

BÀI GIẢNG KHOA HỌC DỮ LIỆU

CHƯƠNG: HIỂU BÀI TOÁN, HIỂU DỮ LIỆU VÀ
TIỀN XỬ LÝ DỮ LIỆU

ThS. Vương Thị Hải Yến
Hà Nội, 02/0/2022

Nội dung

■ Hiểu bài toán

- Năm yếu tố để hiểu bài toán

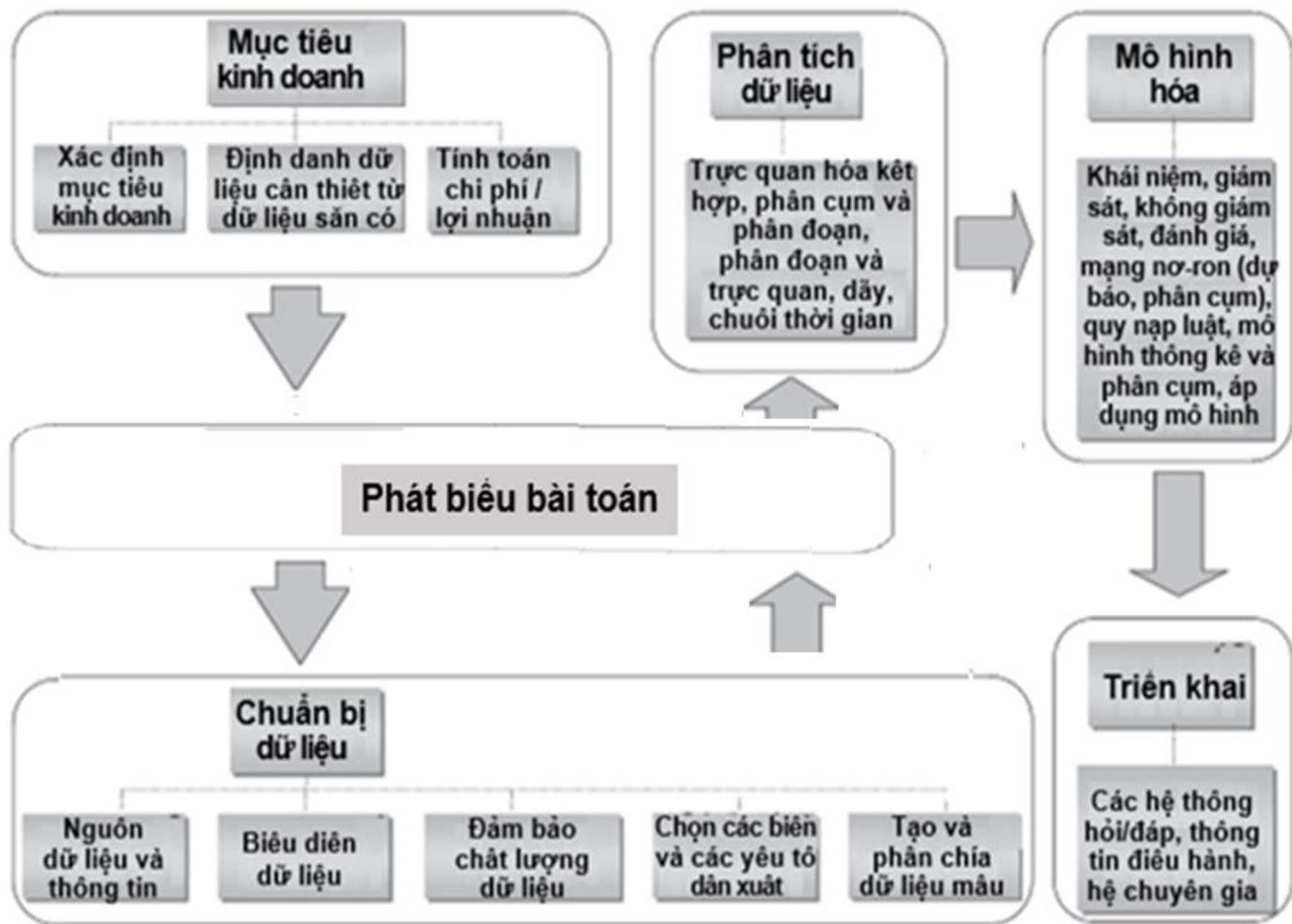
■ Hiểu dữ liệu

- Vai trò của hiểu dữ liệu
- Đối tượng DL và kiểu thuộc tính
- Độ đo tương tự và không tương tự của DL
- Thu thập dữ liệu
- Mô tả thống kê cơ bản của DL
- Trực quan hóa DL
- Đánh giá và lập hồ sơ DL

■ Tiền xử lý dữ liệu

- Vai trò của tiền xử lý dữ liệu
- Làm sạch dữ liệu
- Tích hợp và chuyển dạng dữ liệu
- Rút gọn dữ liệu
- Rời rạc và sinh kiến trúc khái niệm

HIỂU BÀI TOÁN VÀ HIỂU DỮ LIỆU



1. HIỂU BÀI TOÁN: BIẾT ĐƯỢC GÌ?

➤ Đặt vấn đề

- ❖ 5 yếu tố cốt yếu dưới dạng 5 câu hỏi
- ❖ Giải đáp 5 yếu tố này → Đặt được bài toán

➤ Yếu tố 1: Ta đã biết (có) được gì ? Cho INPUT

- ❖ Đây là bước đầu tiên cho mọi trường hợp nghiên cứu
- ❖ Ví dụ 1: Dự báo mục hàng phục vụ bán chéo
 - ❖ Bán chéo (*cross-selling*): bán các sản phẩm bổ sung cho khách hàng hiện tại
 - ❖ Bán sâu (*deep-selling*): tăng tần số hoặc số lượng mua sản phẩm của khách hàng
 - ❖ Bán gia tăng (*up-selling*): bán sản phẩm với số lượng nhiều hơn hoặc giá cao hơn cho khách hàng hiện tại
- ❖ Ví dụ 2: Dự báo khách hàng dịch vụ mạng rời bỏ

Yếu tố 2: Cần quyết định điều gì ?

➤ Nội dung

- ❖ Điều gì thực sự cần phải quyết định
- ❖ Biến quyết định, Đầu ra (Output)
- ❖ Quan trọng: Phân biệt biến đầu ra và biến đầu vào

➤ Trường hợp dễ xác định

- ❖ Ví dụ 1. Bán chéo” Các tập mực hàng đồng xuất hiện cao

➤ Trường hợp khó xác định

- ❖ Ví dụ 2. Dự báo khách hàng dịch vụ mạng rời bỏ: “biến dự báo”, “biến phân lớp” v.v.

Yếu tố 3: Cái gì cố gắng để đạt được

➤ Nội dung

- ❖ Cố tìm gì trong không gian lời giải ?
- ❖ Cái gì cần đạt được ?
- ❖ Hàm mục tiêu, Mô hình mục tiêu
- ❖ Có thể là đa mục tiêu.

➤ Ví dụ

- ❖ Ví dụ 1. Tập con các mục hàng đồng xuất hiện vượt qua một ngưỡng
- ❖ Ví dụ 2. Mô hình dự báo nhận diện lại tốt với dữ liệu kiểm thử

Yếu tố 4: Cái gì cản trở giải bài toán

➤ Nội dung

- ❖ Hạn chế về tài nguyên
- ❖ các ràng buộc

➤ Ví dụ

- ❖ Ví dụ 1. Số mục hàng và giao dịch lớn
- ❖ Ví dụ 2. Dữ liệu mẫu giống nhau song cho kết quả khác nhau

Yếu tố 5: Cái gì tìm hiểu thêm được

➤ Nội dung

- ❖ 4 câu hỏi trên cho xây dựng mô hình
- ❖ Phân tích bối cảnh mô hình rộng hơn: nâng cao ý nghĩa của mô hình. Các khía cạnh phi mô hình

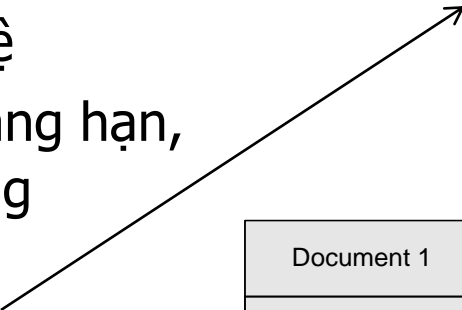
➤ Ví dụ

- ❖ Ví dụ 1. Thay đổi ngưỡng
- ❖ Ví dụ 2. Các phân khúc khách hàng

Kiểu dữ liệu

■ Bản ghi

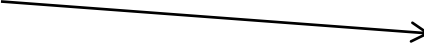
- Bản ghi quan hệ
- Ma trận DL, chẳng hạn, ma trận số, bảng chéo...
- Dữ liệu tài liệu: Tài liệu văn bản dùng vector tần số từ ...
- Dữ liệu giao dịch



	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

■ Đồ thị và mạng

- World Wide Web
- Mạng xã hội và mạng thông tin
- Cấu trúc phân tử



<i>TID</i>	<i>Items</i>
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

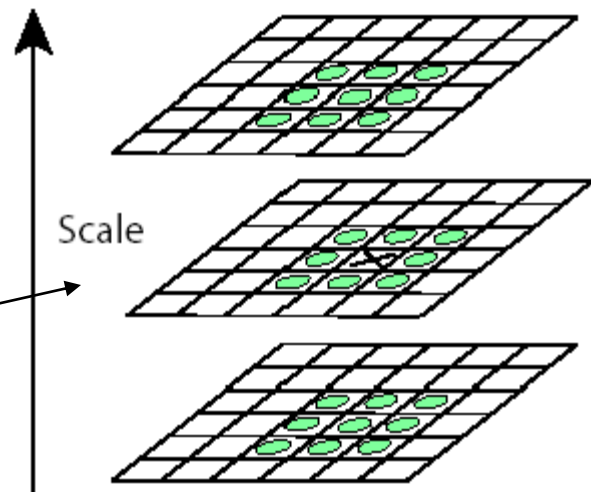
Kiểu dữ liệu

■ Thứ tự

- Dữ liệu thời gian: chuỗi thời gian
- Dữ liệu dãy: dãy giao dịch
- Dữ liệu dãy gene

■ Không gian, ảnh và đa phương tiện:

- DL không gian: bản đồ
- Dữ liệu ảnh,
- Dữ liệu Video: dãy các ảnh
- Dữ liệu audio



Đặc trưng quan trọng của DL có cấu trúc

- Kích thước
 - Tai họa của kích thước lớn
- Thừa
 - Chỉ mang tính hiện diện
- Phân tích
 - Mẫu phụ thuộc quy mô
- Phân bố
 - Tập trung và phân tán

Đối tượng dữ liệu

- Tập DL được tạo nên từ các đối tượng DL.
- Mỗi **đối tượng dữ liệu** (data object) trình bày một thực thể.
- Ví dụ:
 - CSDL bán hàng: Khách hàng, mục lưu, doanh số
 - CSDL y tế: bệnh nhân, điều trị
 - CSDL đại học: sinh viên, giáo sư, môn học
- Tên khác: mẫu (*samples*), ví dụ (*examples*), thể hiện (*instances*), điểm DL (*data points*), đối tượng (*objects*), bộ (*tuples*).
- Đối tượng DL được mô tả bằng các thuộc tính (**attributes**)
- Dòng CSDL → đối tượng DL; cột → thuộc tính.

Thuộc tính

- **Thuộc tính_Attribute** (hoặc chiều_dimension, đặc trưng_features, biến_variables): một trường DL biểu diễn một thuộc tính/đặc trưng của một đối tượng DL.
 - Ví dụ, *ChisoKH*, *tên*, *địa chỉ*
- Kiểu:
 - Định danh
 - Nhị phân
 - Số: định lượng
 - Cỡ khoảng
 - Cỡ tỷ lệ

Kiểu thuộc tính

- **Định danh:** lớp, trạng thái, hoặc “tên đồ vật”
 - *Hair_color* = {*auburn, black, blond, brown, grey, red, white*}
 - Tình trạng hôn nhân (marital status), nghề nghiệp (occupation), số ID (ID numbers), mã zip bưu điện (zip codes)
- **Nhị phân**
 - Thuộc tính định danh hai trạng thái (0 và 1)
 - Nhị phân đối xứng: Cả hai kết quả quan trọng như nhau
 - Chẳng hạn, giới tính
 - Nhị phân phi ĐX: kết quả không quan trọng như nhau.
 - Chẳng hạn, kiểm tra y tế (tích cực/tiêu cực)
 - Quy ước: gán 1 cho kết quả quan trọng nhất (chẳng hạn, dương tính HIV)
- **Có thứ tự**
 - Các giá trị có thứ tự mang nghĩa (xếp hạng) nhưng độ lớn các giá trị liên kết: không được biết
 - *Size* = {*small, medium, large*}, grades, army rankings

Kiểu thuộc tính số

- Số lượng (nguyên hay giá trị thực)
- **Khoảng**
 - Được đo theo kích thước các đơn vị cùng kích thước
 - Các giá trị có thứ tự
 - Chẳng hạn, nhiệt độ theo C° hoặc F° , ngày lịch
 - Không làm điểm "true zero-point"
- **Tỷ lệ**
 - **zero-point** vốn có
 - Các giá trị là một thứ bậc của độ đo so với đơn vị đo lường (10 K° là hai lần cao hơn 5 K°).
 - Ví dụ, nhiệt độ theo *Kelvin*, độ dài đếm được, tổng số đếm được, số lượng tiền

Thuộc tính rời rạc và liên tục

■ Thuộc tính rời rạc

- Chỉ có một tập hữu hạn hoặc hữu hạn đếm được các giá trị
 - Chẳng hạn, mã zip, nghề nghiệp hoặc tập các từ trong một tập tài liệu
- Đôi lúc trình bày như các biến nguyên
- Lưu ý: Thuộc tính nhị phân là trường hợp riêng của thuộc tính rời rạc

■ Thuộc tính liên tục

- Có rất nhiều các giá trị thuộc tính
 - Như nhiệt độ, chiều cao, trọng lượng
- Thực tế, giá trị thực chỉ tính và trình bày bằng sử dụng một hữu hạn chữ số
- Thuộc tính liên tục được trình bày phổ biến như biến dấu phẩy động

Tương tự và phân biệt

■ Tương tự

- Độ đo bằng số cho biết hai đối tượng giống nhau ra sao
- Giá trị càng cao khi hai đối tượng càng giống nhau
- Thường thuộc đoạn $[0,1]$

■ Phân biệt-Dissimilarity (như khoảng cách)

- Độ đo bằng số cho biết hai đối tượng khác nhau ra sao
- Càng thấp khi các đối tượng càng giống nhau
- Phân biệt tối thiểu là 0
- Giới hạn trên tùy

■ Gần-Proximity chỉ dẫn tới tương tự hoặc phân biệt

Đo khoảng cách thuộc tính định danh

- Có thể đưa ra ≥ 2 các trạng thái, như “red, yellow, blue, green” (tổng quát hóa thuộc tính nhị phân)
- Phương pháp 1: Đối sánh đơn giản
 - m : lượng đối sánh, p : tổng số lượng biến
$$d(i, j) = \frac{p - m}{p}$$
- Phương pháp 2: Dùng lượng lớn TT nhị phân
 - Tạo một TT nhị phân mới cho mỗi từ M trạng thái định danh

Khoảng cách DL số: KC Minkowski

- *KC Minkowski*: Một độ đo khoảng cách điển hình

$$d(i, j) = \sqrt[h]{|x_{i1} - x_{j1}|^h + |x_{i2} - x_{j2}|^h + \cdots + |x_{ip} - x_{jp}|^h}$$

với $i = (x_{i1}, x_{i2}, \dots, x_{ip})$ và $j = (x_{j1}, x_{j2}, \dots, x_{jp})$ là hai đối tượng DL p-chiều, và h là bậc (KC này còn được gọi là chuẩn L-h)

- Tính chất
 - $d(i, j) > 0$ nếu $i \neq j$, và $d(i, i) = 0$ (xác định dương)
 - $d(i, j) = d(j, i)$ (đối xứng)
 - $d(i, j) \leq d(i, k) + d(k, j)$ (Bất đẳng thức tam giác)
- Một KC bảo đảm 3 tính chất trên là một **metric**

KC Minkowski: Trường hợp đặc biệt

- $h = 1$: khoảng cách **Manhattan** (khối thành thị, chuẩn L_1)
 - Chẳng hạn, khoảng cách Hamming: số lượng bit khác nhau của hai vector nhị phân

$$d(i, j) = |x_{i_1} - x_{j_1}| + |x_{i_2} - x_{j_2}| + \dots + |x_{i_p} - x_{j_p}|$$

- $h = 2$: Khoảng cách O'colit - **Euclidean** (chuẩn L_2)

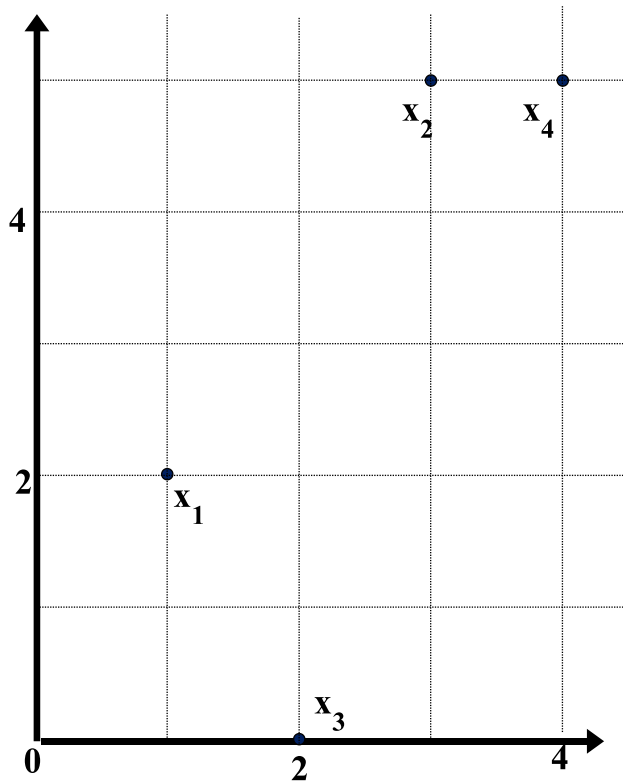
$$d(i, j) = \sqrt{(|x_{i_1} - x_{j_1}|^2 + |x_{i_2} - x_{j_2}|^2 + \dots + |x_{i_p} - x_{j_p}|^2)}$$

- $h \rightarrow \infty$. Khoảng cách **"supremum"** (chuẩn L_{\max} , chuẩn L_∞)
 - Là sự khác biệt cực đại giữa các thành phần (thuộc tính) của các vector

$$d(i, j) = \lim_{h \rightarrow \infty} \left(\sum_{f=1}^p |x_{if} - x_{jf}|^h \right)^{\frac{1}{h}} = \max_f |x_{if} - x_{jf}|$$

Ví dụ: KC Minkowski

DỮ LIỆU	THUỐC TINH 1	THUỐC TINH 2
x_1	1	2
x_2	3	5
x_3	2	0
x_4	4	5



Ma trận phân biệt

Manhattan (L_1)

L	x_1	x_2	x_3	x_4
x_1	0			
x_2	5	0		
x_3	3	6	0	
x_4	6	1	7	0

Euclidean (L_2)

L_2	x_1	x_2	x_3	x_4
x_1	0			
x_2	3.61	0		
x_3	2.24	5.1	0	
x_4	4.24	1	5.39	0

Supremum

L_∞	x_1	x_2	x_3	x_4
x_1	0			
x_2	3	0		
x_3	2	5	0	
x_4	3	1	5	0

Độ tương tự cosine

- Một tài liệu có thể được trình bày bằng hàng nghìn thuộc tính, mỗi ghi nhận tần số của các phần tử (như từ khóa, n-gram) hoặc cụm từ

<i>Document</i>	<i>teamcoach</i>		<i>hockey</i>	<i>baseball</i>	<i>soccer</i>	<i>penalty</i>	<i>score</i>	<i>win</i>	<i>loss</i>	<i>season</i>
Document1	5	0	3	0	2	0	0	2	0	0
Document2	3	0	2	0	1	1	0	1	0	1
Document3	0	7	0	2	1	0	0	3	0	0
Document4	0	1	0	0	1	2	2	0	3	0

- Đối tượng vector khác: đặc trưng gene trong chuỗi phân tử, ...
- Ứng dụng: truy hồi thông tin, phân cấp sinh học, ánh xạ đặc trưng gene, ...
- Độ đo Cosine: d_1 và d_2 : hai two vector (như vector tần suất từ), thì
$$\cos(d_1, d_2) = (d_1 \bullet d_2) / ||d_1|| ||d_2||,$$
với \bullet chỉ tích vector vô hướng, $||d||$: độ dài vector d

Ví dụ: Độ tương tự Cosine

- $\cos(d_1, d_2) = (d_1 \bullet d_2) / (||d_1|| ||d_2||)$,
ở đây \bullet chỉ tích vô hướng, $||d||$: độ dài vector d
- Ví dụ: Tìm độ tương tự giữa hai tài liệu 1 và 2.

$$d_1 = (5, 0, 3, 0, 2, 0, 0, 2, 0, 0)$$

$$d_2 = (3, 0, 2, 0, 1, 1, 0, 1, 0, 1)$$

$$d_1 \bullet d_2 = 5*3 + 0*0 + 3*2 + 0*0 + 2*1 + 0*1 + 0*1 + 2*1 + 0*0 + 0*1 = 25$$

$$||d_1|| = (5*5 + 0*0 + 3*3 + 0*0 + 2*2 + 0*0 + 0*0 + 2*2 + 0*0 + 0*0)^{0.5} = (42)^{0.5} \\ = 6.481$$

$$||d_2|| = (3*3 + 0*0 + 2*2 + 0*0 + 1*1 + 1*1 + 0*0 + 1*1 + 0*0 + 1*1)^{0.5} = (17)^{0.5} \\ = 4.12$$

$$\cos(d_1, d_2) = 0.94$$

Thu thập dữ liệu

- Cách thu thập dữ liệu cần thiết để mô hình hóa Data Acquisition:
 - Trích chọn dữ liệu theo câu hỏi từ CSDL tới tập tin phẳng
 - Ngôn ngữ hỏi bậc cao truy nhập trực tiếp CSDL
 - Kết nối mức thấp để truy nhập trực tiếp CSDL
 - Loại bỏ ràng buộc không gian/thời gian khi di chuyển khối lượng lớn dữ liệu
 - Hỗ trợ việc quản lý và bảo quản dữ liệu tập trung hóa
 - Rút gọn sự tăng không cần thiết của dữ liệu
 - Tạo điều kiện quản trị dữ liệu tốt hơn để đáp ứng mỗi quan tâm đúng đắn

Mô tả thống kê cơ bản của dữ liệu

■ Giá trị kỳ vọng (mean)

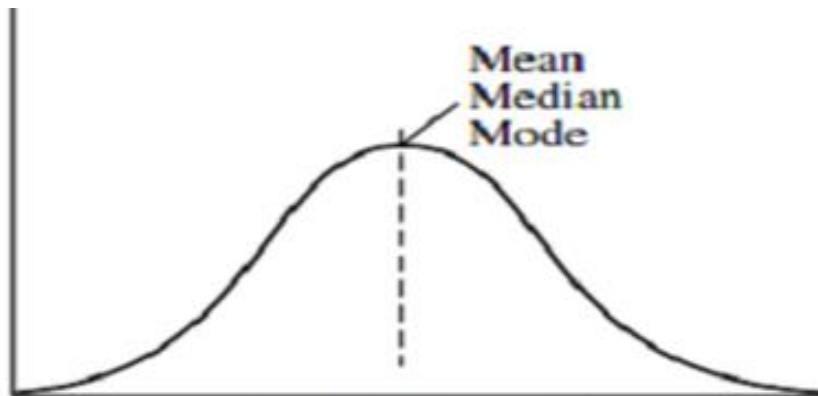
- Xu hướng trung tâm của tập dữ liệu

$$\bar{x} = \frac{\sum_{i=1}^N w_i x_i}{\sum_{i=1}^N w_i} = \frac{w_1 x_1 + w_2 x_2 + \dots + w_N x_N}{w_1 + w_2 + \dots + w_N}$$

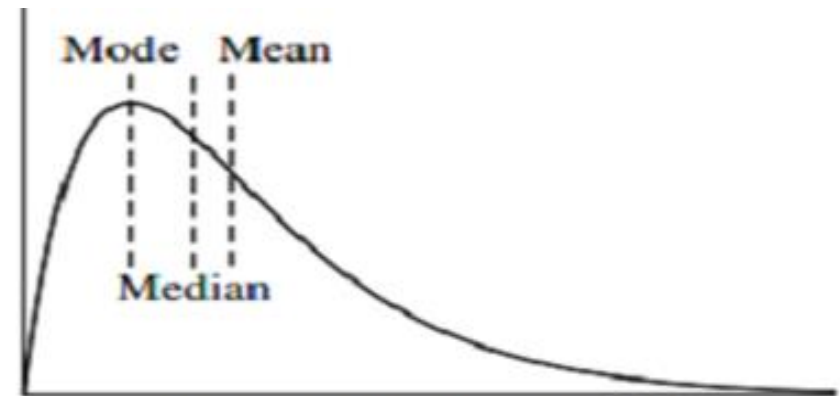
$$\bar{x} = \frac{\sum_{i=1}^N x_i}{N} = \frac{x_1 + x_2 + \dots + x_N}{N}$$

$$\text{median} = L_1 + \left(\frac{N/2 - (\sum \text{freq})l}{\text{freq}_{\text{median}}} \right) \text{width}$$

- Trung vị: (i) xếp lại dãy số, (ii) nếu dãy có $2k+1$ số thì lấy giá trị số thứ $k+1$, nếu có $2k$ số thì trung bình số thứ k và số thứ $k+1$.
- Mode: Tập con dữ liệu xuất hiện với tần số cao nhất. unimodal, bimodal, trimodal, v.v.



(a) symmetric data

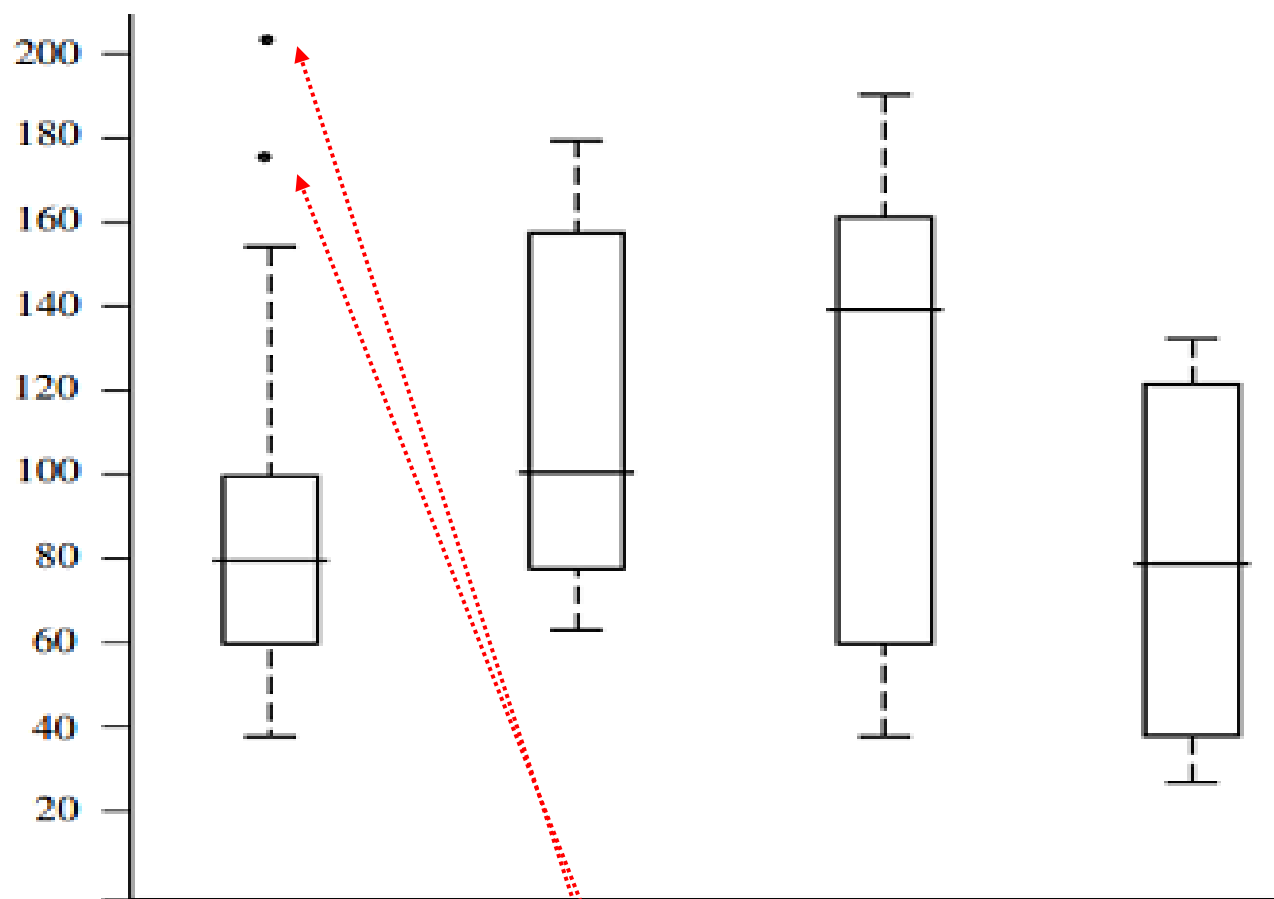


(b) positively skewed data

Một số độ đo thống kê

- Độ lệch chuẩn (Standard deviation)
 - Phân bố dữ liệu xung quanh kỳ vọng
- Cực tiểu (Minimum) và Cực đại (Maximum)
 - Giá trị nhỏ nhất và Giá trị lớn nhất
- Độ đo phân tán
 - [Min, Max]: giá trị k% là giá trị x sao cho $|y \in D: \min \leq y \leq x| / |y \in D| = k\%$
 - Q1=25%, Q2=50%, Q3=75%
 - interquartile range (IQR): Q3-Q1
 - Min, Q1, Median, Q3, Max
- Bảng tần suất (Frequency tables)
 - Phân bố tần suất giá trị của các biến
- Lược đồ (Histograms)
 - Cung cấp kỹ thuật đồ họa biểu diễn tần số giá trị của một biến

Biểu diễn giá trị dữ liệu

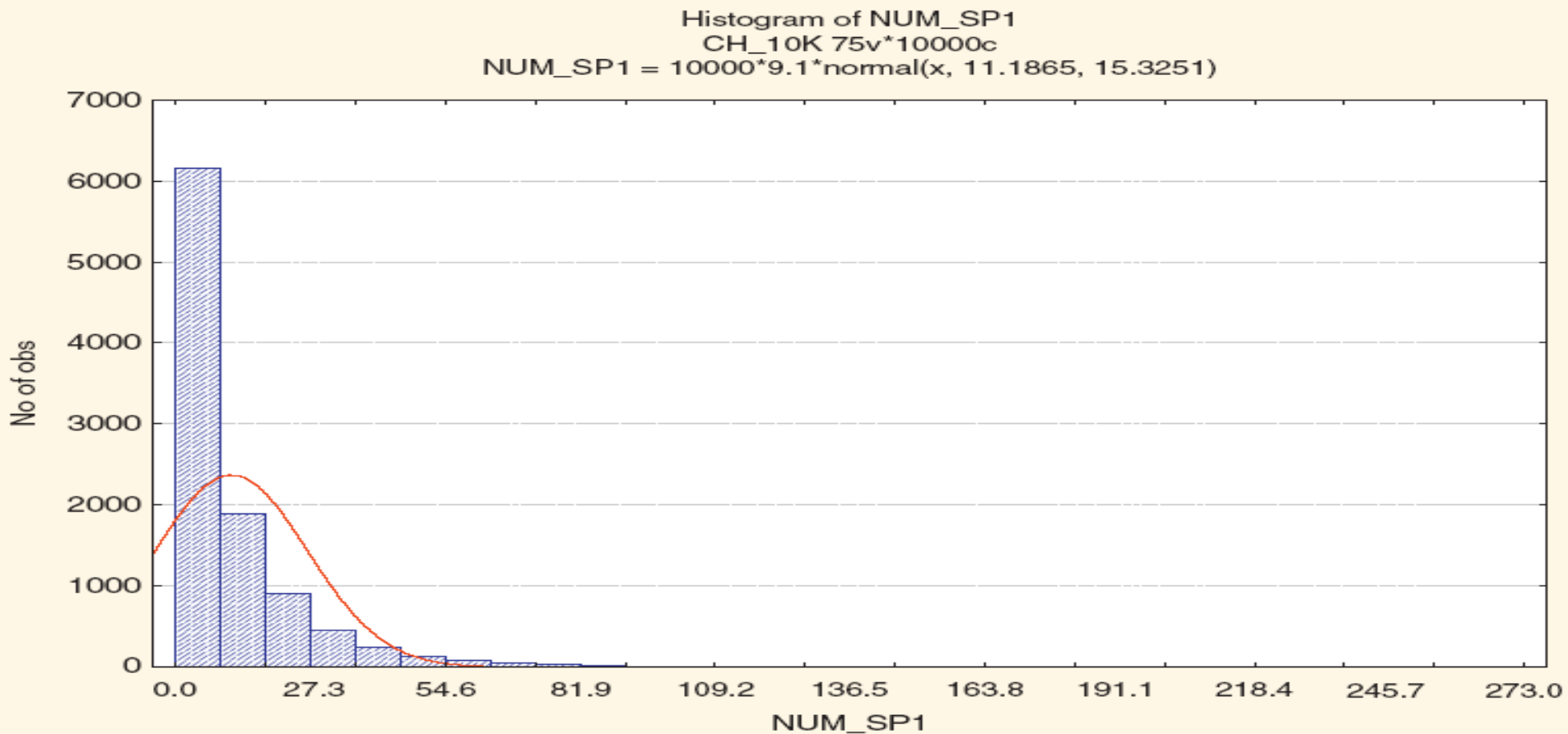


Min, Q1, Median, Q3, Max

$Q1 - 1.5 \cdot IQR$, Q1, Median, Q3, $Q3 + 1.5 \cdot IQR$ nếu nằm ngoài cần kiểm tra là giá trị ngoại lai

Mô tả dữ liệu: trực quan hóa

<i>N</i>	Mean	Min	Max	StDev
10000	9.769700	0.00	454.0000	15.10153



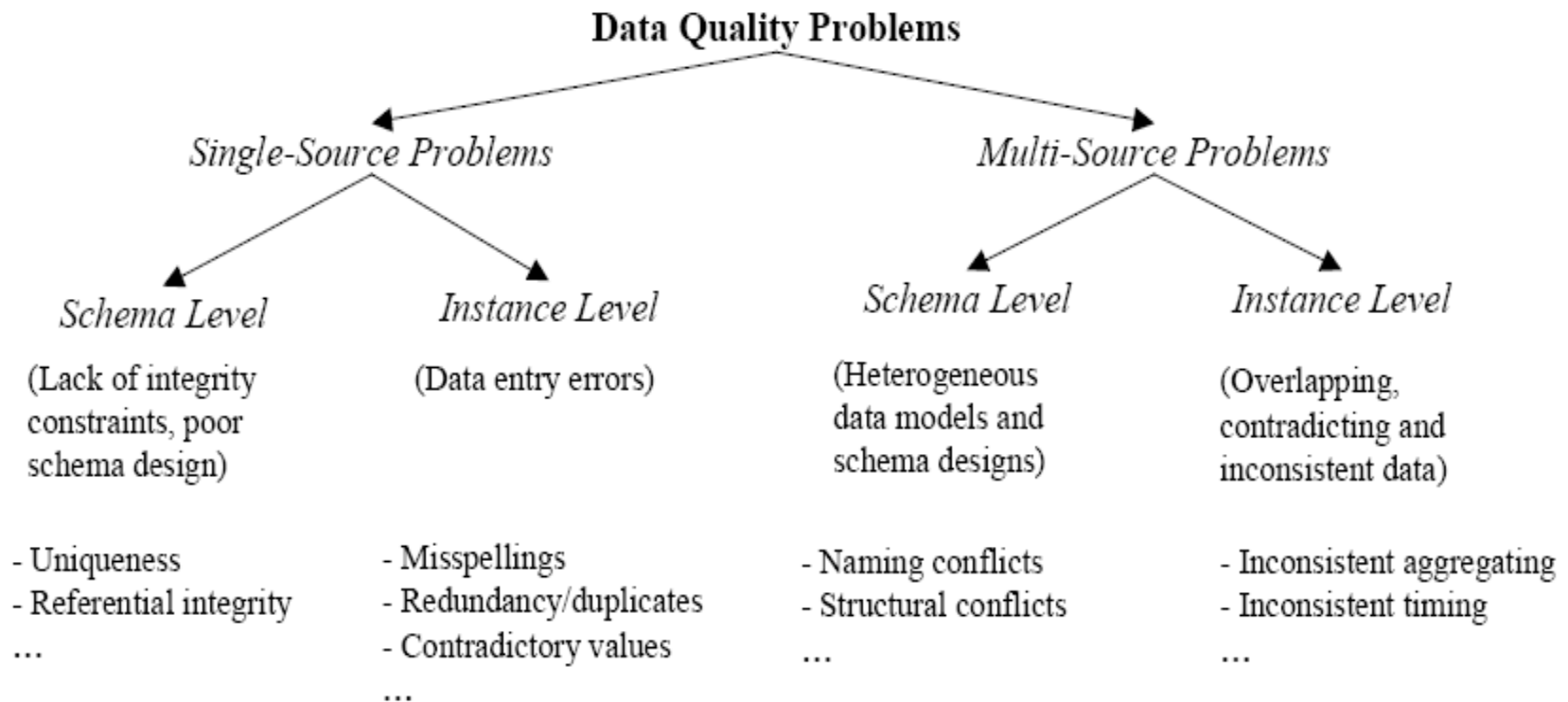
2. Tiền xử lý dữ liệu

- Vai trò của Tiền xử lý dữ liệu
- Làm sạch dữ liệu
- Tích hợp và chuyển dạng dữ liệu
- Rút gọn dữ liệu
- Rời rạc hóa và sinh kiến trúc khái niệm

Vai trò của tiền xử lý

- Không có dữ liệu tốt, không thể có kết quả khai phá tốt!
 - Quyết định chất lượng phải dựa trên dữ liệu chất lượng
 - Chẳng hạn, dữ liệu bội hay thiếu là nguyên nhân thống không chính xác, thậm chí gây hiểu nhầm.
 - Kho dữ liệu cần tích hợp nhất quán của dữ liệu chất lượng
- Phần lớn công việc xây dựng một kho dữ liệu là trích chọn, làm sạch và chuyển đổi dữ liệu —Bill Inmon .
- Dữ liệu có chất lượng cao nếu như phù hợp với mục đích sử dụng trong điều hành, ra quyết định, và lập kế hoạch

Các vấn đề chất lượng dữ liệu [RD00]



- (Thiếu lược đồ toàn vẹn, thiết kế sơ đồ sơ sài) đơn trị, toàn vẹn tham chiếu...
- (Lỗi nhập dữ liệu) sai chính tả, dư thừa/sao, giá trị mâu thuẫn...
- (Mô hình dữ liệu và thiết kế sơ đồ không đồng nhất) xung đột tên, cấu trúc
- (Dữ liệu chồng chéo, mâu thuẫn và không nhất quán) không nhất quán tích hợp và thời gian

[RD00] Erhard Rahm, Hong Hai Do (2000). Data Cleaning: Problems and Current Approaches, *IEEE Data Engineering Bulletin*, **23**(4): 3-13, 2000.

Độ đo đa chiều chất lượng dữ liệu

- Khung đa chiều cấp nhận tốt:
 - Tính chính xác (Accuracy)
 - Tính đầy đủ (Completeness)
 - Tính nhất quán (Consistency)
 - Tính kịp thời (Timeliness)
 - Độ tin cậy (Believability)
 - Giá trị gia tăng (Value added)
 - Biểu diễn được (Interpretability)
 - Tiếp cận được (Accessibility)
- Phân loại bề rộng (Broad categories):
 - Bản chất (intrinsic), ngữ cảnh (contextual), trình diễn (representational), và tiếp cận được (accessibility).

Các bài toán chính trong tiền XL DL

- Làm sạch dữ liệu
 - Điền giá trị thiếu, làm trơn dữ liệu nhiễu, định danh hoặc xóa ngoại lai, và khử tính không nhất quán
- Tích hợp dữ liệu
 - Tích hợp CSDL, khối dữ liệu hoặc tập tin phức
- Chuyển dạng dữ liệu
 - Chuẩn hóa và tổng hợp
- Rút gọn dữ liệu
 - Thu được trình bày thu gọn về kích thước những sản xuất cùng hoặc tương tự kết quả phân tích
- Rời rạc dữ liệu
 - Bộ phận của rút gọn dữ liệu nhưng có độ quan trọng riêng, đặc biệt với dữ liệu số

Một số bài toán cụ thể

- Cách thức làm sạch dữ liệu:
 - Data Cleaning
- Cách thức diễn giải dữ liệu:
 - Data Transformation
- Cách thức nắm bắt giá trị thiếu:
 - Data Imputation
- Trọng số của các trường hợp:
 - Data Weighting and Balancing
- Xử lý dữ liệu ngoại lai và không mong muốn khác:
 - Data Filtering
- Cách thức nắm bắt dữ liệu thời gian/chuỗi thời gian:
 - Data Abstraction
- Cách thức rút gọn dữ liệu để dùng: Data Reduction
 - Bản ghi : Data Sampling
 - Biến: Dimensionality Reduction
 - Giá trị: Data Discretization
- Cách thức tạo biến mới: Data Derivation

Làm sạch dữ liệu

- Nguyên lý chất lượng dữ liệu cần được áp dụng ở mọi giai đoạn quá trình quản lý dữ liệu (nắm giữ, số hóa, lưu trữ, phân tích, trình bày và sử dụng).
 - hai vấn đề cốt lõi để cải thiện chất lượng - phòng ngừa và chỉnh sửa
 - Phòng ngừa liên quan chặt chẽ với thu thập và nhập dữ liệu vào CSDL.
 - Tăng cường phòng ngừa lỗi, vẫn/tồn tại sai sót trong bộ dữ liệu lớn (Maletic và Marcus 2000) và không thể bỏ qua việc xác nhận và sửa chữa dữ liệu
- Vai trò quan trọng
 - “là một trong ba bài toán lớn nhất của kho dữ liệu”—Ralph Kimball
 - “là bài toán “number one” trong kho dữ liệu”—DCI khảo sát
- Các bài toán thuộc làm sạch dữ liệu
 - Xử lý giá trị thiếu
 - Dữ liệu nhiễu: định danh ngoại lai và làm trơn.
 - Chỉnh sửa dữ liệu không nhất quán
 - Giải quyết tính dư thừa tạo ra sau tích hợp dữ liệu.

Xử lý thiếu giá trị

- Bỏ qua bản ghi có giá trị thiếu:
 - Thường làm khi thiếu nhãn phân lớp (giả sử bài toán phân lớp)
 - không hiệu quả khi tỷ lệ số lượng giá trị thiếu lớn (bản giám sát)
- Điền giá trị thiếu bằng tay:
 - tẻ nhạt
 - tính khả thi
- Điền giá trị tự động:
 - Hằng toàn cục: chẳng hạn như “chưa biết - unknown”, có phải một lớp mới
 - Trung bình giá trị thuộc tính các bản ghi hiện có
 - Trung bình giá trị thuộc tính các bản ghi cùng lớp: tinh hơn
 - **Giá trị có khả năng nhất: dựa trên suy luận như công thức Bayes hoặc cây quyết định**

Dữ liệu nhiễu

- Nhiễu:
 - Lỗi ngẫu nhiên
 - Biến dạng của một biến đo được
- Giá trị không chính xác
 - Lỗi do thiết bị thu thập dữ liệu
 - Vấn đề nhập dữ liệu: người dùng hoặc máy có thể sai
 - Vấn đề truyền dữ liệu: sai từ thiết bị gửi/nhận/truyền
 - Hạn chế của công nghệ: ví dụ, phần mềm có thể xử lý không đúng
 - Thiết nhất quán khi đặt tên: cũng một tên song cách viết khác nhau
- Các vấn đề dữ liệu khác yêu cầu làm sạch dữ liệu
 - Bội bản ghi
 - Dữ liệu không đầy đủ
 - Dữ liệu không nhất quán

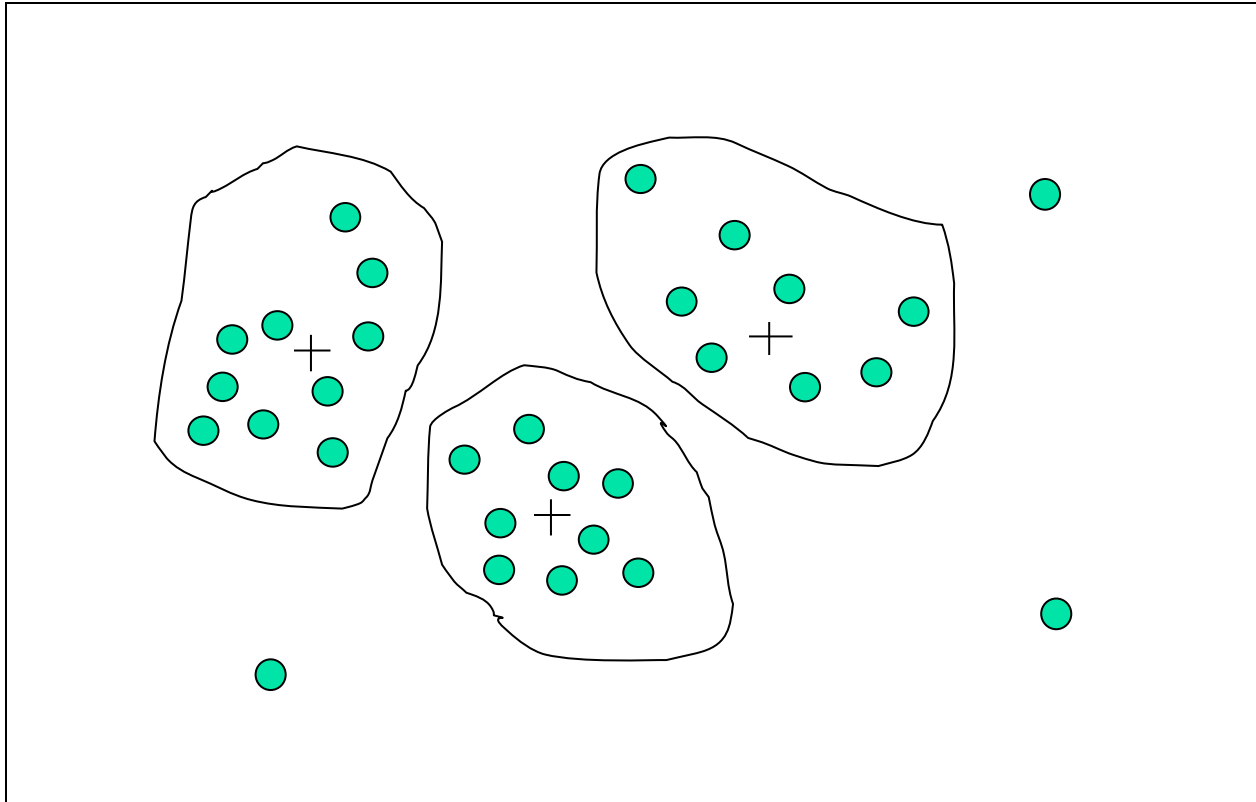
Xử lý dữ liệu nhiễu

- Phương pháp đóng thùng (Binning):
 - Sắp dữ liệu tăng và chia “đều” vào các thùng
 - Làm trơn: theo trung bình, theo trung tuyến, theo biên...
- Phân cụm (Clustering)
 - Phát hiện và loại bỏ ngoại lai (outliers)
- Kết hợp kiểm tra máy tính và con người
 - Phát hiện giá trị nghi ngờ để con người kiểm tra (chẳng hạn, đối phó với ngoại lai có thể)
- Hồi quy
 - Làm trơn: ghép dữ liệu theo các hàm hồi quy

P/pháp xếp thùng

- * Data Smoothing
- * Dữ liệu được xếp theo giá: 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34
- * Chia thùng theo chiều sâu:
 - Bin 1: 4, 8, 9, 15
 - Bin 2: 21, 21, 24, 25
 - Bin 3: 26, 28, 29, 34

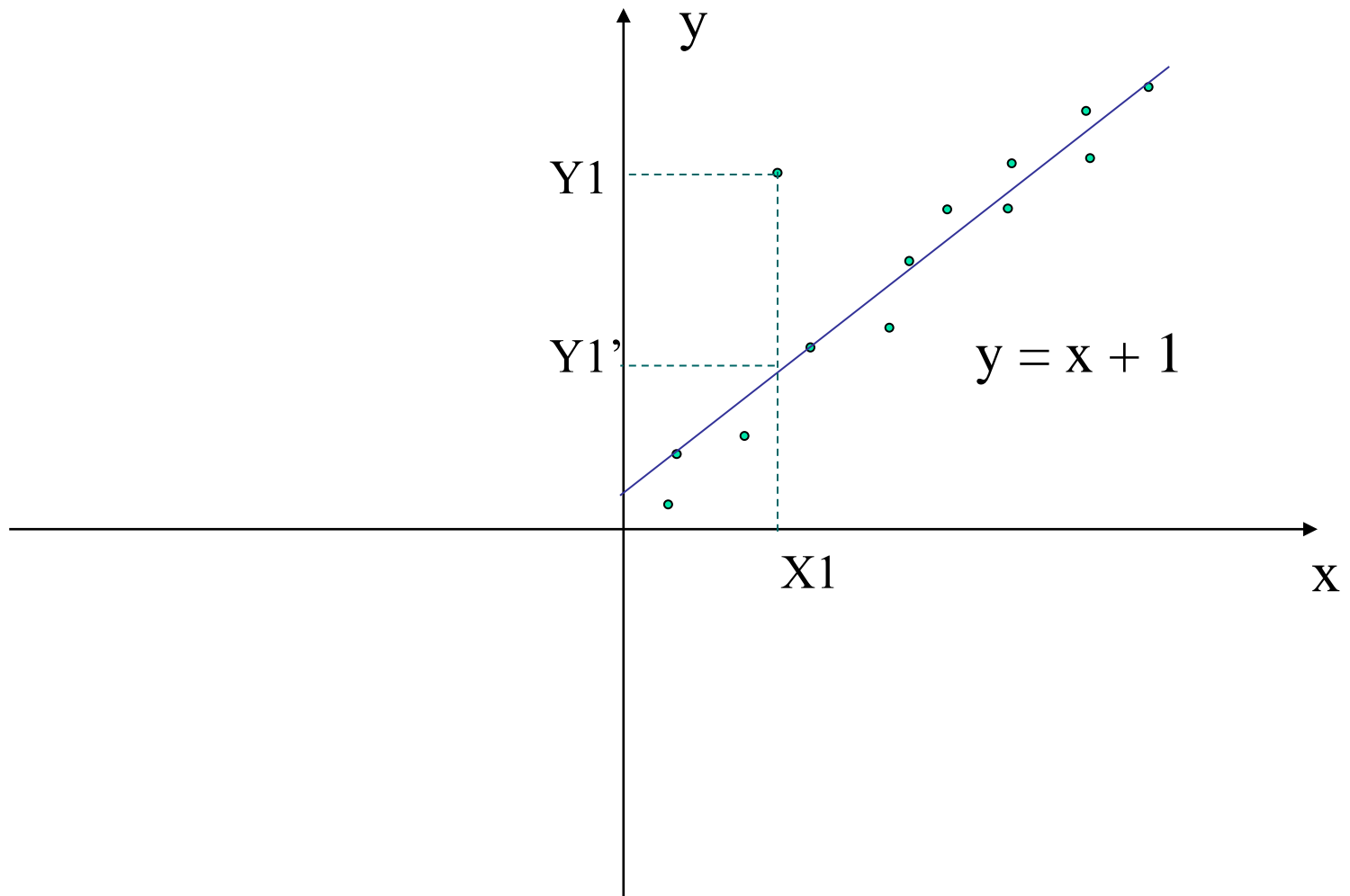
Phân tích cụm (Cluster Analysis)



Cụm: Các phần tử trong cụm là “tương tự nhau”
Làm trơn phần tử trong cụm theo đại diện.

Thuật toán phân cụm: Chương 6.

Hồi quy (Regression)



Tích hợp dữ liệu

- Tích hợp dữ liệu (Data integration):
 - Kết hợp dữ liệu từ nhiều nguồn thành một nguồn lưu trữ chung
- Tích hợp sơ đồ
 - Tích hợp siêu dữ liệu từ các nguồn khác nhau
 - Vấn đề định danh thực thể: xác định thực thể thực tế từ nguồn dữ liệu phức, chẳng hạn, $A.cust-id \equiv B.cust-#$
- Phát hiện và giải quyết vấn đề thiết nhất quá dữ liệu
 - Cùng một thực thể thực sự: giá trị thuộc tính các nguồn khác nhau là khác nhau
 - Nguyên nhân: trình bày khác nhau, cỡ khác nhau, chẳng hạn, đơn vị quốc tế khác với Anh quốc

Nắm bắt dư thừa trong tích hợp dữ liệu

- Dư thừa dữ liệu: thường có khi tích hợp từ nhiều nguồn khác nhau
 - Một thuộc tính có nhiều tên khác nhau ở các CSDL khác nhau
 - Một thuộc tính: thuộc tính “nguồn gốc” trong CSDL khác, chẳng hạn, doanh thu hàng năm
- Dữ liệu dư thừa có thể được phát hiện khi phân tích tương quan
- Tích hợp cần trọng dữ liệu nguồn phức giúp giảm/tránh dư thừa, thiếu nhất quán và tăng hiệu quả tốc độ và chất lượng

Chuyển dạng dữ liệu

- Làm trơn (Smoothing): loại bỏ nhiễu từ dữ liệu
- Tổng hợp (Aggregation): tóm tắt, xây dựng khối dữ liệu
- Tổng quát hóa (Generalization): leo kiến trúc khái niệm
- Chuẩn hóa (Normalization): thu nhỏ vào miền nhỏ, riêng
 - Chuẩn hóa min-max
 - Chuẩn hóa z-score
 - Chuẩn hóa tỷ lệ thập phân
- Xây dựng thuộc tính/đặc trưng
 - Thuộc tính mới được xây dựng từ các thuộc tính đã có

Chuyển đổi dữ liệu: Chuẩn hóa

- Chuẩn hóa min-max

$$v' = \frac{v - \min_A}{\max_A - \min_A} (\text{new_max}_A - \text{new_min}_A) + \text{new_min}_A$$

- Chuẩn hóa z-score

$$v' = \frac{v - \text{mean}_A}{\text{stand_dev}_A}$$

- Chuẩn hóa tỷ lệ thập phân

$$v' = \frac{v}{10^j} \quad j : \text{số nguyên nhỏ nhất mà } \text{Max}(| \quad |) < 1 \quad v'$$

Chiến lược rút gọn dữ liệu

- Kho dữ liệu chứa tới hàng TB
 - Phân tích/khai phá dữ liệu phức mất thời gian rất dài khi chạy trên tập toàn bộ dữ liệu
- Rút gọn dữ liệu
 - Có được trình bày gọn của tập dữ liệu mà nhỏ hơn nhiều về khối lượng mà sinh ra cùng (hoặc hầu như cùng) kết quả.
- Chiến lược rút gọn dữ liệu
 - Tập hợp khối dữ liệu
 - Giảm đa chiều – loại bỏ thuộc tính không quan trọng
 - Nén dữ liệu
 - Giảm tính số hóa – dữ liệu thành mô hình
 - Rời rạc hóa và sinh cây khái niệm

Rút gọn chiều

- Rút gọn đặc trưng (như., lựa chọn tập con thuộc tính):
 - Lựa chọn tập nhỏ nhất các đặc trưng mà phân bố xác suất của các lớp khác nhau cho giá trị khi cho giá trị của các lớp này gần như phân bố vốn có đã cho giá trị của các đặc trưng
 - Rút gọn # của các mẫu trong tập mẫu dễ dàng hơn để hiểu dữ liệu
- Phương pháp Heuristic (có lực lượng mũ # phép chọn):
 - Không ngoan chọn chuyển tiếp từ phía trước
 - Kết hợp chọn chuyển tiếp và loại bỏ lạc hậu.
 - Rút gọn câu quyết định

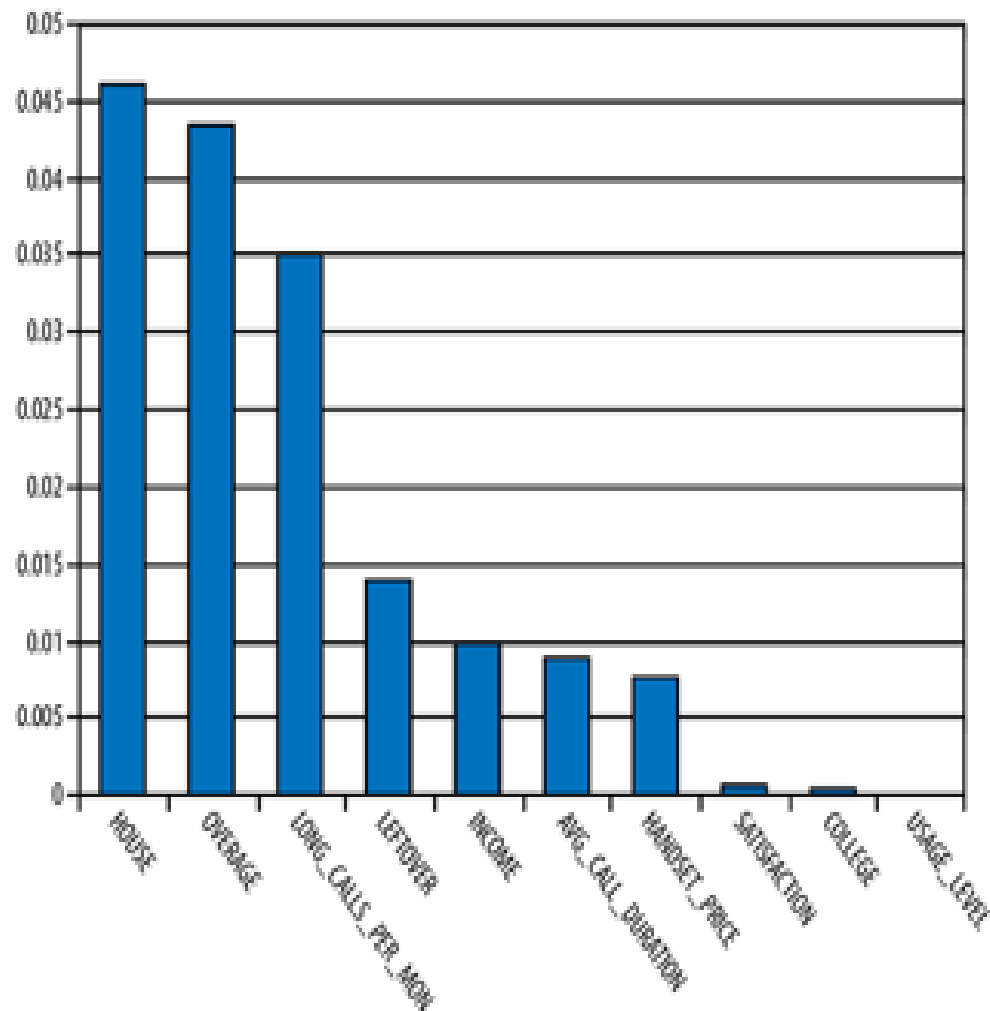
Ví dụ thuộc tính: Bài toán rời bỏ dịch vụ

<i>Biến</i>		<i>Giải thích</i>
COLLEGE	Bằng ĐH	Khách hàng được đào tạo bậc đại học hay không? Biến này nhận giá trị YES (có) và NO (không)
INCOME	Thu nhập	Thu nhập hàng năm là tổng số tiền thu nhập mà khách hàng có trong một năm
OVERAGE	TB phụ thu	Trung bình phụ thu mỗi tháng
LEFTOVER	T/bình phút dư	Trung bình số phút còn dư mỗi tháng
HOUSE	Giá trị nhà	Giá trị ước tính nhà của khách hàng từ điều tra dân số
HANDSET_PRICE		Giá trị điện thoại cầm tay mà khách hàng sử dụng
LONG_CALLS_PER_MONTH		Trung bình số cuộc gọi dài (15 phút trở lên) theo tháng
AVERAGE_CALL_DURATION		Thời gian trung bình một cuộc gọi
TGTB một cuộc gọi		
SATISFACTION	Độ hài lòng	Mức độ hài lòng của khách hàng theo báo cáo
REPORTED_USAGE_LEVEL		Mức sử dụng do người dùng tự đánh giá
LEAVE	<i>(biến mục tiêu)</i>	Khách hàng đã ở lại hay rời mạng? Biến này nhận một trong hai giá trị là STAY (ở lại) và CHURN (rời bỏ)

Công ty điện thoại di động: các thuộc tính như liệt kê
“Lớp” liên quan tới **leave (rời bỏ)**

Rời bỏ dịch vụ

Rank	Info. gain	Attribute name
1	0.0461	HOUSE
2	0.0436	OVERAGE
3	0.0350	LONG_CALLS_PER_MON
4	0.0136	LEFTOVER
5	0.0101	INCOME
6	0.0089	AVG_CALL_DURATION
7	0.0076	HANDSET_PRICE
8	0.0003	SATISFACTION
9	0.000	COLLEGE
10	0.000	USAGE_LEVEL



Độ quan trọng các thuộc tính: Tiến hành lại sau mỗi bước

Nén dữ liệu (Data Compression)

- Nén xâu văn bản
 - Tồn tại lý thuyết phong phú và thuật toán điển hình
 - **Mạnh:** Không tổn thất điển hình
 - **Yếu:** chỉ các thao tác hạn hẹp mà không mở rộng
- Nén Audio/video
 - Nén tổn thất điển hình, với tinh lọc cải tiến
 - Vài trường hợp mảnh tín hiệu nhỏ được tái hợp không cần dựng toàn bộ
- Chuỗi thời gian mà không là audio
 - Ngắt điển hình và thay đổi chậm theo thời gian

Phân tích thành phần chính PCA

- **Principal Component Analysis**
- Cho N vector dữ liệu k -chiều, tìm c ($\leq k$) vector trực giao tốt nhất để trình diễn dữ liệu.
 - Tập dữ liệu gốc được rút gọn thành N vector dữ liệu c chiều: c thành phần chính (chiều được rút gọn).
- Mỗi vector dữ liệu là tổ hợp tuyến tính của các vector thành phần chính.
- Chỉ áp dụng cho dữ liệu số.
- Dùng khi số chiều vector lớn.

Phân cụm

- Phân tập DL thành các cụm, và chỉ cần lưu trữ đại diện của cụm
- Có thể rất hiệu quả nếu DL là được phân cụm mà không chứa dữ liệu “bẩn”
- Có thể phân cụm phân cấp và được lưu trữ trong cấu trúc cây chỉ số đa chiều
- Tồn tại nhiều lựa chọn cho xác định phân cụm và thuật toán phân cụm

Rút gọn mẫu

- **Sampling**
- Cho phép một thuật toán khai phá chạy theo độ phức tạp tựa tuyến tính theo cỡ của DL
- Lựa chọn một tập con **trình diễn** dữ liệu
 - Lấy mẫu ngẫu nhiên đơn giản có hiệu quả rất tồi nếu có DL lệch
- Phát triển các phương pháp lấy mẫu thích nghi
 - Lấy mẫu phân tầng:
 - Xấp xỉ theo phần trăm của mỗi lớp (hoặc bộ phận nhận diện được theo quan tâm) trong CSDL tổng thể
 - Sử dụng kết hợp với dữ liệu lệch
- Lấy mẫu có thể không rút gọn được CSDL.

Rời rạc hóa

- Ba kiểu thuộc tính:
 - Định danh — giá trị từ một tập không có thứ tự
 - Thứ tự — giá trị từ một tập được sắp
 - Liên tục — số thực
- Rời rạc hóa:
 - Chia miền thuộc tính liên tục thành các đoạn
 - Một vài thuật toán phân lớp chỉ chấp nhận thuộc tính phân loại.
 - Rút gọn cỡ DL bằng rời rạc hóa
 - Chuẩn bị cho phân tích tiếp theo

Rời rạc hóa và kiến trúc khái niệm

■ Rời rạc hóa

- Rút gọn số lượng giá trị của thuộc tính liên tục bằng cách chia miền giá trị của thuộc tính thành các đoạn. Nhãn đoạn sau đó được dùng để thay thế giá trị thực.

■ Phân cấp khái niệm

- Rút gọn DL bằng tập hợp và thay thế các khái niệm mức thấp (như giá trị số của thuộc tính tuổi) bằng khái niệm ở mức cao hơn (như trẻ, trung niên, hoặc già)

Rời rạc hóa & kiến trúc khái niệm DL số

- Phân thùng (xem làm trơn khử nhiễu)
- Phân tích sơ đồ (đã giới thiệu)
- Phân tích cụm (đã giới thiệu)
- Rời rạc hóa dựa theo Entropy
- Phân đoạn bằng phân chia tự nhiên

Sinh kiến trúc khái niệm tự động

- Một vài kiến trúc khái niệm có thể được sinh tự động dựa trên phân tích số lượng các giá trị phân biệt theo thuộc tính của tập DL đã cho
 - Thuộc tính có giá trị phân biệt nhất được đặt ở cấp độ phân cấp thấp nhất
 - Lưu ý: Ngoài trừ, các ngày trong tuần, tháng, quý, năm

