  - Can be used to predict outcome in new situation
  - Can be used to understand and explain how prediction is derived
    (*may be even more important*)
- Methods originate from artificial intelligence, statistics, and research on databases

Structural descriptions

- Example: if-then rules

```
If tear production rate = reduced
   then recommendation = none
Otherwise, if age = young and astigmatic = no
   then recommendation = soft
```

Age	Spectacle prescription	Astigmatism	Tear production rate	Recommended lenses
Young	Myope	No	Reduced	None
Young	Hypermetrope	No	Normal	Soft
Pre-presbyopic	Hypermetrope	No	Reduced	None
Presbyopic	Myope	Yes	Normal	Hard
…	…	…	…	…

Can machines really learn?

- Definitions of "learning" from dictionary:

  To get knowledge of by study, experience, or being taught — Difficult to measure

  To become aware by information or from observation
  To commit to memory
  To be informed of, ascertain; to receive instruction — Trivial for computers

- Operational definition:

  Things learn when they change their behavior in a way that makes them perform better in the future. — Does a slipper learn?

- Does learning imply intention?

21

Introduction

- What is data mining?
- Data Mining Goals
- Stages of the Data Mining Process
- Data Mining Techniques
- Knowledge Representation Methods
- Applications
- Example: weather data

22

Knowledge Representation Methods

- Tables
- Data cube
- Linear models
- Trees
- Rules
- Instance-based Representation
- Clusters

23

Knowledge Representation Methods

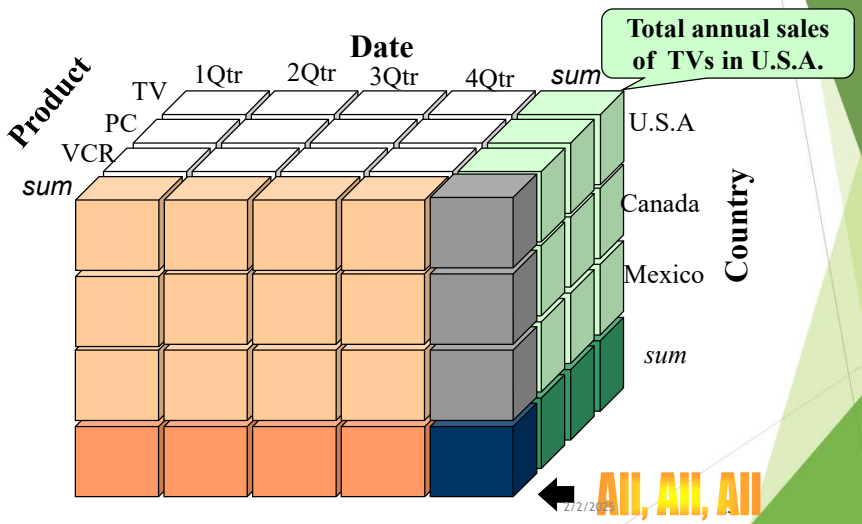
- Decision table for the weather problem:

Outlook	Humidity	Play
Sunny	High	No
Sunny	Normal	Yes
Overcast	High	Yes
Overcast	Normal	Yes
Rainy	High	No
Rainy	Normal	No

24

Knowledge Representation Methods

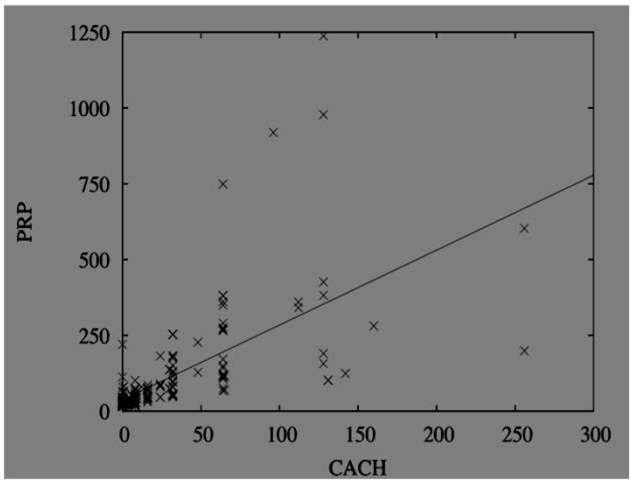
- A sample data cube:



25

Knowledge Representation Methods

- A linear regression function for the CPU performance data

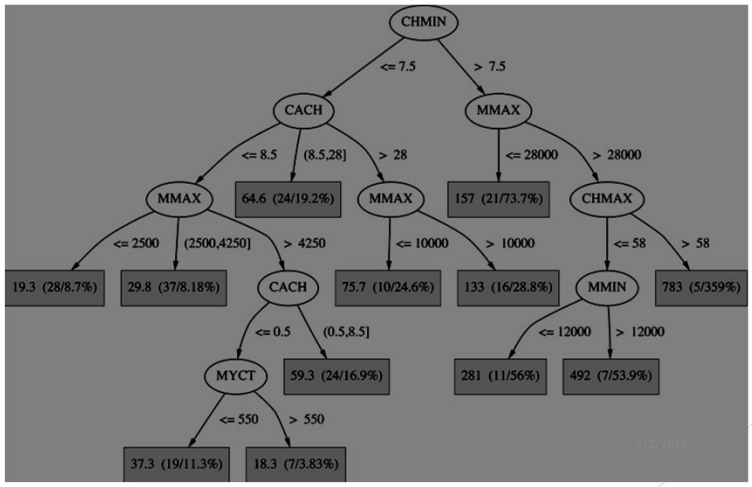


PRP = 37.06 + 2.47CACH

26

Knowledge Representation Methods

- Regression tree for the CPU data



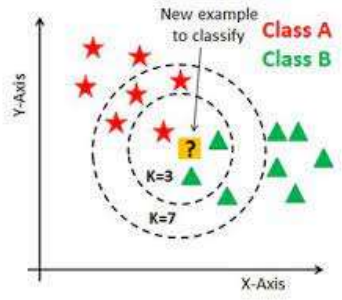
Knowledge Representation Methods

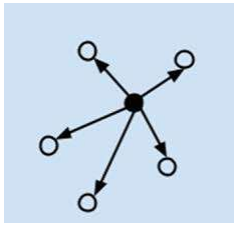
- If-then Rules

```
If tear production rate = reduced
   then recommendation = none
Otherwise, if age = young and astigmatic = no
   then recommendation = soft
```

Knowledge Representation Methods

- Instance-based representation

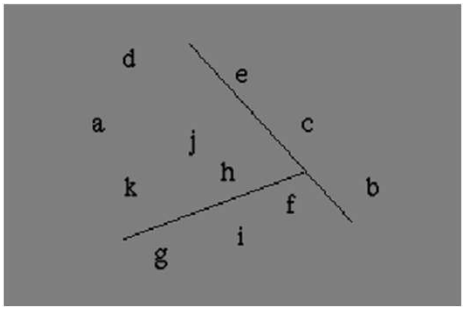


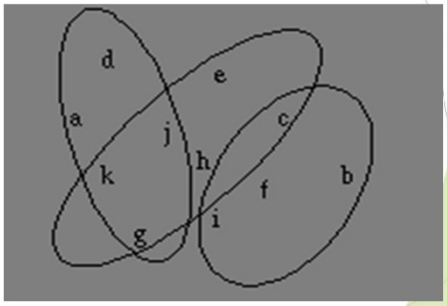


29

Knowledge Representation Methods

- Clusters





30

Introduction

- What is data mining?
- Data Mining Goals
- Stages of the Data Mining Process
- Data Mining Techniques
- Knowledge Representation Methods
- Applications
- Example: weather data

Applications

- The result of learning—or the learning method itself—is deployed in practical applications
  - Processing loan applications
  - Screening images for oil slicks
  - Electricity supply forecasting
  - Diagnosis of machine faults
  - Marketing and sales
  - Separating crude oil and natural gas
  - Reducing banding in rotogravure printing
  - Finding appropriate technicians for telephone faults
  - Scientific applications: biology, astronomy, chemistry
  - Automatic selection of TV programs
  - Monitoring intensive care patients

Processing loan applications (American Express)

- Given: questionnaire with financial and personal information
- Question: should money be lent?
- Simple statistical method covers 90% of cases
- Borderline cases referred to loan officers
- But: 50% of accepted borderline cases defaulted!
- Solution: reject all borderline cases?
  - No! Borderline cases are most active customers

33

Enter machine learning

- 1000 training examples of borderline cases
- 20 attributes:
  - age
  - years with current employer
  - years at current address
  - years with the bank
  - other credit cards possessed,…
- Learned rules: correct on 70% of cases
  - human experts only 50%
- Rules could be used to explain decisions to customers

34

Screening images

- Given: radar satellite images of coastal waters
- Problem: detect oil slicks in those images
- Oil slicks appear as dark regions with changing size and shape
- Not easy: lookalike dark regions can be caused by weather conditions (e.g. high wind)
- Expensive process requiring highly trained personnel

2/2/2025

35

35

Enter machine learning

- Extract dark regions from normalized image
- Attributes:
  - size of region
  - shape, area
  - intensity
  - sharpness and jaggedness of boundaries
  - proximity of other regions
  - info about background
- Constraints:
  - Few training examples—oil slicks are rare!
  - Unbalanced data: most dark regions aren't slicks
  - Regions from same image form a batch
  - Requirement: adjustable false-alarm rate

2/2/2025

36

36

Load forecasting

- Electricity supply companies need forecast of future demand for power
- Forecasts of min/max load for each hour ®significant savings
- Given: manually constructed load model that assumes “normal” climatic conditions
- Problem: adjust for weather conditions
- Static model consists of:
  - base load for the year
  - load periodicity over the year
  - effect of holidays

2/2/2025

37

37

Enter machine learning

- Prediction corrected using “most similar” days
- Attributes:
  - temperature
  - humidity
  - wind speed
  - cloud cover readings
  - plus difference between actual load and predicted load
- Average difference among three “most similar” days added to static model
- Linear regression coefficients form attribute weights in similarity function

2/2/2025

38

38

Diagnosis of machine faults

- Diagnosis: classical domain of expert systems
- Given: Fourier analysis of vibrations measured at various points of a device's mounting
- Question: which fault is present?
- Preventative maintenance of electromechanical motors and generators
- Information very noisy
- So far: diagnosis by expert/hand-crafted rules

2/2/2025

39

39

Enter machine learning

- Available: 600 faults with expert's diagnosis
- ~300 unsatisfactory, rest used for training
- Attributes augmented by intermediate concepts that embodied causal domain knowledge
- Expert not satisfied with initial rules because they did not relate to his domain knowledge
- Further background knowledge resulted in more complex rules that were satisfactory
- Learned rules outperformed hand-crafted ones

2/2/2025

40

40

Marketing and sales I

- Companies precisely record massive amounts of marketing and sales data
- Applications:
  - Customer loyalty:
    identifying customers that are likely to defect by detecting changes in their behavior
    (e.g. banks/phone companies)
  - Special offers:
    identifying profitable customers
    (e.g. reliable owners of credit cards that need extra money during the holiday season)

41

Marketing and sales II

- Market basket analysis
  - Association techniques find groups of items that tend to occur together in a transaction
    (used to analyze checkout data)
- Historical analysis of purchasing patterns
- Identifying prospective customers
  - Focusing promotional mailouts
    (targeted campaigns are cheaper than mass-marketed ones)

42

Introduction

- What is data mining?
- Data Mining Goals
- Stages of the Data Mining Process
- Data Mining Techniques
- Knowledge Representation Methods
- Applications
- Example: weather data

The weather problem

- Conditions for playing a certain game

Outlook	Temperature	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	Normal	False	Yes
...	...	...	...	...

```
If outlook = sunny and humidity = high then play = no
If outlook = rainy and windy = true then play = no
If outlook = overcast then play = yes
If humidity = normal then play = yes
If none of the above then play = yes
```

Classification vs. association rules

- Classification rule:
  predicts value of a given attribute (the classification of an example)

```
If outlook = sunny and humidity = high
   then play = no
```

- Association rule:
  predicts value of arbitrary attribute (or combination)

```
If temperature = cool then humidity = normal
If humidity = normal and windy = false
   then play = yes
If outlook = sunny and play = no
   then humidity = high
If windy = false and play = no
   then outlook = sunny and humidity = high
```


45

Weather data with mixed attributes

- Some attributes have numeric values

Outlook	Temperature	Humidity	Windy	Play
Sunny	85	85	False	No
Sunny	80	90	True	No
Overcast	83	86	False	Yes
Rainy	75	80	False	Yes
...	...	...	...	...

```
If outlook = sunny and humidity > 83 then play = no
If outlook = rainy and windy = true then play = no
If outlook = overcast then play = yes
If humidity < 85 then play = yes
If none of the above then play = yes
```


46


Summary
What is Data Mining?
What kinds of Data can be mined?
Which Technologies are used?
Which kinds of applications are targeted?
2/2/2025
47

47