



TÀI LIỆU LÝ THUYẾT KTDL & UD

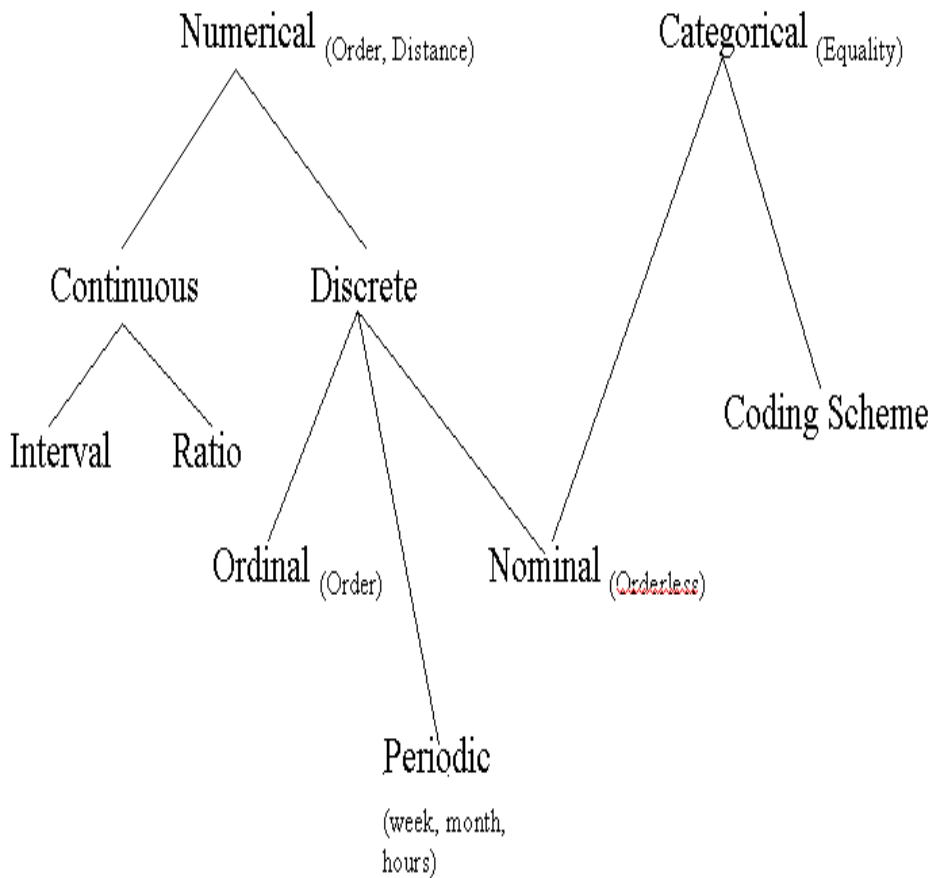
CHUẨN BỊ DỮ LIỆU

Giảng viên: ThS. Lê Ngọc Thành
Email: lnthanh@fit.hcmus.edu.vn

Nội dung

- **Tại sao cần chuẩn bị dữ liệu?**
- Làm sạch dữ liệu (data cleaning)
- Chọn lọc dữ liệu (data selection)
- Rút gọn dữ liệu (data reduction)
- Biến đổi dữ liệu (data transformation)

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
 -

Chất lượng dữ liệu

- ✓ **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 = ""

- ✓ **Tạp, nhiều (noise)** : chứa lỗi hoặc các sai biệt

VD : Lương = "-100 000"

- ✓ **Mâu thuẫn** : có sự không thống nhất trong mã hoặc trong tên

VD : Tuổi =42 , Ngày sinh = 03/07/1997;
US=USA?

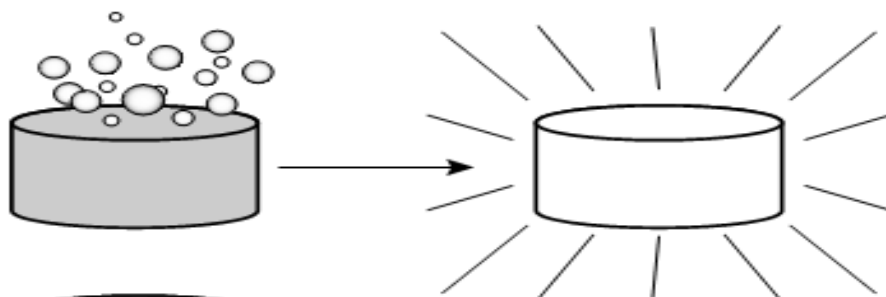
Hệ quả chất lượng dữ liệu

- Quyết định đúng đắn phải dựa trên các dữ liệu chính xác
 - VD : việc trùng lặp hoặc thiếu dữ liệu có thể dẫn tới việc thống kê không chính xác, thậm chí làm lạc lối.
- Kho dữ liệu cần sự tích hợp đồng nhất các DL chất lượng

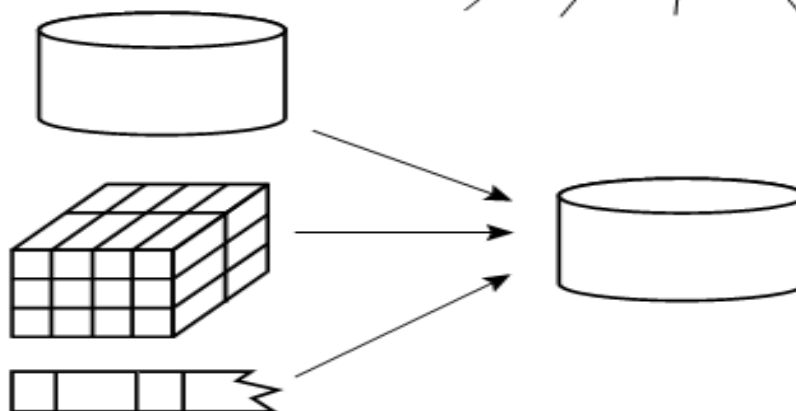
“Dữ liệu không chất lượng → khai thác không tốt”

Giải pháp? (1/2)

Data cleaning



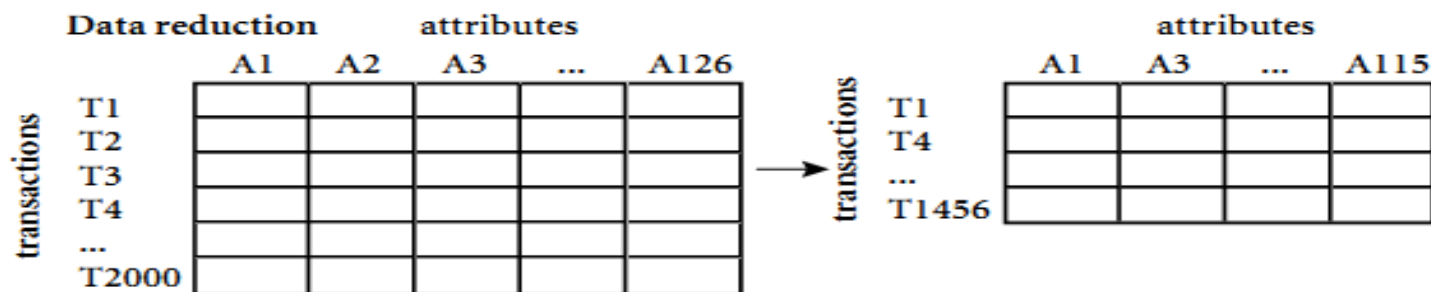
Data integration



Data transformation

-2, 32, 100, 59, 48 → -0.02, 0.32, 1.00, 0.59, 0.48

Data reduction



Giải pháp? (2/2)

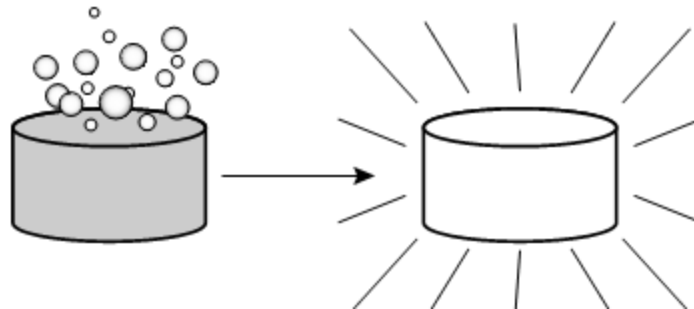
- ✓ **Cần làm sạch DL (Data Cleaning)**
 - Điền các giá trị thiếu, khử DL nhiễu, xác định và loại bỏ DL sai biệt, DL nhiễu và giải quyết DL mâu thuẫn
- ✓ **Cần chọn lọc/ Tích hợp DL (Data Intergration)**
 - Tổng hợp, tích hợp DL từ nhiều CSDL, tập tin khác nhau .
- ✓ **Cần biến đổi DL (Data transformation)**
 - Chuẩn hoá và tổng hợp (aggregation) .
- ✓ **Cần rút gọn DL**
 - Giảm kích thước DL nhưng đảm bảo kết quả phân tích .

Nội dung

- Tại sao cần chuẩn bị dữ liệu?
- **Làm sạch dữ liệu (data cleaning)**
- Chọn lọc dữ liệu (data selection)
- Rút gọn dữ liệu (data reduction)
- Biến đổi dữ liệu (data transformation)

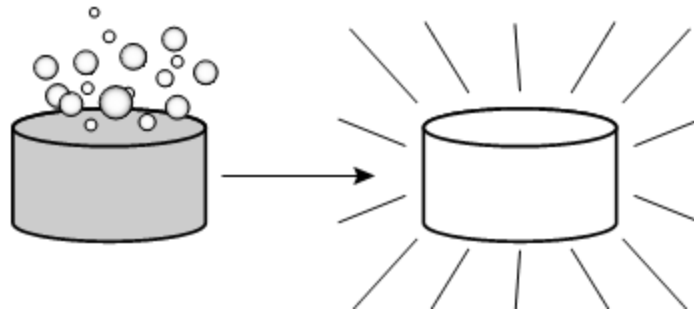
Làm sạch dữ liệu

- Làm sạch dữ liệu là vấn đề quan trọng bậc nhất
- Làm sạch dữ liệu là quá trình:
 - Điền các giá trị thiếu (missing data)
 - Xác định và loại bỏ dữ liệu sai biệt, dữ liệu nhiễu (noisy data)
 - Giải quyết dữ liệu mâu thuẫn



Làm sạch dữ liệu

- Làm sạch dữ liệu là vấn đề quan trọng bậc nhất
- Làm sạch dữ liệu là quá trình:
 - Điền các giá trị thiếu
 - Xác định và loại bỏ dữ liệu sai biệt, dữ liệu nhiễu
 - Giải quyết dữ liệu mâu thuẫn



Điền giá trị thiếu (1/2)

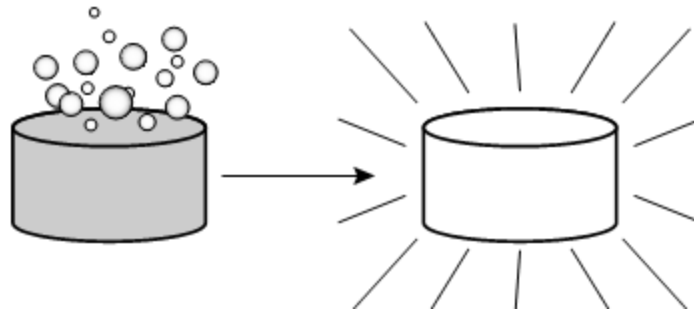
- Bỏ qua các mẫu tin có giá trị thiếu:
 - Thường dùng khi thiếu nhãn của lớp (trong phân lớp)
 - Dễ, nhưng không hiệu quả, đặc biệt khi tỷ lệ giá trị thiếu của thuộc tính cao.
- Điền các giá trị thiếu bằng tay: vô vị và không khả thi
- Điền các giá trị thiếu tự động:
 - Thay thế bằng hằng số chung. VD,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

Điền giá trị thiếu (2/2)

- 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)

Làm sạch dữ liệu

- Làm sạch dữ liệu là vấn đề quan trọng bậc nhất
- Làm sạch dữ liệu là quá trình:
 - Điền các giá trị thiếu
 - Xác định và loại bỏ dữ liệu sai biệt, dữ liệu nhiễu
 - Giải quyết dữ liệu mâu thuẫn



Khử 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,...
 - Phương pháp gom nhóm (Clustering):
 - Phát hiện và loại bỏ các khác biệt
 - Phương pháp hồi qui (Regression):
 - Đưa DL vào hàm hồi qui

Khử nhiễu – pp chia giỏ (1/4)

- Phương pháp 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

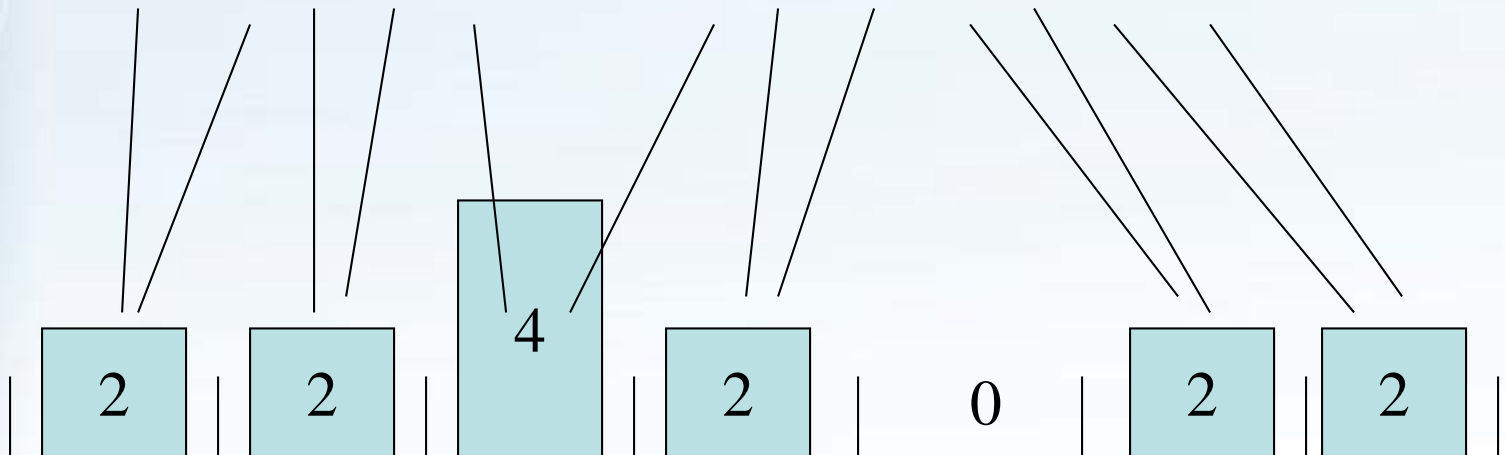
Khử nhiễu – pp chia giỏ (2/4)

- Ví dụ chia giỏ theo độ rộng:

Giá trị nhiệt độ với $N = 7$:

64 65 68 69 70 71 72 72 75 75 80 81 83 85

Đếm



[64,67) [67,70) [70,73) [73,76) [76,79) [79,82) [82,85]

Biên trái \leq giá trị $<$ Biên phải

*Chia vùng giá trị thành N khoảng cùng kích thước.
Độ rộng của từng khoảng = (giá trị lớn nhất - giá trị nhỏ nhất)/ N .*

Khử nhiễu – pp chia giỏ (3/4)

- Nhưng không tốt cho DL bị lệch

Đếm



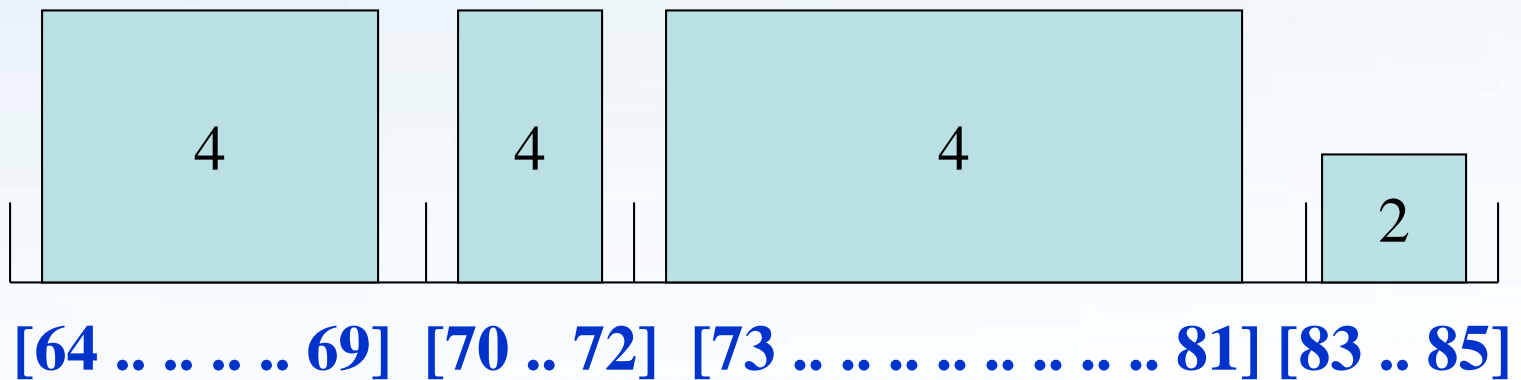
Khử nhiễu – pp chia giỏ (4/4)

- Chia giỏ theo độ sâu:

Giá trị nhiệt độ với $N = 4$:

64 65 68 69 70 71 72 72 75 75 80 81 83 85

Đếm



Độ sâu = 4, ngoại trừ giỏ cuối cùng

Chia vùng giá trị thành N khoảng mà mỗi khoảng có chứa gần như cùng số lượng mẫu

Khử nhiễu với giỏ đã chia

- Sắp xếp DL giá (\$) :
4, 8, 15, 21, 21, 24, 25, 28, 34
 - Phân chia thành giỏ có cùng độ sâu (equal-depth) với $N = 3$
 - Bin 1: 4, 8, 15
 - Bin 2: 21, 21, 24
 - Bin 3: 25, 28, 34
- Làm gì với giỏ đã chia?

Khử nhiễu với giỏ đã chia

- Bin 1: 4, 8, 15
- Bin 2: 21, 21, 24
- Bin 3: 25, 28, 34

Bảng trung vị giỏ:

- Bin 1: 8, 8, 8
- Bin 2: 21, 21, 21
- Bin 3: 28, 28, 28

Bảng giá trị TB giỏ:

- Bin 1: 9, 9, 9
- Bin 2: 22, 22, 22
- Bin 3: 29, 29, 29

Bảng biên giỏ :

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

Làm tròn

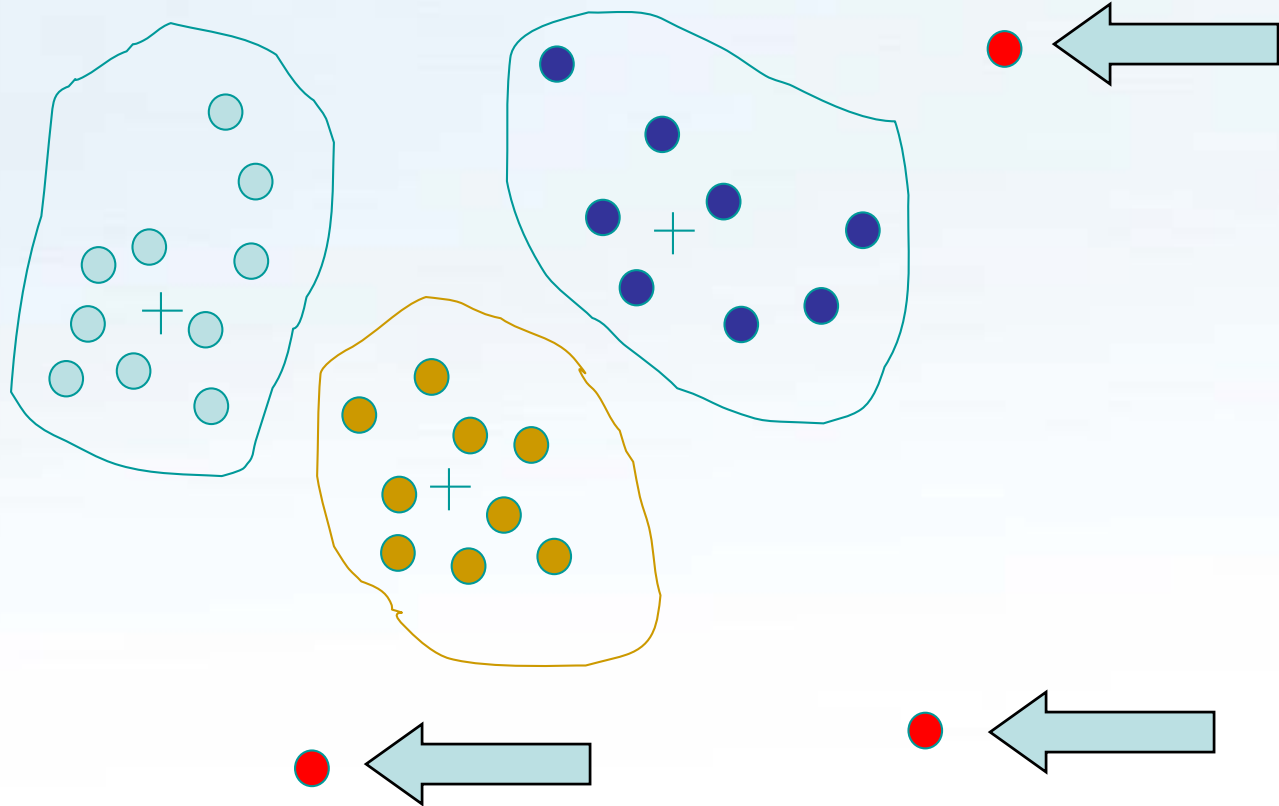
Bài tập khử nhiễu với giỏ

- Cho DL giá (\$) :
15, 17, 19, 25, 29, 31, 33, 41, 42, 45, 45, 47, 52, 52, 64
- Dùng phương pháp chia giỏ theo độ rộng và độ sâu với số giỏ là 4 để:
 - Tính giá trị của giỏ theo làm tròn trung vị.
 - Tính giá trị của giỏ theo làm tròn biên giỏ.
 - Tính giá trị của giỏ theo làm tròn TB giỏ.
 - Nhận xét kết quả đạt được.

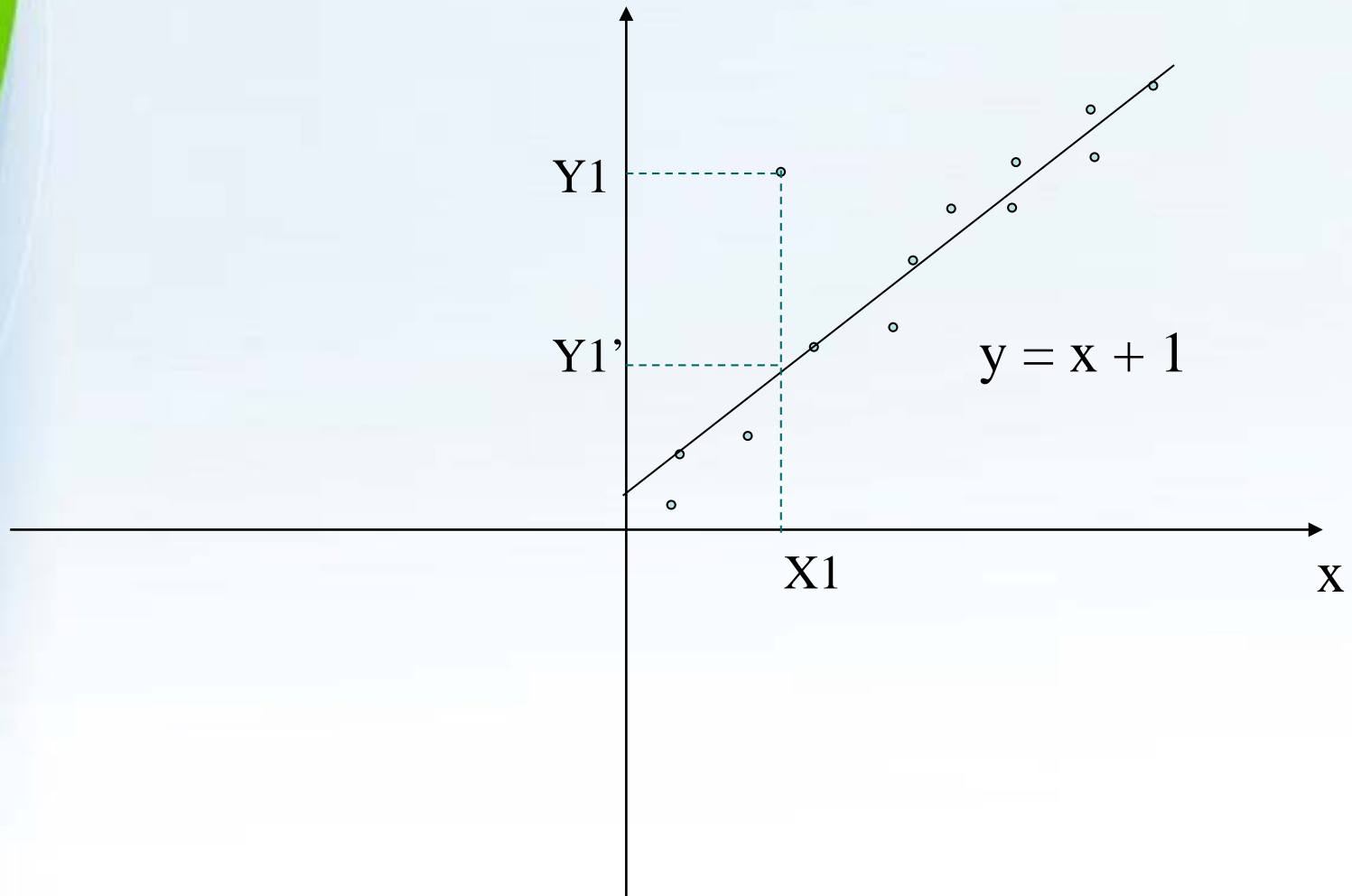
Khử nhiễu?

- 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,...
 - Phương pháp gom nhóm (Clustering):
 - Phát hiện và loại bỏ các khác biệt
 - Phương pháp hồi qui (Regression):
 - Đưa DL vào hàm hồi qui

Khử nhiễu – pp gom nhóm

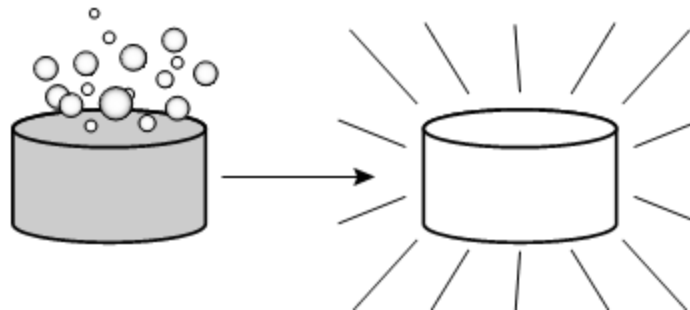


Khử nhiễu – pp hồi quy



Làm sạch dữ liệu

- Làm sạch dữ liệu là vấn đề quan trọng bậc nhất
- Làm sạch dữ liệu là quá trình:
 - Điền các giá trị thiếu
 - Xác định và loại bỏ dữ liệu sai biệt, dữ liệu nhiễu
 - Giải quyết dữ liệu mâu thuẫn



Giải quyết mâu thuẫn

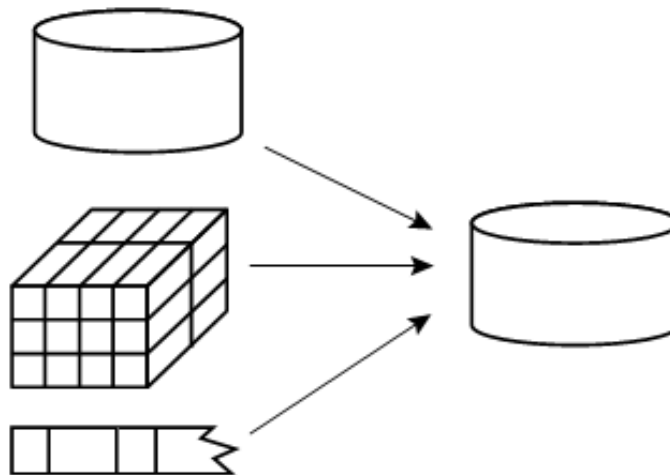
- Đọc thêm trong tài liệu tham khảo để trả lời câu hỏi:
 - Làm thế nào để xử lý DL mâu thuẫn?
 - Cho ví dụ từng phương pháp giải quyết mâu thuẫn.

Nội dung

- Tại sao cần chuẩn bị dữ liệu?
- Làm sạch dữ liệu (data cleaning)
- **Chọn lọc dữ liệu (data selection)**
- Rút gọn dữ liệu (data reduction)
- Biến đổi dữ liệu (data transformation)

Chọn lọc dữ liệu

- Chọn lựa và tập hợp DL từ nhiều nguồn khác nhau vào trong một CSDL
- Những vấn đề gì xảy ra khi chọn lựa và tổng hợp dữ liệu?



Quá trình chọn lọc dữ liệu (1/4)

- Quá trình:
 - Chỉ chọn những DL cần thiết cho tiến trình khai thác DL.
 - So khớp lược đồ dữ liệu
 - Loại bỏ DL dư thừa và trùng lặp
 - Phát hiện và giải quyết các mâu thuẫn trong DL

Quá trình chọn lọc dữ liệu (2/4)

- So khớp lược đồ dữ liệu
 - 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)

Quá trình chọn lọc dữ liệu (3/4)

- Loại bỏ dữ liệu dư thừa, trùng lặp
 - 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. $X \uparrow \leftrightarrow Y \uparrow$
 - $r<0$: tương quan nghịch . $X \downarrow \leftrightarrow Y \uparrow$

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

Quá trình chọn lọc dữ liệu (4/4)

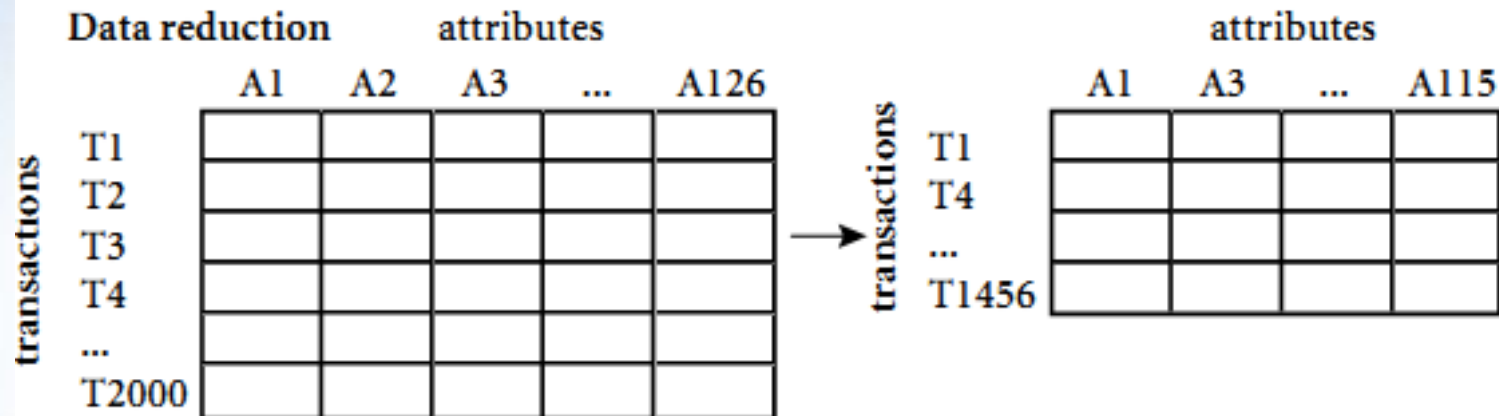
- Giải quyết mâu thuẫn trong dữ liệu
 - Ví dụ: trọng lượng được đo bằng kg hoặc pound
 - Xác định chuẩn và ánh xạ dựa trên siêu dữ liệu (metadata)

Nội dung

- Tại sao cần chuẩn bị dữ liệu?
- Làm sạch dữ liệu (data cleaning)
- Chọn lọc dữ liệu (data selection)
- **Rút gọn dữ liệu (data reduction)**
- Biến đổi dữ liệu (data transformation)

Rút gọn dữ liệu

- Dữ liệu có thể quá lớn đối với một số ứng dụng KTDL: tốn thời gian.
- Rút gọn dữ liệu là quá trình thu gọn dữ liệu (kích thước) sao cho vẫn thu được cùng (hoặc gần như cùng) kết quả phân tích.



Các phương pháp rút gọn

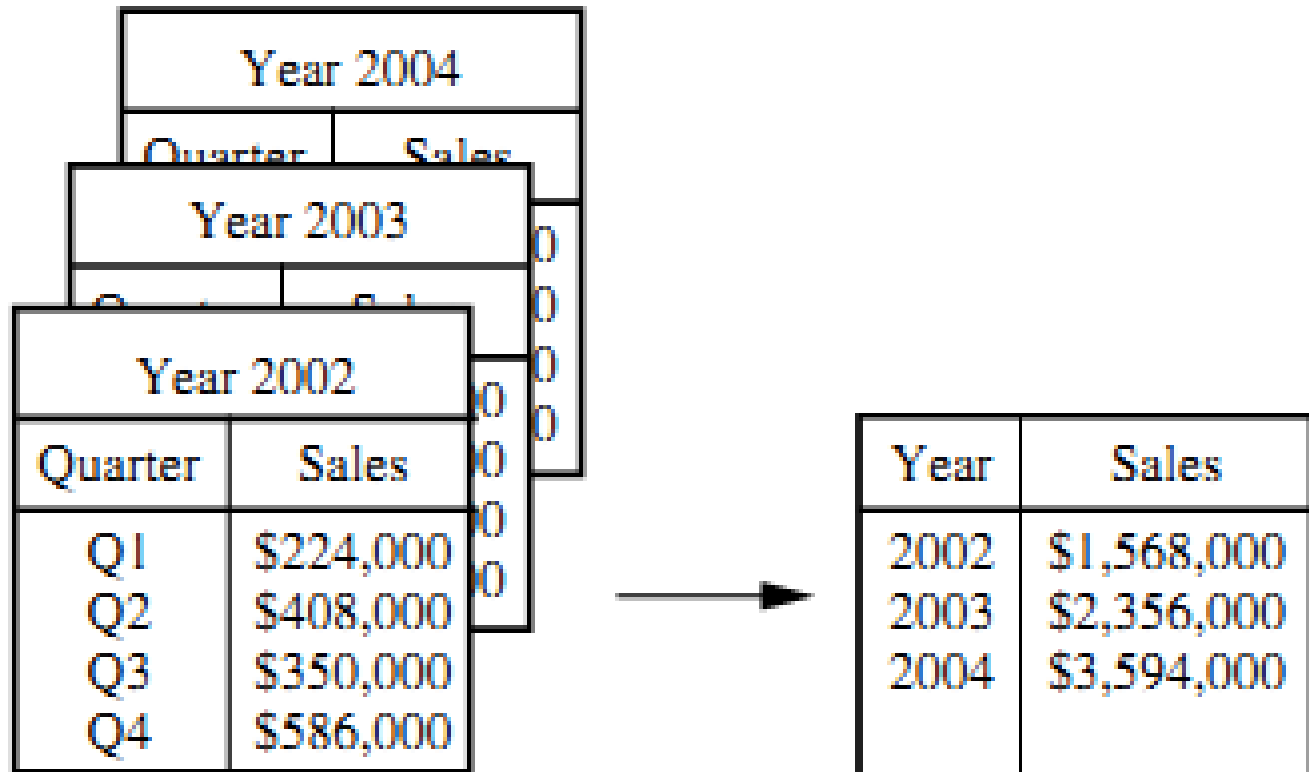
- Các phương pháp:
 - Tổng hợp
 - Giảm chiều dữ liệu
 - Nén dữ liệu
 - Giảm số lượng
 - Rời rạc hóa và phân cấp khái niệm



Rút gọn – Tổng hợp (1/3)

- Tổng hợp
 - 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 dữ liệu cấp thấp vào dữ liệu cấp cao :
 - Giảm kích thước tập dữ liệu : giảm số thuộc tính
 - Tăng tính lý thú của mẫu

Rút gọn – Tổng hợp (2/3)



The diagram illustrates the process of data consolidation. On the left, three overlapping tables represent quarterly sales data for the years 2002, 2003, and 2004. The 2002 table is fully visible, showing quarterly sales figures. The 2003 and 2004 tables are partially visible behind it. An arrow points from these three tables to a single consolidated table on the right, which summarizes the total annual sales for each year.

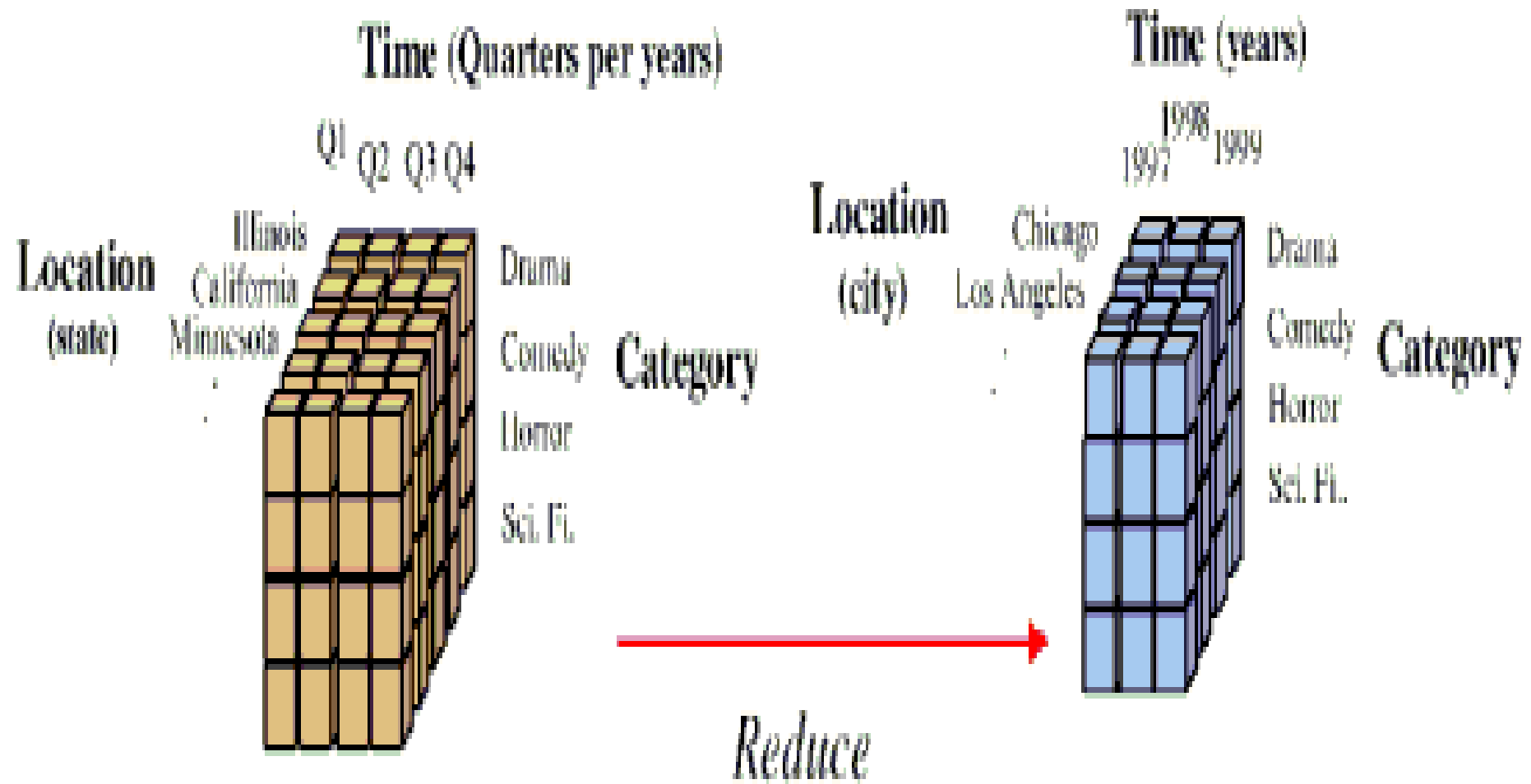
Year 2004	
Quarter	Sales
Q1	\$224,000
Q2	\$408,000
Q3	\$350,000
Q4	\$586,000

Year 2003	
Quarter	Sales
Q1	\$224,000
Q2	\$408,000
Q3	\$350,000
Q4	\$586,000

Year 2002	
Quarter	Sales
Q1	\$224,000
Q2	\$408,000
Q3	\$350,000
Q4	\$586,000

Year	Sales
2002	\$1,568,000
2003	\$2,356,000
2004	\$3,594,000

Rút gọn – Tổng hợp (3/3)



Rút gọn – Giảm chiều (1/6)

- Giảm chiều dữ liệu
 - Chọn lựa đặc trưng (tập con các thuộc tính)
 - Chọn m từ n thuộc tính, $m \leq n$
 - 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

Rút gọn – Giảm chiều (2/6)

- Giảm chiều dữ liệu bằng cách nào?
 - Vét cạn
 - Có 2d tập con thuộc tính của d thuộc tính
 - Độ phức tạp tính toán quá cao
 - PP Heuristic
 - Stepwise forward selection
 - Stepwise backward elimination
 - Kết hợp cả hai
 - Cây quyết định qui nạp

Rút gọn – Giảm chiều (3/6)

– 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 = $\{\}$
 - » $B1 = \{A1\}$
 - » $B2 = \{A1, A4\}$
 - » $B3 = \{A1, A4, A6\}$

Rút gọn – Giảm chiều (4/6)

– 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 $= \{A1, A2, A3, A4, A5, A6\}$
 - » $B1 = \{A1, A3, A4, A5, A6\}$
 - » $B2 = \{A1, A4, A5, A6\}$
 - » $B3 = \{A1, A4, A6\}$

Rút gọn – Giảm chiều (5/6)

– PP Heuristic - Kết hợp

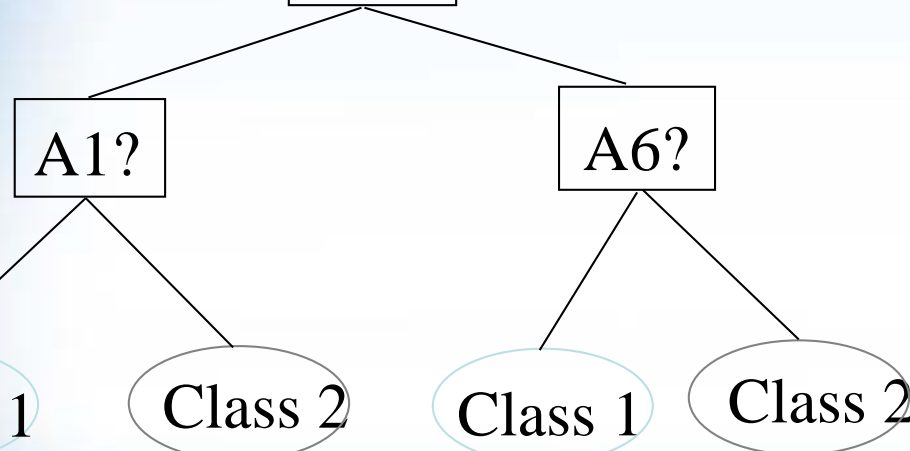
- Đầu tiên : chọn thuộc tính đơn tốt nhất và loại thuộc tính đơn xấu nhất
- Chọn tiếp thuộc tính tốt nhất và loại tiếp thuộc tính xấu nhất trong số còn lại, ...
- Ví dụ : tập thuộc tính ban đầu $\{A1, A2, A3, A4, A5, A6\}$
 - Tập rút gọn ban đầu $= \{A1, A2, A3, A4, A5, A6\}$
 - » $B1 = \{A1, A3, A4,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

Rút gọn – Giảm chiều (6/6)

– 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}

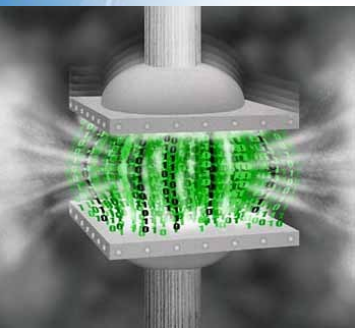
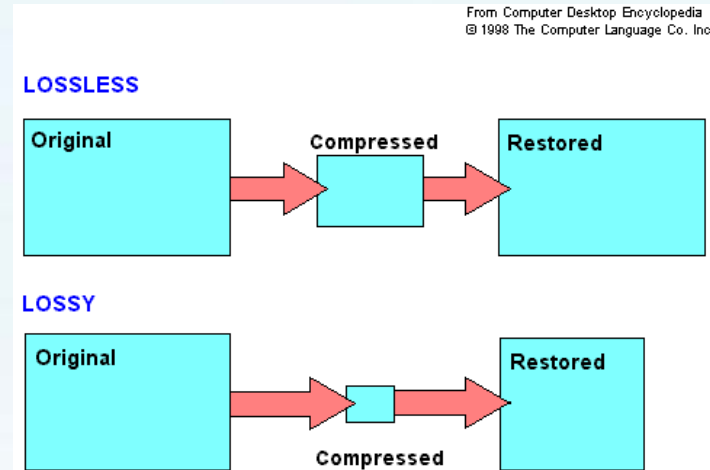
A4 ? \Rightarrow Tập rút gọn = {A1, A4, A6}



Rút gọn – Nén

- Nén dữ liệu:

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



Rút gọn – Giảm số lượng

- Giảm số lượng (numerosity reduction): chọn dạng biểu diễn khác của dữ liệu (“nhỏ hơn”)
- Một số phương pháp:
 - PP tham số:
 - Sử dụng mô hình toán học để lưu các tham số
 - 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

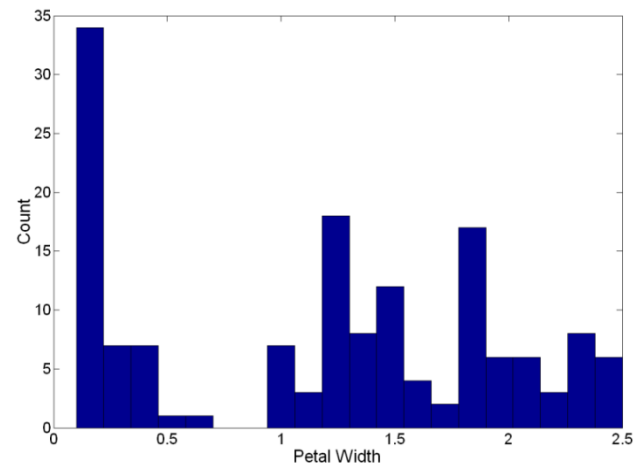
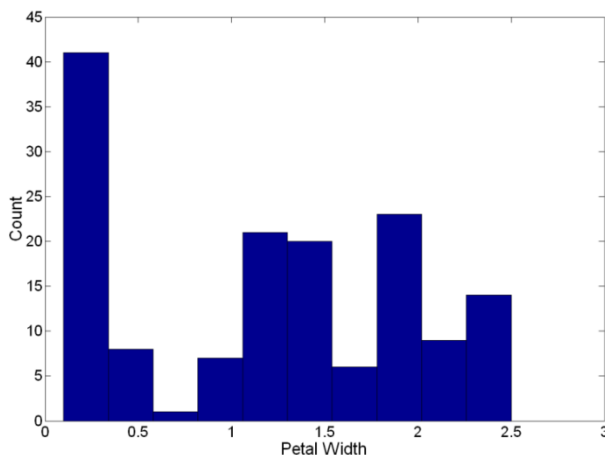
Rút gọn – Giảm số lượng

- 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 : $p(a, b, c, d) = \alpha a^b \beta a^c \chi a^d \delta b^c d$

Rút gọn – Giảm số lượng

– PP biểu đồ (histogram)

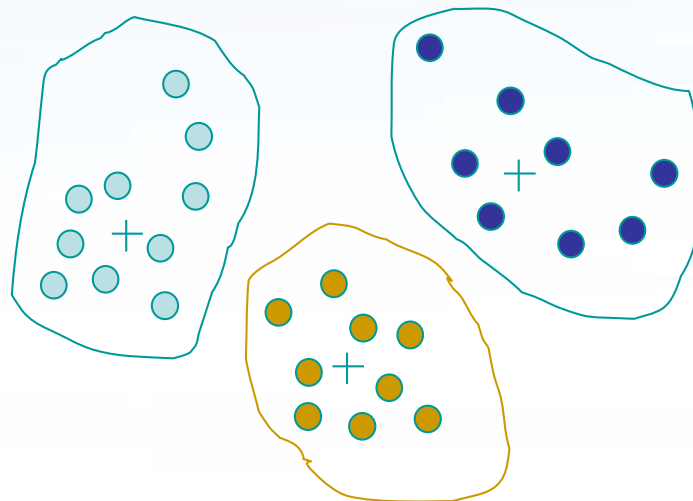
- PP thông dụng để rút gọn DL
- Phân chia DL vào các giỏ và chiều cao của cột là số đối tượng nằm trong mỗi giỏ. Chỉ lưu giá trị trung bình của mỗi giỏ.
- Hình dáng của biểu đồ tùy thuộc vào số lượng giỏ



Rút gọn – Giảm số lượng

– PP gom nhóm

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



Rút gọn – Giảm số lượng

– PP lấy mẫu (sampling)

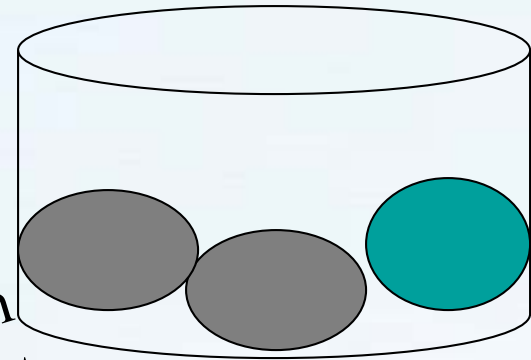
- Dùng tập mẫu ngẫu nhiên nhỏ hơn nhiều để thay thế cho tập dữ liệu lớn.
- PP lấy mẫu ngẫu nhiên không thay thế (SRSWOR)
- PP lấy mẫu ngẫu nhiên có thay thế (SRSWR)
- PP lấy mẫu theo nhóm/phân cấp

Rút gọn – Giảm số lượng

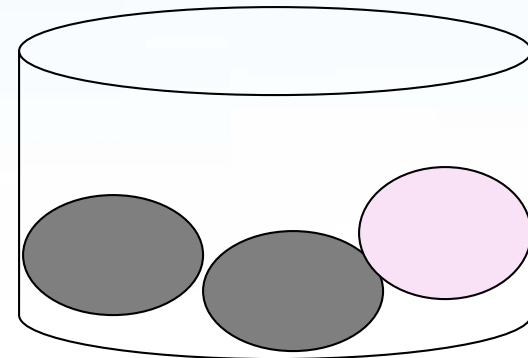


Raw Data

SRSWOR
(simple random
sample without
replacement)

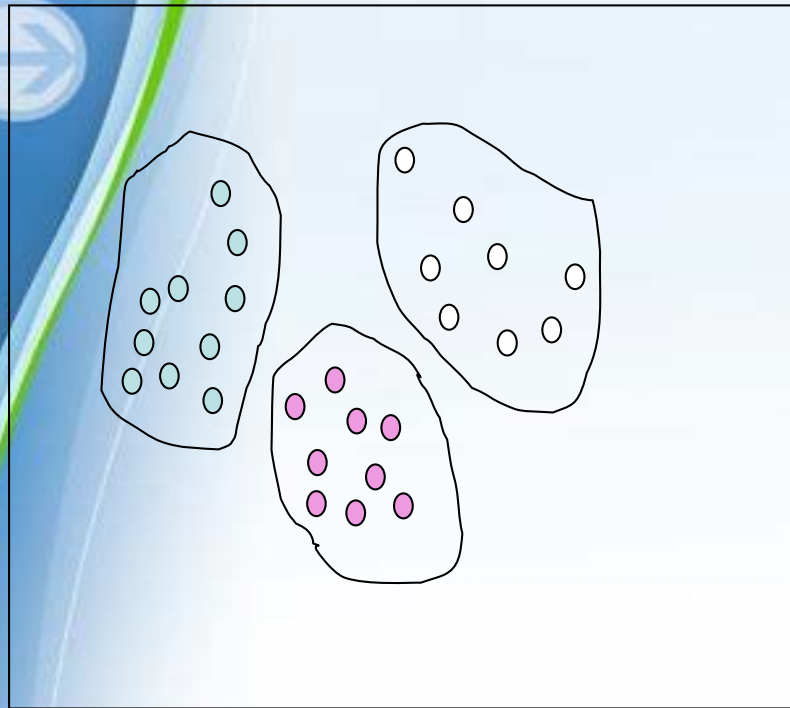


SRSWR

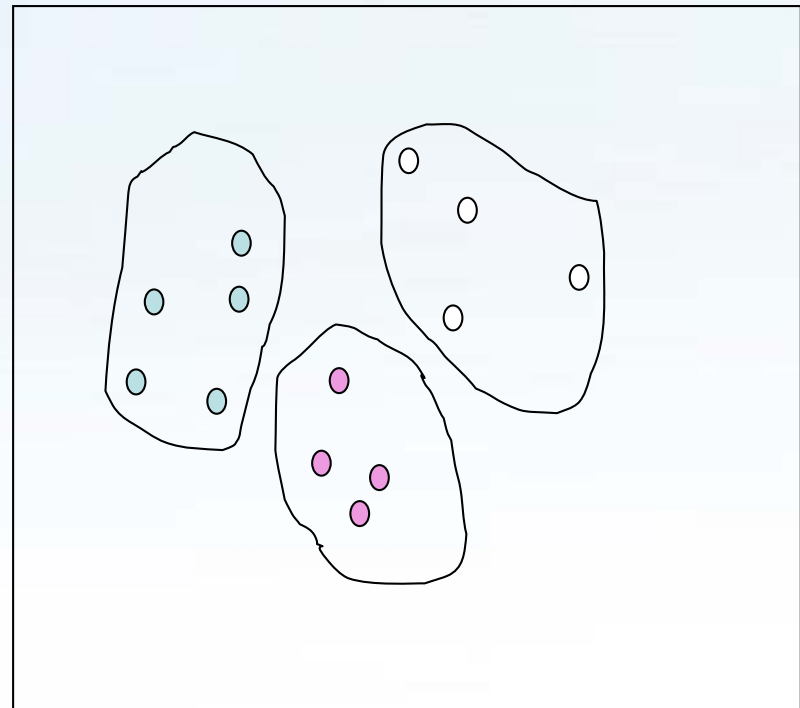


Rút gọn – Giảm số lượng

Raw Data



Cluster/Stratified Sample



Rút gọn – Rời rạc và phân cấp

- 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
 - Dành cho dữ liệu dạng số liên tục.
 - Phương pháp: chia giỏ, phân tích biểu đồ, gom nhóm, rời rạc hoá theo entropy, phân đoạn tự nhiên.

Rút gọn – Rời rạc và phân cấp

- Phân cấp khái niệm:
 - Tập hợp và thay thế khái niệm cấp thấp bằng khái niệm cấp cao hơn.
 - Dành cho dữ liệu dạng phi số: tạo sơ đồ phân cấp.

Outlook	Temperature	Humidity	Windy	Play
sunny	85	85	false	no
sunny	80	90	true	no
overcast	83	78	false	yes
rain	70	96	false	yes
rain	68	80	false	yes
rain	65	70	true	no
overcast	64	65	true	yes
sunny	72	95	false	no
sunny	69	70	false	yes
rain	75	80	false	yes
sunny	75	70	true	yes
overcast	72	90	true	yes
overcast	81	75	false	yes
rain	71	80	true	no

Attributes:

Outlook (overcast, rain, sunny)

Temperature real

Humidity real

Windy (true, false)

Play (yes, no)

Standard
Spreadsheet
Format

Outlook	Outlook	Outlook	Temp	Humidity	Windy	Windy	Play	Play
overcast	rain	sunny			TRUE	FALSE	yes	no
0	0	1	85	85	0	1	1	0
0	0	1	80	90	1	0	0	1
1	0	0	83	78	0	1	1	0
0	1	0	70	96	0	1	1	0
0	1	0	68	80	0	1	1	0
0	1	0	65	70	1	0	0	1
1	0	0	64	65	1	0	1	0
-	-	-	-	-	-	-	-	-
-	-	-	-	-	-	-	-	-

Outlook	Temperature	Humidity	Windy	Play
sunny	85	85	false	no
sunny	80	90	true	no
overcast	83	78	false	yes
rain	70	96	false	yes
rain	68	80	false	yes
rain	65	70	true	no
overcast	64	65	true	yes
sunny	72	95	false	no
sunny	69	70	false	yes
rain	75	80	false	yes
sunny	75	70	true	yes
overcast	72	90	true	yes
overcast	81	75	false	yes
rain	71	80	true	no

Attributes:

Outlook (overcast, rain, sunny)

Temperature real

Humidity real

Windy (true, false)

Play (yes, no)

Standard
Spreadsheet
Format

[illegible]

Rút gọn – Rời rạc và phân cấp

- 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 thành trẻ, trung niên, già

Nội dung

- Tại sao cần chuẩn bị dữ liệu?
- Làm sạch dữ liệu (data cleaning)
- Chọn lọc dữ liệu (data selection)
- Rút gọn dữ liệu (data reduction)
- **Biến đổi dữ liệu (data transformation)**

Biến đổi dữ liệu

- Biến đổi dữ liệu: chuyển đổi dữ liệu thành dạng phù hợp và thuận tiện cho các thuật toán KTDL
- Quá trình biến đổi dữ liệu:
 - Làm trơn (smoothing)
 - Tích hợp (aggregation)
 - Tổng quát hóa (generalization)
 - Chuẩn hóa (normalization)
 - Xây dựng thuộc tính (attribute construction)

Quá trình biến đổi dữ liệu

- Làm trơn: là quá trình bỏ đi nhiễu từ dữ liệu.
- Tích hợp: tóm tắt hay tích hợp dữ liệu.
- Tổng quát hóa: thay thế khái niệm mức thấp bằng các khái niệm mức cao.
- Chuẩn hóa: dữ liệu thuộc tính nên được đưa về phạm vi giá trị nhỏ như từ 0 tới 1.
- Xây dựng thuộc tính: thuộc tính mới được hình thành và thêm vào tập thuộc tính cho trước

Tóm tắt

- Dữ liệu thường thiếu, nhiễu, mâu thuẫn và nhiều chiều. Dữ liệu tốt là chìa khóa tạo ra các mô hình giá trị và đáng tin cậy.
- Chuẩn bị DL gồm các quá trình:
 - Làm sạch
 - Lựa chọn
 - Rút gọn
 - Biến đổi

Câu hỏi cuối bài 1

- Tại sao chuẩn bị DL là công việc cấp thiết và tốn nhiều thời gian?
- Các cách giải quyết vấn đề thiếu giá trị trong các mẫu tin của CSDL?
- 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
 - Khử nhiễu DL trên bằng giá trị TB của giỏ với số giỏ $n=4$. Nhận xét hiệu quả của kỹ thuật này với DL trên.
 - Có thể áp dụng các kỹ thuật nào để khử nhiễu DL ?
 - Dùng DL trên vẽ biểu đồ cùng chiều rộng (equal-width histogram) với độ rộng = 10

Câu hỏi cuối bài 2

- Tại sao cần phải chọn lựa/tích hợp dữ liệu? Hãy nêu quá trình chọn lựa dữ liệu.
- Tại sao cần phải rút gọn dữ liệu? Quá trình rút gọn dữ liệu có thể làm mất mát thông tin hay không? Nếu có hãy nêu cách khắc phục.
- Hãy tìm hiểu các quá trình biến đổi dữ liệu. Cho ví dụ cho từng hướng biến đổi.

Tài liệu tham khảo

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

Hỏi & Đáp

