KHAI THÁC DỮ LIỆU & ỨNG DỤNG *(DATA MINING)*

GV : ThS. NGUYỄN HOÀNG TÚ ANH

BÀI 2
QUI TRÌNH CHUẨN BỊ DỮ LIỆU

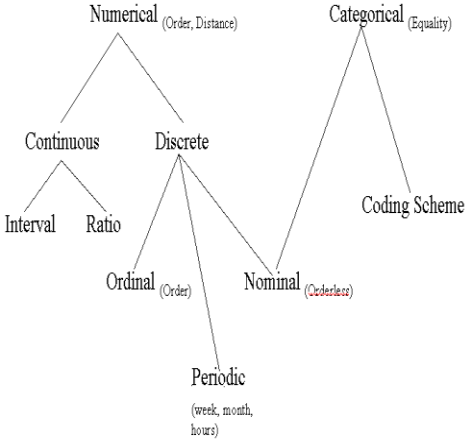
NỘI DUNG

1. **Tại sao cần chuẩn bị dữ liệu?**
2. Làm sạch dữ liệu (data cleaning)
3. Chọn lọc dữ liệu (data selection)
4. Rút gọn dữ liệu (data reduction)
5. Mã hoá dữ liệu

3

CÁC KIỂU DỮ LIỆU

- Dữ liệu dạng thuộc tính - giá trị (Attribute-value data)
- Các kiểu dữ liệu
  - số (numeric), phi số (categorical)
  - Tĩnh, động (thời gian)
- Các dạng dữ liệu khác
  - DL phân tán
  - DL văn bản
  - DL web, siêu DL
  - Hình ảnh, audio/video
  - ....



4

CHUẨN BỊ DỮ LIỆU

- **Dữ liệu trong thực tế có chất lượng xấu**
  - **DL thiếu, không đầy đủ** : thiếu giá trị của thuộc tính, thiếu các thuộc tính quan tâm, hoặc chỉ chứa DL tích hợp
    - VD : tuổi, cân nặng = “”
  - ***DL bị tạp, nhiễu (noise)*** *: chứa lỗi hoặc các sai biệt*
    - *VD : Lương =“-100 000”*
  - **DL mâu thuẫn** : có sự không thống nhất trong mã hoặc trong tên
    - VD : Tuổi =42 , Ngày sinh = 03/07/1997; US=USA?

5

Tại sao DL có chất lượng xấu ?

- **Bài tập theo nhóm số 3**
  - **Tình huống** : Bạn đi phỏng vấn xin việc làm tại phòng quản lý thông tin của công ty **ĐIỆN TỬ X** (**gồm rất nhiều chi nhánh trên toàn quốc**).
    - ***Người phỏng vấn đặt ra vấn đề*** **: Bạn cần thu thập DL bán hàng của tất cả các chi nhánh trong quí 1/2009 để phân tích kết quả kinh doanh.** *Những vấn đề gì cần đối mặt và hướng giải quyết. Dựa trên nội dung phỏng vấn để xác định xem người xin tuyển dụng có đáp ứng được yêu cầu của công ty không?*
    - **Nội dung của cuộc phỏng vấn tập trung vào bài toán thu thập, chuẩn bị dữ liệu và chất lượng dữ liệu. Không phỏng vấn về việc sử dụng dữ liệu để phân tích kết quả kinh doanh như thế nào**.

6

Tại sao DL có chất lượng xấu ?

- **Bài tập theo nhóm số 3**
  - **Cách thực hiện :**
    - Mỗi nhóm sẽ chia làm **3 nhóm nhỏ** : nhóm phỏng vấn, nhóm đi phỏng vấn và nhóm quan sát. Các nhóm này sẽ thực hiện phỏng vấn và đi phỏng vấn chéo với nhóm khác (theo danh sách đã công bố).
    - ***Ví dụ :*** *nhóm A có nhóm A1 – phỏng vấn, A2– đi phỏng vấn và A3 - quan sát. Tương tự với nhóm B.* Khi đó nhóm A1 sẽ phỏng vấn nhóm B2 (theo cặp nếu có nhiều hơn 1 người trong nhóm) và nhóm A3 quan sát . Nhóm B1 sẽ phỏng vấn nhóm A2 (theo cặp nếu có nhiều hơn 1 người trong nhóm) và nhóm B3 quan sát. *Trong trường hợp số người quan sát nhiều hơn 1 thì sẽ chia ra quan sát ở cả 2 cuộc phỏng vấn trong một Group.*

Tại sao DL có chất lượng xấu ?

- **Bài tập theo nhóm số 3**
  - **Cách thực hiện :**
    - Mỗi nhóm sẽ chia làm **3 nhóm nhỏ** : nhóm phỏng vấn, nhóm đi phỏng vấn và nhóm quan sát.
    - ***Cách chia nhóm*** *:*
      - Nếu **nhóm có 4 SV** thì chia ra : 1SV- phỏng vấn, 1SV-đi phỏng vấn và 2 SV-quan sát (SV quan sát sẽ chia ra quan sát ở cả 2 cuộc phỏng vấn trong một Group)
      - Nếu **nhóm có 3 SV** thì chia ra : 1 SV - phỏng vấn, 1 SV - đi phỏng vấn và 1 SV - quan sát.

8

Tại sao DL có chất lượng xấu ?

- **Bài tập theo nhóm số 3**
  - **Cách thực hiện :**
    - Mỗi nhóm sẽ chia làm **3 nhóm nhỏ**. Các nhóm này sẽ thực hiện phỏng vấn và đi phỏng vấn chéo với nhóm khác.
    - ***Nhóm quan sát thực hiện việc ghi lại biên bản phỏng vấn : thông tin về người phỏng vấn, người đi phỏng vấn, người quan sát, nhóm, các câu hỏi, trả lời liên quan đến nội dung thu thập DL và kết quả cuộc phỏng vấn và tự đánh giá chất lượng cuộc phỏng vấn .***
    - ***Tiêu chí đánh giá bài tập số 3* : thông qua chất lượng câu hỏi, câu trả lời có nhắm đúng mục tiêu và nội dung phỏng vấn hay không. Đánh giá qua biên bản phỏng vấn và nhận xét tự đánh giá.**

Tại sao DL có chất lượng xấu ?

- **Bài tập theo nhóm số 3**
  - **Thời gian thực hiện phỏng vấn : 7'.**
- **Một số câu hỏi gợi ý :**
  1. **Sau khi thu thập DL từ các chi nhánh, bạn có thể gặp những vấn đề gì?**
  2. **Ví dụ ?**
  3. ***Lý do ?***
- ***Mỗi quan sát viên đều phải có một biên bản phỏng vấn và nộp chung theo Group. Lưu ý : ghi rõ các thông tin liên quan đến nhóm và kết quả có tuyển dụng hay không. Viết ngắn gọn, súc tích.***

CHUẨN BỊ DỮ LIỆU

- "DL không chất lượng, không cho kết quả khai thác tốt"
  - Quyết định đúng đắn phải dựa trên các DL chính xác
    - VD : việc trùng lắp hoặc thiếu DL có thể dẫn tới việc thống kê không chính xác, thậm chí làm lạc lối.
  - Nhà kho DL cần sự tích hợp đồng nhất các DL chất lượng

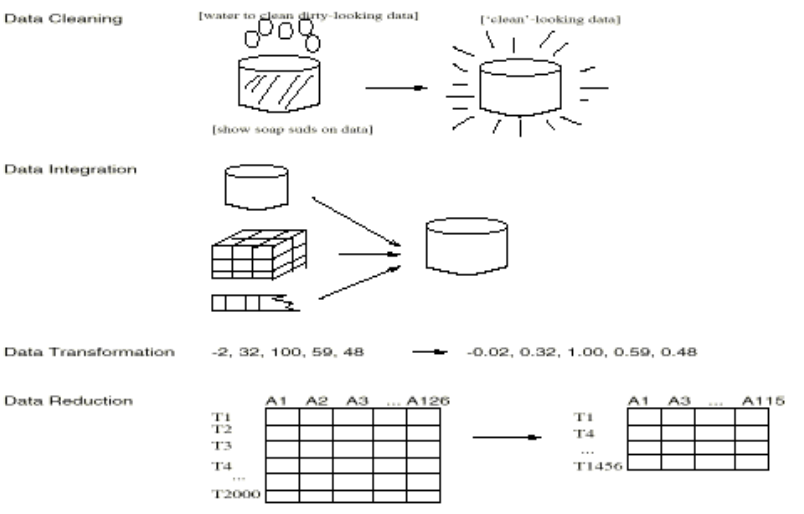
11

CHUẨN BỊ DỮ LIỆU

- **Các bước của quá trình chuẩn bị DL ?**
  - **Làm sạch DL**
    - Điền các giá trị thiếu, khử DL nhiễu, xác định và loại bỏ DL sai biệt, DL nhiễu và giải quyết DL mâu thuẫn
  - **Chọn lọc/ Tích hợp DL**
    - Tổng hợp, tích hợp DL từ nhiều CSDL, tập tin khác nhau .
  - **Biến đổi DL/ Mã hoá DL**
    - Chuẩn hoá và tổng hợp (aggregation) .
  - **Rút gọn DL**
    - Giảm kích thước DL nhưng đảm bảo kết quả phân tích .

12

CHUẨN BỊ DỮ LIỆU



NỘI DUNG

1. Tại sao cần chuẩn bị dữ liệu ?
2. **Làm sạch dữ liệu (data cleaning)**
3. Chọn lọc dữ liệu (data selection)
4. Rút gọn dữ liệu (data reduction)
5. Mã hoá dữ liệu

14

LÀM SẠCH DỮ LIỆU

- Làm sạch DL là vấn đề quan trọng bậc nhất của nhà kho DL
- Các nhiệm vụ của công đoạn làm sạch DL
  - Điền các giá trị còn thiếu
  - Xác định các sai biệt và khử DL tạp, nhiễu
  - Sửa chữa các DL mâu thuẫn

15

ĐIỀN DỮ LIỆU THIẾU

- Bỏ qua các mẫu tin có giá trị thiếu
  - Thường dùng khi thiếu nhãn của lớp (trong phân lớp)
  - Dễ, nhưng không hiệu quả, đặc biệt khi tỷ lệ giá trị thiếu của thuộc tính cao.
- Điền các giá trị thiếu bằng tay : vô vị + không khả thi
- Điền các giá trị thiếu tự động :
  - Thay thế bằng hằng số chung: VD : “không biết”. Có thể thành lớp mới trong DL

16

ĐIỀN DỮ LIỆU THIẾU

- Điền các giá trị thiếu tự động :
  - Thay thế bằng giá trị trung bình của thuộc tính
  - *Thay thế bằng giá trị trung bình của thuộc tính trong một lớp*
  - Thay thế bằng giá trị có nhiều khả năng nhất : suy ra từ công thức Bayesian, cây quyết định hoặc thuật giải EM (Expectation Maximization)

17

ĐIỀN DỮ LIỆU THIẾU

- **Tình huống:**
  - Thu thập DL về sinh viên thuộc tất cả các trường của ĐHQG Tp.HCM (Vd : để phân tích mức sống SV)
  - Các thuộc tính nào có thể có trong CSDL ?
  - ***Ví dụ thuộc tính bị thiếu giá trị là thuộc tính "Tiền thuê nhà"***
  - Cách giải quyết?

18

DỮ LIỆU NHIỄU

- **Các phương pháp cơ bản khử nhiễu :**
  - **Phương pháp chia giỏ (Binning)** :
    - Sắp xếp và chia DL vào các giỏ có cùng độ sâu (equal-depth)
    - Khử nhiễu bằng giá trị TB, trung tuyến, biên giỏ,…
  - **Gom nhóm (Clustering)** :
    - Phát hiện và loại bỏ các khác biệt
  - **Phương pháp hồi qui (Regression)** :
    - Đưa DL vào hàm hồi qui

19

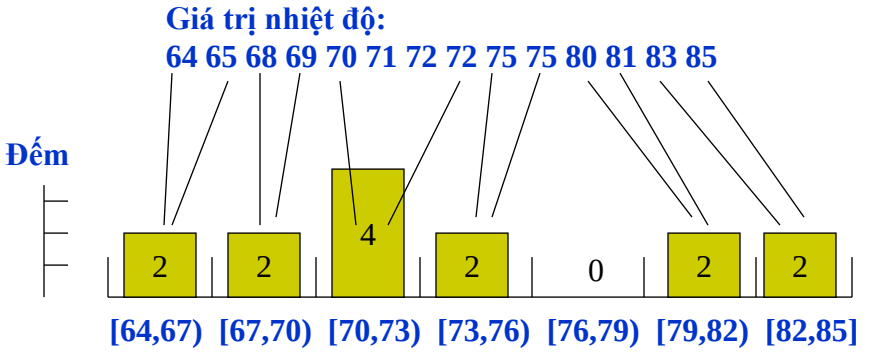
DỮ LIỆU NHIỄU

- **Phương pháp rời rạc hóa : chia giỏ (Binning)**
  - Chia theo độ rộng (Equal-width - khoảng cách) :
    - Chia vùng giá trị thành N khoảng cùng kích thước
    - Độ rộng của từng khoảng = (giá trị lớn nhất - giá trị nhỏ nhất)/N
  - Chia theo độ sâu (Equal-depth – tần suất) :
    - Chia vùng giá trị thành N khoảng mà mỗi khoảng có chứa gần như cùng số lượng mẫu

20

DỮ LIỆU NHIỄU

- **Phương pháp rời rạc hóa : chia giỏ** theo độ rộng (Equal-width – khoảng cách) :

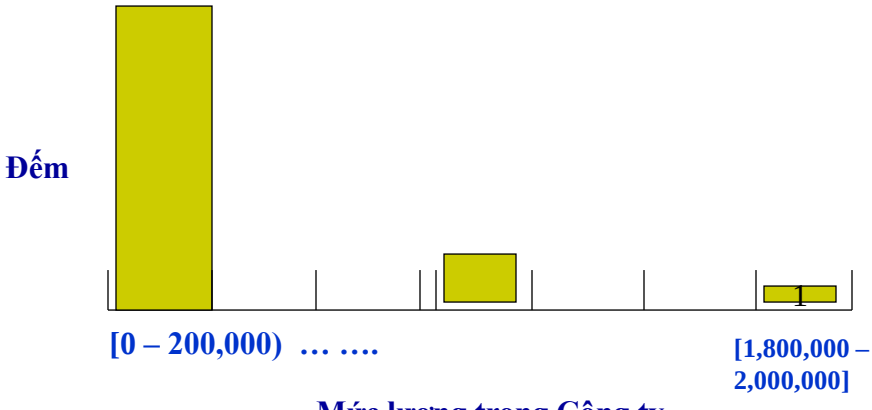


Biên trái giỏ <= giá trị < Biên phải giỏ

21

DỮ LIỆU NHIỄU

- **Phương pháp rời rạc hóa : chia giỏ** theo độ rộng (Equal-width – khoảng cách) : ***không tốt cho DL bị lệch***



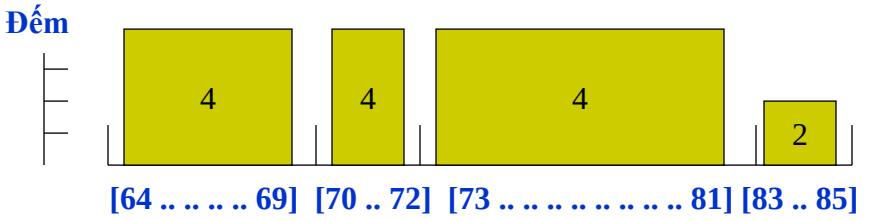
Mức lương trong Công ty

22

DỮ LIỆU NHIỄU

- **Phương pháp rời rạc hóa : chia giỏ** theo độ sâu (Equal-depth – tần suất) :

Giá trị nhiệt độ:
64 65 68 69 70 71 72 72 75 75 80 81 83 85



Độ sâu = 4, ngoại trừ giỏ cuối cùng

VÍ DỤ PHƯƠNG PHÁP CHIA GIỎ

- **Sắp xếp DL giá ($) :**
  4, 8, 15, 21, 21, 24, 25, 28, 34

* Phân chia thành giỏ có cùng độ sâu (equal-depth) : độ sâu = 3
- Bin 1: 4, 8, 15
- Bin 2: 21, 21, 24
- Bin 3: 25, 28, 34

* **Làm trơn =**

Bằng trung tuyến giỏ:
- Bin 1: 8, 8, 8
- Bin 2: 21, 21, 21
- Bin 3: 28, 28, 28

Bằng giá trị TB giỏ:
- Bin 1: 9, 9, 9
- Bin 2: 22, 22, 22
- Bin 3: 29, 29, 29

Bằng biên giỏ :
- Bin 1: 4, 4, 15
- Bin 2: 21, 21, 24
- Bin 3: 25, 25, 34

Bài tập phương pháp chia giỏ

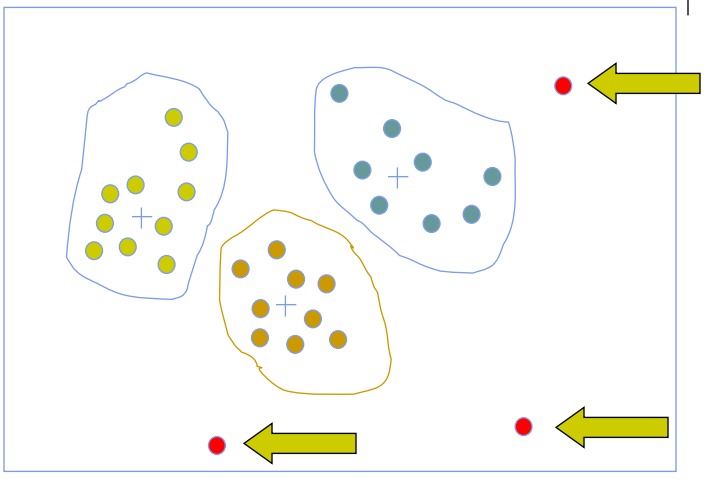
- **Thời gian : 8’**
- **Cho DL giá ($) :**
  15, 17, 19, 25, 29, 31, 33, 41, 42, 45, 45, 47, 52, 52, 64

SỐ GIỎ : 4

- Dùng phương pháp phân chia lần lượt theo **độ rộng** và theo **độ sâu**.
  - **Tính giá trị của giỏ theo phương pháp làm trơn theo trung tuyến :**
    - ***Nhóm:***
  - **Tính giá trị của giỏ theo phương pháp làm trơn theo biên giỏ :**
    - ***Nhóm:***
  - **So sánh kết quả hai phương pháp phân chia**

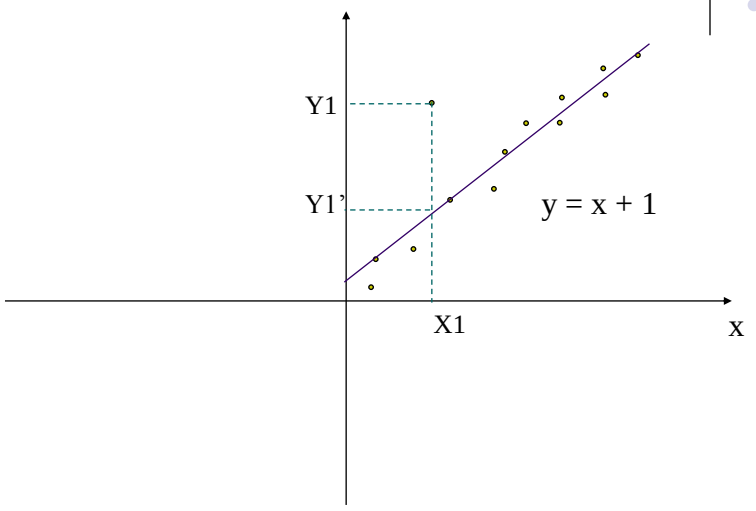
25

PHƯƠNG PHÁP GOM NHÓM



26

PHƯƠNG PHÁP HỒI QUI



27

DỮ LIỆU MÂU THUẪN

- Tự nghiên cứu trong tài liệu tham khảo để tìm câu trả lời : Làm thế nào để xử lý DL mâu thuẫn ?

28

NỘI DUNG

1. Tại sao cần chuẩn bị dữ liệu ?
2. Làm sạch dữ liệu (data cleaning)
3. **Chọn lọc dữ liệu (data selection)**
4. Rút gọn dữ liệu (data reduction)
5. Mã hoá dữ liệu

CHỌN LỌC DỮ LIỆU

- Tập hợp DL từ nhiều nguồn khác nhau vào trong một CSDL
  - Chỉ chọn những DL cần thiết cho tiến trình khai thác DL.
- Sơ đồ tập hợp DL
- Loại bỏ DL dư thừa và trùng lặp
- Phát hiện và giải quyết các mâu thuẫn trong DL

CHỌN LỌC DỮ LIỆU

Sơ đồ tập hợp DL

- Bài toán nhận diện thực thể
  - Làm thế nào để các thực thể từ nhiều nguồn DL trở nên tương xứng
  - US=USA; customer_id = cust_number
- Sử dụng siêu DL(metadata)

31

CHỌN LỌC DỮ LIỆU

Loại bỏ DL dư thừa, trùng lắp

- Một thuộc tính là thừa nếu nó có thể suy ra từ các thuộc tính khác
- Cùng một thuộc tính có thể có nhiều tên trong các CSDL khác nhau
- Một số mẫu tin DL bị lặp lại
- Dùng phép phân tích tương quan
  - r=0: X và Y không tương quan
  - r>0 : tương quan thuận. X↑↔Y↑
  - r<0: tương quan nghịch . X↓↔ Y ↑

$$r = \frac{\sum(x_i - \bar{X})(y_i - \bar{Y})}{\sqrt{\sum(x_i - \bar{X})^2 \sum(y_i - \bar{Y})^2}}$$

32

CHỌN LỌC DỮ LIỆU

- **Giải quyết mâu thuẫn trong DL**
  - Ví dụ : trọng lượng được đo bằng kg hoặc pound
  - Xác định chuẩn và ánh xạ dựa trên siêu dữ liệu (meta data)

33

NỘI DUNG

1. Tại sao cần chuẩn bị dữ liệu ?
2. Làm sạch dữ liệu (data cleaning)
3. Chọn lọc dữ liệu (data selection)
4. **Rút gọn dữ liệu (data reduction)**
5. Mã hoá dữ liệu

34

RÚT GỌN DỮ LIỆU

- DL có thể quá lớn đối với 1 số chương trình KTDL: Tốn nhiều thời gian.
- Rút gọn DL : DL được rút gọn (kích thước) sao cho vẫn thu được cùng (*hoặc gần như cùng*) kết quả phân tích.
- ***Các phương pháp*** :
  - Tổng hợp và tổng quát hóa
  - Giảm chiều DL
  - Nén DL
  - Giảm số lượng
  - Rời rạc hóa

35

RÚT GỌN DỮ LIỆU

Tổng hợp và tổng quát hóa

- Tổ hợp từ 2 thuộc tính (đối tượng) trở lên thành 1 thuộc tính (đối tượng)
  - *VD : các thành phố tổng hợp vào vùng, khu vực, nước, …*
- Tổng hợp/ tổng quát DL cấp thấp vào DL cấp cao :
  - Giảm kích thước tập DL : giảm số thuộc tính
  - Tăng tính lý thú của mẫu

36

RÚT GỌN DỮ LIỆU

Giảm chiều DL

- Chọn lựa đặc trưng (tập con các thuộc tính)
  - Chọn m từ n thuộc tính, m ≤ n
  - Loại bỏ các thuộc tính không liên quan, dư thừa
- Cách xác định thuộc tính không liên quan ?
  - Số liệu thống kê
  - Độ lợi thông tin

37

RÚT GỌN DỮ LIỆU

Giảm chiều DL bằng cách nào?

- **Vét cạn**
  - Có 2^d tập con thuộc tính của *d* thuộc tính
  - *Độ phức tạp tính toán quá cao*
- **PP Heuristic**
  - Stepwise forward selection
  - *Stepwise backward elimitation*
  - Kết hợp cả hai
  - *Cây quyết định qui nạp*

38

RÚT GỌN DỮ LIỆU

- ***PP Heuristic - Stepwise forward***
  - Đầu tiên : chọn thuộc tính đơn tốt nhất
  - *Chọn tiếp thuộc tính tốt nhất trong số còn lại, …..*
  - Ví dụ : tập thuộc tính ban đầu {A1,A2,A3,A4,A5,A6}
    - ***Tập rút gọn ban đầu*** ={}
      B1= {A1}
      B2= {A1,A4}
      B3= {A1,A4, A6}

39

RÚT GỌN DỮ LIỆU

- ***PP Heuristic - Stepwise backward***
  - Đầu tiên : loại thuộc tính đơn xấu nhất
  - *Loại tiếp thuộc tính xấu nhất trong số còn lại, …*
  - Ví dụ : tập thuộc tính ban đầu {A1,A2,A3,A4,A5,A6}
    - ***Tập rút gọn ban đầu*** ={A1,A2,A3,A4,A5,A6}
      B1= {A1,A3,A4,A5,A6}
      B2= {A1,A4,A5,A6}
      B3= {A1,A4, A6}

40

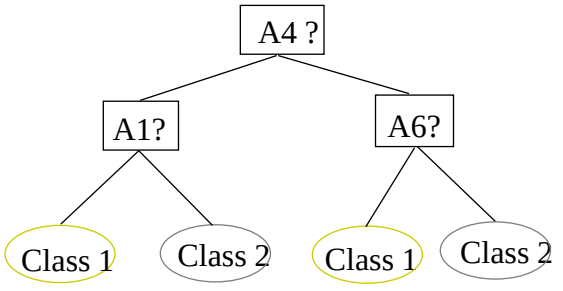
RÚT GỌN DỮ LIỆU

- ***PP Heuristic - Kết hợp***
  - Đầu tiên : chọn thuộc tính đơn tốt nhất và loại thuộc tính đơn xấu nhất
  - *Chọn tiếp thuộc tính tốt nhất và loại tiếp thuộc tính xấu nhất trong số còn lại, …*
  - Ví dụ : tập thuộc tính ban đầu {A1,A2,A3,A4,A5,A6}
    - ***Tập rút gọn ban đầu*** ={A1,A2,A3,A4,A5,A6}
      B1= {A1,A3,A4,A5,A6}
      B2= {A1,A4,A5,A6}
      B3= {A1,A4, A6}

41

RÚT GỌN DỮ LIỆU

- ***PP Heuristic – Cây quyết định qui nạp***
  - Đầu tiên : xây dựng cây quyết định
  - *Loại các thuộc tính không xuất hiện trên cây*
  - Ví dụ : tập thuộc tính ban đầu {A1,A2,A3,A4,A5,A6}
    ⇒***Tập rút gọn*** **= {A1, A4, A6}**



42

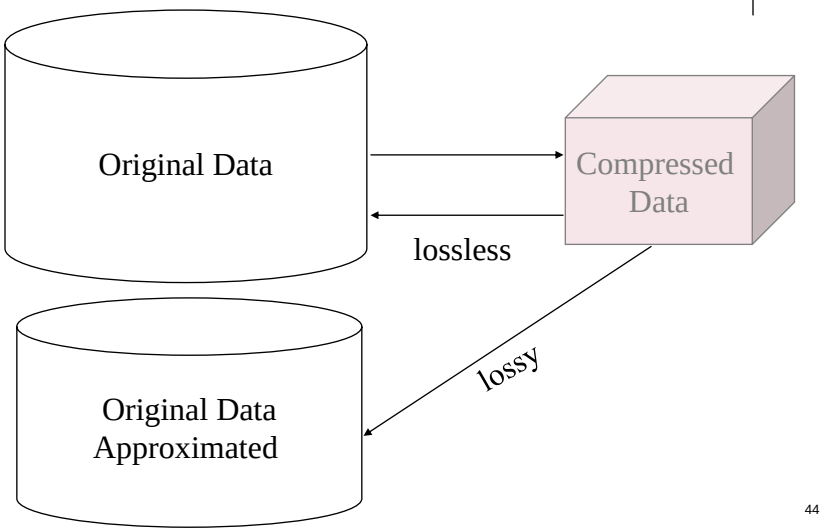
RÚT GỌN DỮ LIỆU

Nén DL

- Mã hoá hoặc biến đổi dữ liệu
- Nén không mất thông tin (lossless)
  - *DL có thể phục hồi lại*
- Nén có mất thông tin (lossy)
  - *DL không thể phục hồi lại hoàn toàn*
- Dùng biến đổi wavelet, phân tích thành phần cơ bản (principal component analysis-PCA), …

43

RÚT GỌN DỮ LIỆU



44

RÚT GỌN DỮ LIỆU

- **Giảm số lượng (numerosity reduction)**
  - Chọn dạng biểu diễn DL khác, "nhỏ hơn"
  - PP tham số :
    - Sử dụng mô hình toán học để lưu giữ các tham số (của DL)
    - *Mô hình hồi qui và log-tuyến tính*
  - PP không tham số :
    - Không sử dụng mô hình toán học mà lưu biểu diễn rút gọn
    - *Biểu đồ, gom nhóm, lấy mẫu*

45

RÚT GỌN DỮ LIỆU

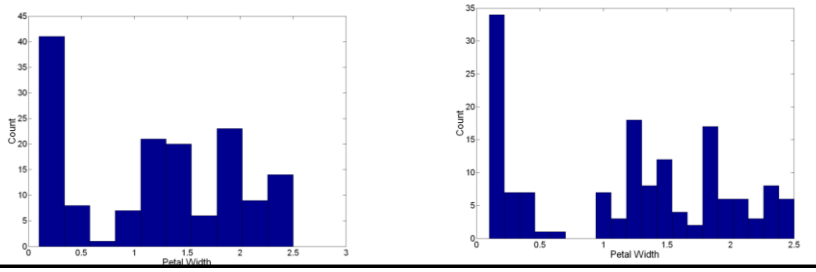
- **Giảm số lượng (tt)**
  - PP hồi qui tuyến tính : $Y = \alpha + \beta X$ **(chỉ lưu α, β)**
  - *PP hồi qui bội :* $Y = b_0 + b_1 X_1 + b_2 X_2$
  - Mô hình log-tuyến tính :
    - *Xác suất :* $p(a, b, c, d) = \alpha_{ab} \beta_{ac} \chi_{ad} \delta_{bcd}$

46

RÚT GỌN DỮ LIỆU

Giảm số lượng (tt)

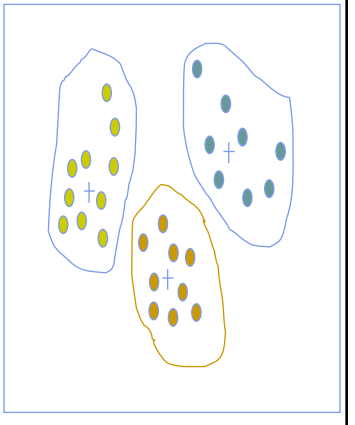
- **PP biểu đồ (histogram)**
  - *PP thông dụng để rút gọn DL*
  - Phân chia DL vào các giỏ và chiều cao của cột là số đối tượng nằm trong mỗi giỏ. Chỉ lưu giá trị trung bình của mỗi giỏ.
  - Hình dáng của biểu đồ tùy thuộc vào số lượng giỏ
- ***Ví dụ : Chiều dài cánh hoa (10 và 20 giỏ)***



RÚT GỌN DỮ LIỆU

Giảm số lượng (tt)

- PP gom nhóm
  - *Phân chia DL vào các nhóm và lưu biểu diễn của nhóm .*
  - Rất hiệu quả nếu DL tập trung thành nhóm nhưng ngược lại khi DL rải rác
  - *Rất nhiều thuật toán gom nhóm.*



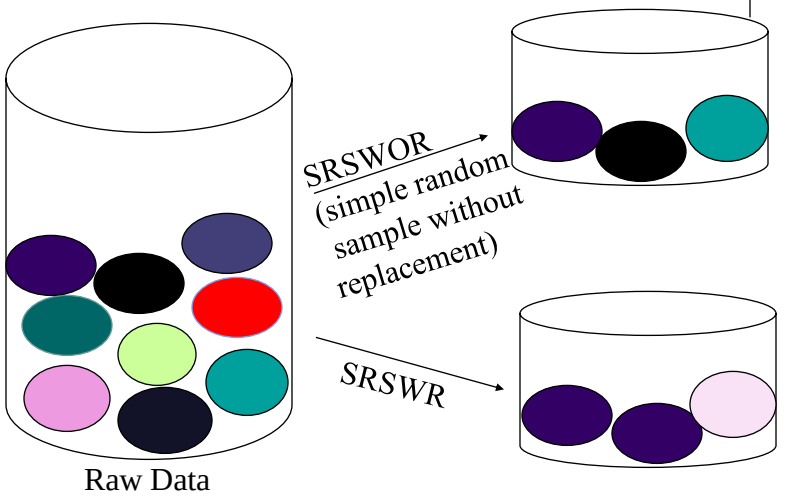
48

RÚT GỌN DỮ LIỆU

- **Giảm số lượng (tt)**
  - PP lấy mẫu (sampling)
    - *Dùng tập mẫu ngẫu nhiên nhỏ hơn nhiều để thay thế cho tập DL lớn.*
    - PP lấy mẫu ngẫu nhiên không thay thế (SRSWOR)
    - *PP lấy mẫu ngẫu nhiên có thay thế (SRSWR)*
    - PP lấy mẫu theo nhóm/phân cấp

49

RÚT GỌN DỮ LIỆU

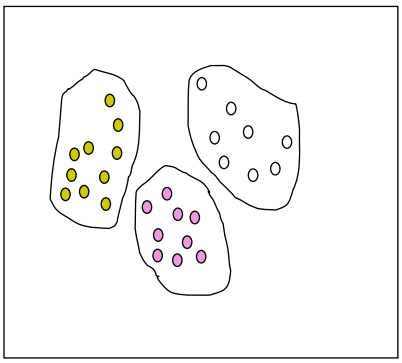


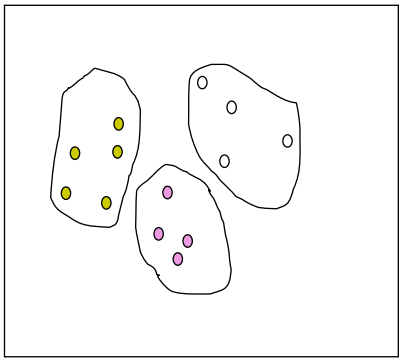
50

RÚT GỌN DỮ LIỆU

Raw Data

Cluster/Stratified Sample





51

NỘI DUNG

1. Tại sao cần chuẩn bị dữ liệu ?
2. Làm sạch dữ liệu (data cleaning)
3. Chọn lọc dữ liệu (data selection)
4. Rút gọn dữ liệu (data reduction)
5. **Mã hoá dữ liệu**

52

MÃ HÓA DỮ LIỆU

- **Mã hoá** : chuyển đổi DL thành dạng phù hợp và thuận tiện cho các thuật toán KTDL
  - Rời rạc hóa :
    - *Biến đổi miền giá trị thuộc tính (liên tục) bằng cách chia miền giá trị thành từng khoảng. Lưu nhãn của khoảng thay cho các giá trị thực.*
  - Phân cấp khái niệm :
    - *Tập hợp và thay thế khái niệm cấp thấp bằng khái niệm cấp cao hơn.*

53

MÃ HÓA DỮ LIỆU

- **PP mã hóa**
  - DL dạng số :
    - Chia giỏ
    - *Phân tích biểu đồ*
    - Gom nhóm
    - *Rời rạc hoá theo entropy*
    - Phân đoạn tự nhiên
  - DL dạng phi số :
    - *Tạo sơ đồ phân cấp.*

54

MÃ HÓA DỮ LIỆU

- **Ví dụ** :
  - Chuyển đổi giá trị logic thành 1,0
  - *Chuyển đổi giá trị ngày tháng thành số*
  - Chuyển đổi các cột có giá trị số lớn thành tập các giá trị trong vùng nhỏ hơn, chẳng hạn chia chúng cho hệ số nào đó
  - *Nhóm các giá trị có cùng ngữ nghĩa như : Hoạt động trước CMT8 là nhóm 1; từ 01/08/45 – 31/06/54 ; nhóm 2; từ 01/07/54 – 30/4/75 là nhóm 3, …*
  - Thay thế giá trị của Tuổi = trẻ, trung niên, già

55

TÓM TẮT

1. Thực tế DL - thiếu, nhiễu, mâu thuẫn và nhiều chiều
2. Chuẩn bị DL là vấn đề quan trọng của DM
3. Chuẩn bị DL gồm :
   - Làm sạch DLvà lựa chọn
   - Rút gọn DL
   - Mã hóa DL
4. Dữ liệu tốt là chìa khóa tạo ra các mô hình giá trị và đáng tin cậy.
5. Đây là lĩnh vực nghiên cứu còn nhiều thách thức

56

CÁC CÔNG VIỆC CẦN LÀM

1. ***Hoàn tất việc đăng ký báo cáo xemina :***
   - **Hạn chót : 26/9/2008 (qua Moodle)**

2. ***Chuẩn bị bài 3 : Khai thác tập phổ biến và luật kết hợp***
   - Xem nội dung các bài tập nhóm thuộc bài 3 – Phần 1.
   - **Chuẩn bị các BT nhóm chương 3**
   - **Cách thực hiện :**
     - **Đọc slide, xem các ví dụ**
     - **Tham khảo trên Internet và tài liệu tham khảo**

57

BÀI TẬP

1. Tại sao chuẩn bị DL là công việc cấp thiết và tốn nhiều thời gian ?
2. *Các cách giải quyết vần đề thiếu giá trị trong các mẫu tin của CSDL?*
3. Giả sử CSDL có thuộc tính Tuổi với các giá trị trong các mẫu tin (tăng dần): 13,15,16,16,19,20,20,21,22,22,25,25,25,25,30,33, 33,35,35, 35,35,36,40,45,46,52,70.
   a) *Khử nhiễu DL trên bằng giá trị TB của giỏ. Nhận xét hiệu quả của kỹ thuật này với DL trên.*
   b) Có thể áp dụng các kỹ thuật nào để khử nhiễu DL ?
   c) *Dùng DL trên vẽ biểu đồ cùng chiều rộng (equal-width histogram) với độ rộng = 10*

58

TÀI LIỆU THAM KHẢO

1. E.Rahm, H.H.Do. Data cleaning : Problems and Current Approaches. IEEE bulletin of Technical Committee on Data engineering, Vol. 23, N.4, 2000
2. J.Han, M.Kamber, Chương 2 – Data mining : Concepts and Techniques

59

Q & A

60