



CS 412 Intro. to Data Mining


Chapter 3. Data Preprocessing

Jiawei Han, Computer Science, Univ. Illinois at Urbana-Champaign, 2017



Chapter 3: Data Preprocessing

การเตรียมข้อมูล หรือ การประมวลผลข้อมูล

- ❑ Data Preprocessing: An Overview 
- ❑ Data Cleaning
- ❑ Data Integration
- ❑ Data Reduction and Transformation
- ❑ Dimensionality Reduction
- ❑ Summary

What is Data Preprocessing? — Major Tasks

เป็นขั้นตอนสำหรับการจัดข้อมูลที่ไม่เกี่ยวข้องออกไป

□ Data cleaning

- Handle missing data, smooth noisy data, identify or remove outliers, and resolve inconsistencies

ข้อมูลที่ผิดไปเลย

□ Data integration รวมเดต้าจากหลายๆแหล่ง

- Integration of multiple databases, data cubes, or files

□ Data reduction ลดมิติข้อมูล

- Dimensionality reduction เลือก Data ที่สำคัญ
- Numerosity reduction
- Data compression เปลี่ยนแปลงชนิดของข้อมูลเพื่อให้เข้ากับข้อมูลชนิดอื่นๆได้

□ Data transformation and data discretization


แปลงข้อมูลในขั้นตอนการคัดเลือก ให้เหมาะสำหรับขั้นตอนการทำเหมืองข้อมูล

- Normalization
- Concept hierarchy generation

Why Preprocess the Data? — Data Quality Issues

- ❑ Measures for data quality: A multidimensional view
 - ❑ Accuracy: correct or wrong, accurate or not คนกรอกหลายคน , เปลี่ยนเครื่อง
 - ❑ Completeness: not recorded, unavailable, ... กรอกข้อมูลไม่ครบถ้วน
 - ❑ Consistency: some modified but some not, dangling, ... ทำ Normalization ไม่ดี ข้อมูลก็จะไปอยู่ในหลายที่
 - ❑ Timeliness: timely update? Data ที่กรอกในเวลาหนึ่ง และเวลาต่อมา มีการเปลี่ยนแปลง
 - ❑ Believability: how trustable the data are correct?
 - ❑ Interpretability: how easily the data can be understood?

Chapter 3: Data Preprocessing

- ❑ Data Preprocessing: An Overview
- ❑ Data Cleaning 
- ❑ Data Integration
- ❑ Data Reduction and Transformation
- ❑ Dimensionality Reduction
- ❑ Summary

Data Cleaning

การทำความสะอาดข้อมูล

- ❑ Data in the Real World Is Dirty: Lots of potentially incorrect data, e.g., instrument faulty, human or computer error, and transmission error ข้อมูลที่ผิดไม่ว่าจะจากคอมพิวเตอร์ หรือ คน
- ❑ Incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
 - ❑ e.g., *Occupation* = “ ” (missing data)
- ❑ Noisy: containing noise, errors, or outliers
 - ❑ e.g., *Salary* = “-10” (an error)
- ❑ Inconsistent: containing discrepancies in codes or names, e.g.,
 - ❑ *Age* = “42”, *Birthday* = “03/07/2010” อายุ & วันเกิด ไม่ตรงกัน
 - ❑ Was rating “1, 2, 3”, now rating “A, B, C”
 - ❑ discrepancy between duplicate records
- ❑ Intentional (e.g., *disguised missing* data)
 - ❑ Jan. 1 as everyone’s birthday?

Incomplete (Missing) Data

ข้อมูลไม่สมบูรณ์

ไม่ได้กรอกข้อมูล

- ❑ Data is not always available
 - ❑ E.g., many tuples have no recorded value for several attributes, such as customer income in sales data
- ❑ Missing data may be due to เกิดจาก
 - ❑ Equipment malfunction เครื่องเกิด
 - ❑ Inconsistent with other recorded data and thus deleted เครื่องทำงานไม่สม่ำเสมอ
 - ❑ Data were not entered due to misunderstanding
 - ❑ Certain data may not be considered important at the time of entry
 - ❑ Did not register history or changes of the data
- ❑ Missing data may need to be inferred

How to Handle Missing Data?

- ❑ Ignore the tuple: usually done when class label is missing (when doing classification)—not effective when the % of missing values per attribute varies considerably
- ❑ Fill in the missing value manually: tedious + infeasible? เต็ม + ลบออก
- ❑ Fill in it automatically with กรอกแทนโดยการสร้าง class ใหม่
 - ❑ a global constant : e.g., “unknown”, a new class?!
 - ❑ the attribute mean เอาค่ากลางมาแทนค่า missing
 - ❑ the attribute mean for all samples belonging to the same class: smarter แทนด้วย mean ที่อยู่ใน class เดียวกัน
 - ❑ **the most probable value: inference-based such as Bayesian formula or decision tree**

Noisy Data ข้อมูลรบกวน

- ❑ **Noise:** random error or variance in a measured variable
- ❑ **Incorrect attribute values** may be due to
 - ❑ Faulty data collection instruments
 - ❑ Data entry problems
 - ❑ Data transmission problems
 - ❑ Technology limitation
 - ❑ Inconsistency in naming convention
- ❑ **Other data problems**
 - ❑ Duplicate records
 - ❑ Incomplete data
 - ❑ Inconsistent data

Noisy Data

ข้อมูลรบกวน เป็นข้อมูลที่มีความผิดพลาด

หรือความคลาดเคลื่อน สาเหตุจาก

- อุปกรณ์เก็บรวบรวมข้อมูลทำหน้าที่ผิดพลาด

- ปัญหาการบันทึก หรือ ป้อนค่าข้อมูล

- ปัญหาการส่งข้อมูลผิดพลาด

- ข้อจำกัดทางเทคโนโลยี

How to Handle Noisy Data?

- ❑ Binning ทำให้ค่าข้อมูลที่เราสนใจลำดับเป็นไปอย่างราบรื่น โดยพิจารณาจากค่าที่อยู่รอบๆ
 - ❑ First sort data and partition into (equal-frequency) bins
 - ❑ Then one can **smooth by bin means, smooth by bin median, smooth by bin boundaries**, etc.
- ❑ Regression ข้อมูลสามารถปรับให้เรียงได้โดยการปรับข้อมูลให้เหมาะสมกับฟังก์ชัน
 - ❑ Smooth by fitting the data into regression functions
- ❑ Clustering ค่าผิดปกติอาจถูกตรวจสอบพบโดยการจัดกลุ่ม โดยมีย่านที่คล้ายกันจะถูกจัดเป็นกลุ่ม
 - ❑ Detect and remove outliers
- ❑ Semi-supervised: Combined computer and human inspection
 - ❑ Detect suspicious values and check by human (e.g., deal with possible outliers)