



BÀI GIẢNG NHẬP MÔN KHAI PHÁ DỮ LIỆU

CHƯƠNG 2. HIỂU BÀI TOÁN, HIỂU DỮ LIỆU VÀ TIỀN XỬ LÝ DỮ LIỆU

TS. Trần Mai Vũ
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ
ĐẠI HỌC QUỐC GIA HÀ NỘI



Nội dung

■ Hiểu bài toán

- Năm yếu tố để hiểu bài toán

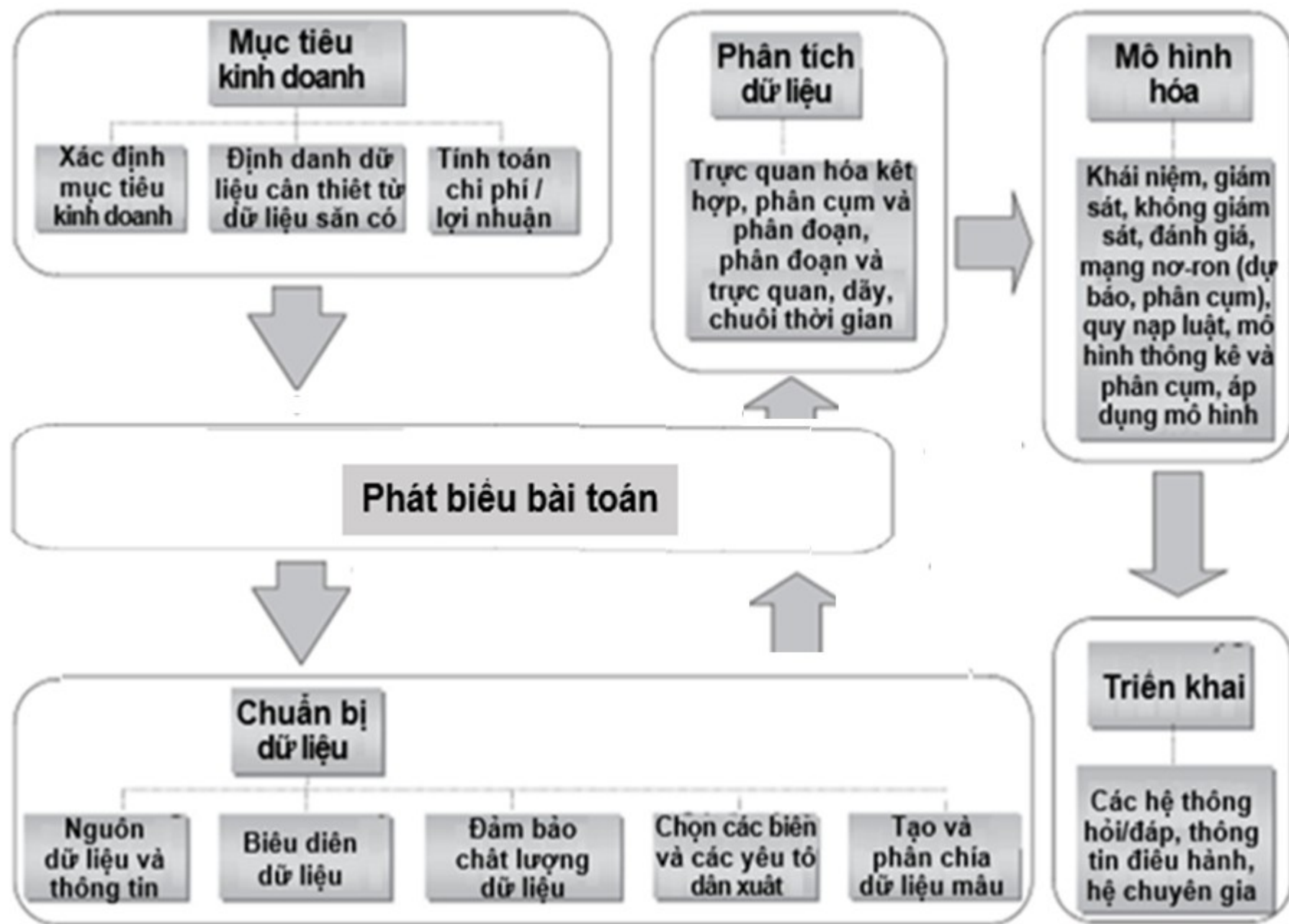
■ Hiểu dữ liệu

- 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.

được

➤ Nội dung

- ❖ Cỗ 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

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

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

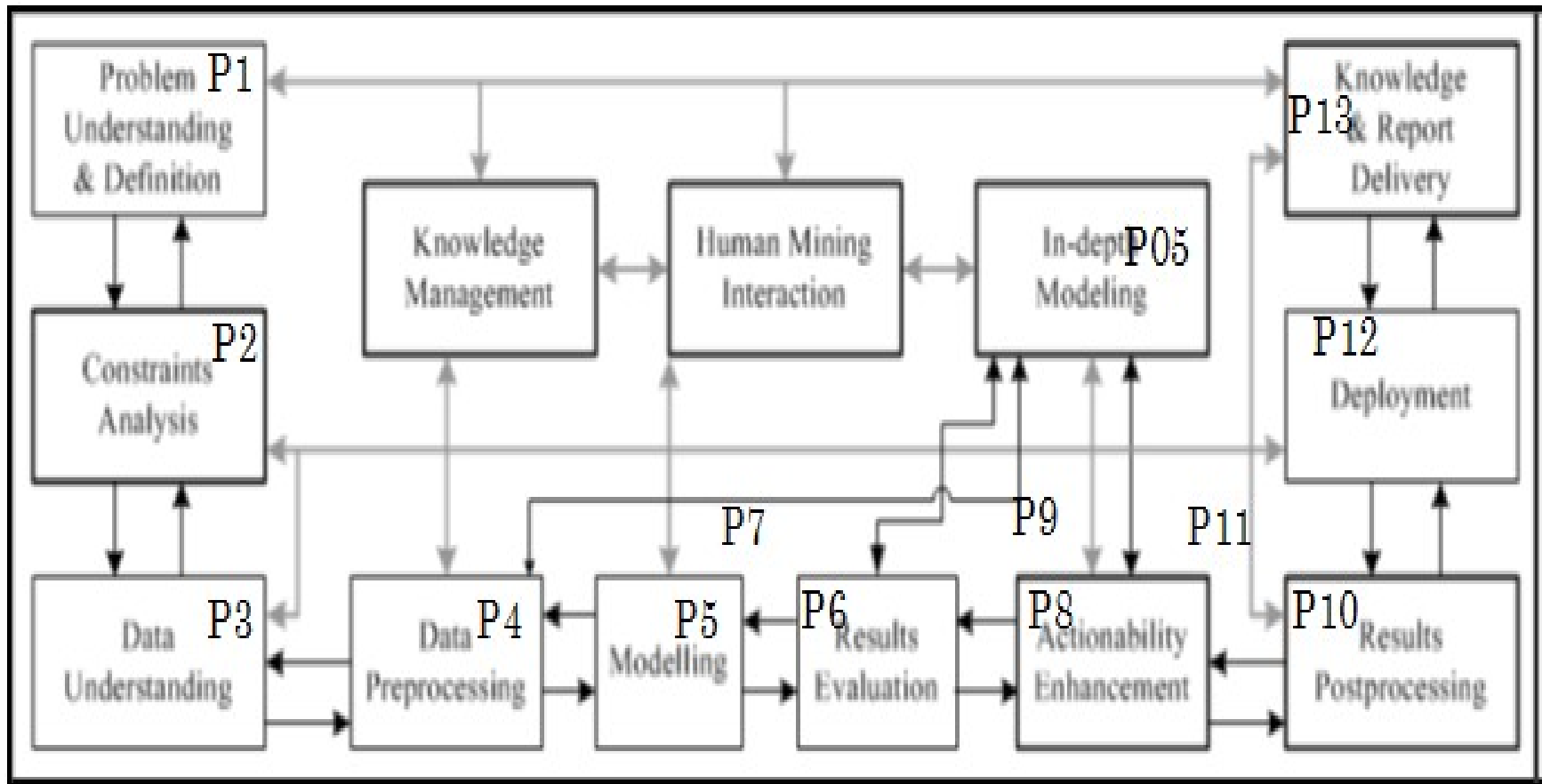
➤ Ví dụ

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

Phiên bản 2006	Phiên bản 2011
Chapter 1 Introduction	Chapter 1 Introduction
Chapter 2 Data Preprocessing	Chapter 2 Getting to Know Your Data
	Chapter 3 Data Preprocessing
Chapter 3 Data Warehouse and OLAP Technology: An Overview	Chapter 4 Data Warehousing and Online Analytical
Chapter 4 Data Cube Computation and Data Generalization	Chapter 5 Data Cube Technology
Chapter 5 Mining Frequent Patterns, Associations, and Correlations	Chapter 6 Mining Frequent Patterns, Associations, and Correlations: Basic Concepts and Methods
Chapter 6 Classification and Prediction	Chapter 7 Advanced Pattern Mining
Chapter 7 Cluster Analysis	Chapter 8 Classification: Basic Concepts
	Chapter 9 Classification: Advanced Methods
	Chapter 10 Cluster Analysis: Basic Concepts and Methods
	Chapter 11 Advanced Cluster Analysis
Chapter 8 Mining Stream, Time-Series, and Sequence Data	Chapter 12 Outlier Detection
Chapter 9 Graph Mining, Social Network Analysis, and Multirelational Data Mining	
Chapter 10 Mining Object, Spatial, Multimedia, Text, and Web Data	
Chapter 11 Applications and Trends in Data Mining	Chapter 13 Data Mining Trends and Research Frontiers

- Thay đổi đáng kể phiên bản 2006 tới 2011
 - Phiên bản 2011 nhấn mạnh **Hiểu dữ liệu !**

Một mô hình KPDL hướng ứng dụng



■ Khai phá DL hướng miền ứng dụng [CYZ10]

- Bước P1 “Hiểu và định nghĩa vấn đề”, Bước P2 “Phân tích ràng buộc”
- Bước P3 “Hiểu dữ liệu”, Bước P4 “Tiền xử lý dữ liệu”



Vấn đề và ràng buộc

■ Vấn đề

- Câu hỏi mục tiêu kinh doanh (Xem chương 1)
- Thường từ 1-3 mục tiêu cụ thể
- Phạm vi dữ liệu liên quan tới câu hỏi
- Đặt bài toán sơ bộ: biến mục tiêu, dữ liệu điều kiện, mô tả sơ bộ ràng buộc dữ liệu điều kiện tới biến mục tiêu

■ Phân tích ràng buộc

- Ràng buộc kinh doanh: Làm rõ hơn mối liên quan giữa dữ liệu với mục tiêu kinh doanh
- Ràng buộc nội tại: Ràng buộc dữ liệu về kiểu, ràng buộc liên quan 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	n	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

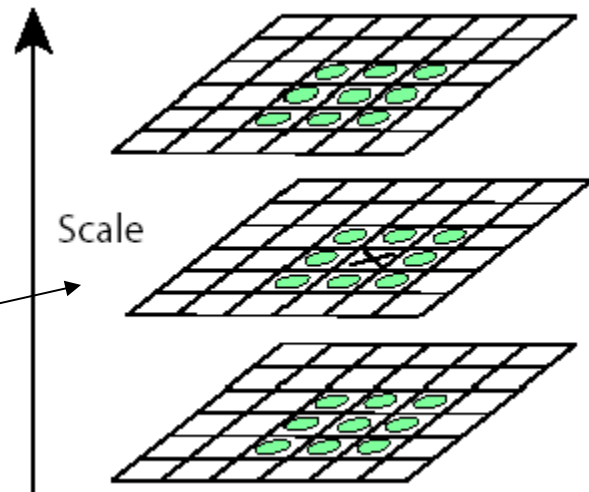
<i>TID</i>	<i>Items</i>
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

■ 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 ($10 K^{\circ}$ là hai lần cao hơn $5 K^{\circ}$).
 - 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



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
- Phương pháp 2: Dùng lượng lớn TT nhị phân
 - Tạo một TT nhị phân mới cho mỗi từ M trạng thái định danh

Đo khoảng cách thuộc tính nhị phân

- Bảng kê cho dữ liệu nhị phân
- Đo khoảng cách các biến nhị phân đối xứng:
- Đo khoảng cách các biến nhị phân không đối xứng:
- Hệ số Jaccard (đo tương tự cho các biến nhị phân không ĐX):

	1	0	sum
1	q	r	$q + r$
0	s	t	$s + t$
sum	$q + s$	$r + t$	p

$$d(i, j) = \frac{r + s}{q + r + s + t}$$

$$d(i, j) = \frac{r + s}{q + r + s}$$

$$sim_{Jaccard}(i, j) = \frac{q}{q + r + s}$$

- Chú ý: Hệ số Jaccard giống độ “gắn kết” (coherence):

$$coherence(i, j) = \frac{sup(i, j)}{sup(i) + sup(j) - sup(i, j)} = \frac{q}{(q + r) + (q + s) - q}$$

Phân biệt giữa các biến nhị phân

■ Ví dụ

Name	Gender	Fever	Cough	Test-1	Test-2	Test-3	Test-4
Jack	M	Y	N	P	N	N	N
Mary	F	Y	N	P	N	P	N
Jim	M	Y	P	N	N	N	N

- Giới tính (Gender): thuộc tính nhị phân đối xứng
- Các thuộc tính còn lại: nhị phân phi đối xứng
- Cho giá trị Y và P là 1, và giá trị N là 0:

$$d(jack, mary) = \frac{0 + 1}{2 + 0 + 1} = 0.33$$

$$d(jack, jim) = \frac{1 + 1}{1 + 1 + 1} = 0.67$$

$$d(jim, mary) = \frac{1 + 2}{1 + 1 + 2} = 0.75$$

Chuẩn hóa dữ liệu số

- Z-score: $Z = \frac{X - \mu}{\sigma}$
 - X: DL thô sẽ được chuẩn hóa, μ : trung bình mẫu (kỳ vọng_ của tập số, σ : độ lệch chuẩn
 - Khoảng cách giữa DL thô và kỳ vọng theo đơn vị độ lệch chuẩn
 - Âm (-) khi DL thô nhỏ thua kỳ vọng, “+” khi lớn hơn above
- Một cách khác: Tính độ lệch tuyệt đối trung bình

trong đó

$$s_f = \frac{1}{n}(|x_{1f} - m_f| + |x_{2f} - m_f| + \dots + |x_{nf} - m_f|)$$
$$m_f = \frac{1}{n}(x_{1f} + x_{2f} + \dots + x_{nf}).$$

- Độ chuẩn hóa (z-score):
- Dùng độ lệch tuyệt đối trung bình là mạnh mẽ hơn so với độ lệch chuẩn

Khoảng cách DL số: KC Minkowski

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

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

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)

25 ■ 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

$$d(i, j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \dots + |x_{ip} - x_{jp}|$$

- $h = 2$: Khoảng cáchƠclit - **Euclidean** (chuẩn L_2)

$$d(i, j) = \sqrt{(|x_{i1} - x_{j1}|^2 + |x_{i2} - x_{j2}|^2 + \dots + |x_{ip} - x_{jp}|^2)}$$

- $h \rightarrow \infty$: Khoảng cách **"supremum"** (chuẩn L_{\max} , chuẩn L_∞)

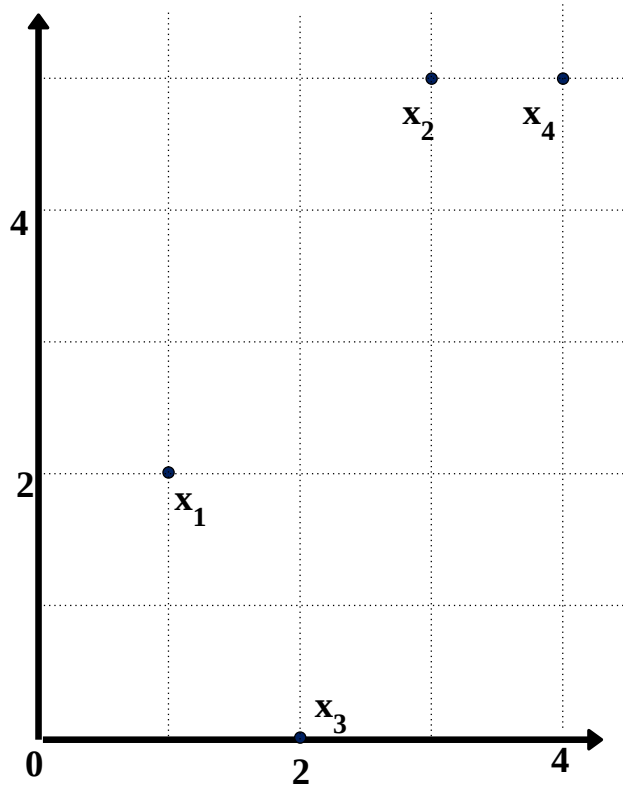
- Là sự khác biệt cực đại giữa các thành phần (thuộc tính) của các vector

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



Ví dụ: KC Minkowski

DỮ LIỆU	THUỘC TÍNH 1	THUỘC TÍNH 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

- Một biến có thứ tự có thể rời rạc hoặc liên tục
- Thứ tự là quan trọng, chẳng hạn như “hạng”
- Có thể coi cỡ-khoảng
 - Thay x_{if} bằng hạng của nó $r_{if} \in \{1, \dots, M_f\}$
 - Ánh xạ phạm vi biến vào $[0, 1]$ khi thay thế đối tượng i thành biến f :

$$z_{if} = \frac{r_{if} - 1}{M_f - 1}$$

- Tính toán độ phân biệt sử dụng phương pháp với biến cỡ-khoảng

Thuộc tính có kiểu pha trộn

- Một CSDL chứa một kiểu thuộc tính
 - Định danh, nhị phân đối xứng, nhị phân phi đối xứng, số, thứ tự
- Có thể sử dụng công thức trọng số để kết hợp tác động của chúng

$$d(i, j) = \frac{\sum_{f=1}^p \delta_{ij}^{(f)} d_{ij}^{(f)}}{\sum_{f=1}^p \delta_{ij}^{(f)}}$$

- f là nhị phân hay định danh:
 $d_{ij}^{(f)} = 0$ nếu $x_{if} = x_{jf}$, hoặc $d_{ij}^{(f)} = 1$ ngược lại
- f là số: sử dụng khoảng cách đã chuẩn hóa
- f là thứ bậc
 - Tính toán hạng r_{if} và
 - Cho z_{if} như cỡ-khoảng

$$z_{if} = \frac{r_{if} - 1}{M_f - 1}$$

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)

hàng chữ từ

Document	team	coach	hockey	baseball	soccer	penalty	score	win	loss	season
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

$$\cos(d_1, d_2) = (d_1 \cdot d_2) / \|d_1\| \|d_2\| ,$$

với \cdot chỉ tích vector vô hướng, $\|d\|$: độ dài vector d



Ví dụ: Độ tương tự Cosine

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

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

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

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

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

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

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

So sánh hai phân bố XS: Phân kỳ KL

- Phân kỳ *Kullback-Leibler* (KD) : Đo sự khác biệt hai phân bố xác suất trên cùng biến x
 - Từ lý thuyết thông tin: liên quan chặt với *entropy tương đối*, *phân kỳ thông tin*, và *thông tin để phân biệt*
- $D_{KL}(p(x), q(x))$: phân kỳ của $q(x)$ từ $p(x)$, đo độ mất mát thông tin khi $q(x)$ được dùng để xấp xỉ $p(x)$

- Dạng rời rạc:
$$D_{KL}(p(x), q(x)) = \sum_{x \in X} q(x) \ln \frac{q(x)}{p(x)}$$

- Phân kỳ KL đo số kỳ vọng các bit yêu cầu thêm để mã hóa ví dụ từ $p(x)$ (phân bố “true”) khi dùng một mã dựa trên $q(x)$, được biểu diễn như một lý thuyết mô hình mô tả hoặc xấp xỉ $p(x)$

- Dạng liên tục:
$$D_{KL}(p(x), q(x)) = \int_{-\infty}^{\infty} q(x) \ln \frac{q(x)}{p(x)} dx$$

- Phân kỳ KL : không là độ đo khoảng cách, không là metric: phi đối xứng, không bảo đảm bất đẳng thức tam giác

$$D_{KL}(p(x), q(x)) = \sum_{x \in X} q(x) \ln \frac{q(x)}{p(x)}$$

- Dựa trên công thức, $D_{KL}(P, Q) \geq 0$ và $D_{KL}(P, Q) = 0$ $P = Q$.
- Xem xét $p = 0$ hoặc $q = 0$
 - $\lim_{q \rightarrow 0} q \log q = 0$
 - Khi $p = 0$ nhưng $q \neq 0$, $D_{KL}(p, q)$ được định nghĩa là ∞ : một sự kiện e là khả năng ($p(e) > 0$), và dự báo q là không thể tuyệt đối ($q(e) = 0$), thì hai phân bố là **khác biệt tuyệt đối**
- Thực tế: P và Q được cung cấp từ phân bố tần suất, không xem xét khả năng của cái không nhìn thấy: **làm trơn** (*smoothing*) là **cần thiết**
- Ví dụ: $P : (a : 3/5, b : 1/5, c : 1/5)$. $Q : (a : 5/9, b : 3/9, d : 1/9)$
 - Đưa vào một hằng số rất nhỏ ϵ , chẳng hạn, $\epsilon = 10^{-3}$
 - Tập mẫu được quan sát trong P , $SP = \{a, b, c\}$, $SQ = \{a, b, d\}$, $SU = \{a, b, c, d\}$
 - Làm trơn, bổ sung ký hiệu thiếu cho mỗi phân bố với xác suất ϵ
 - $P' : (a : 3/5 - \epsilon/3, b : 1/5 - \epsilon/3, c : 1/5 - \epsilon/3, d : \epsilon)$
 - $Q' : (a : 5/9 - \epsilon/3, b : 3/9 - \epsilon/3, c : \epsilon, d : 1/9 - \epsilon/3)$.



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)

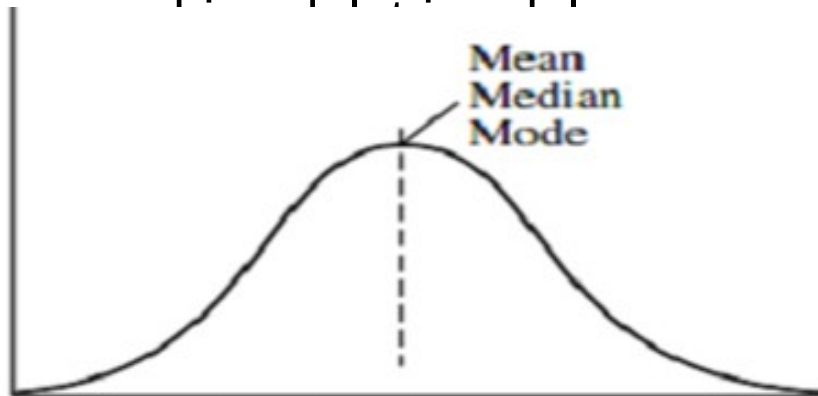
- Xu hướng trung tâm của tập dữ liệu

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

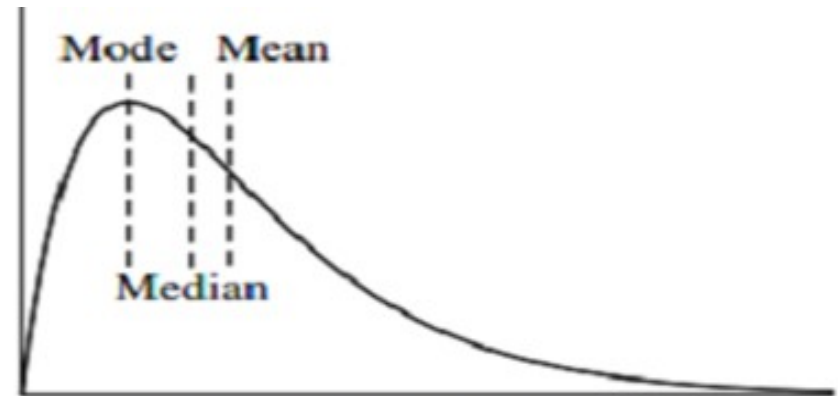
$$\bar{x} = \frac{\sum_{i=1}^N x_i}{N} = \frac{x_1 + x_2 + \dots + x_N}{N}$$

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

- 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,



(a) symmetric data



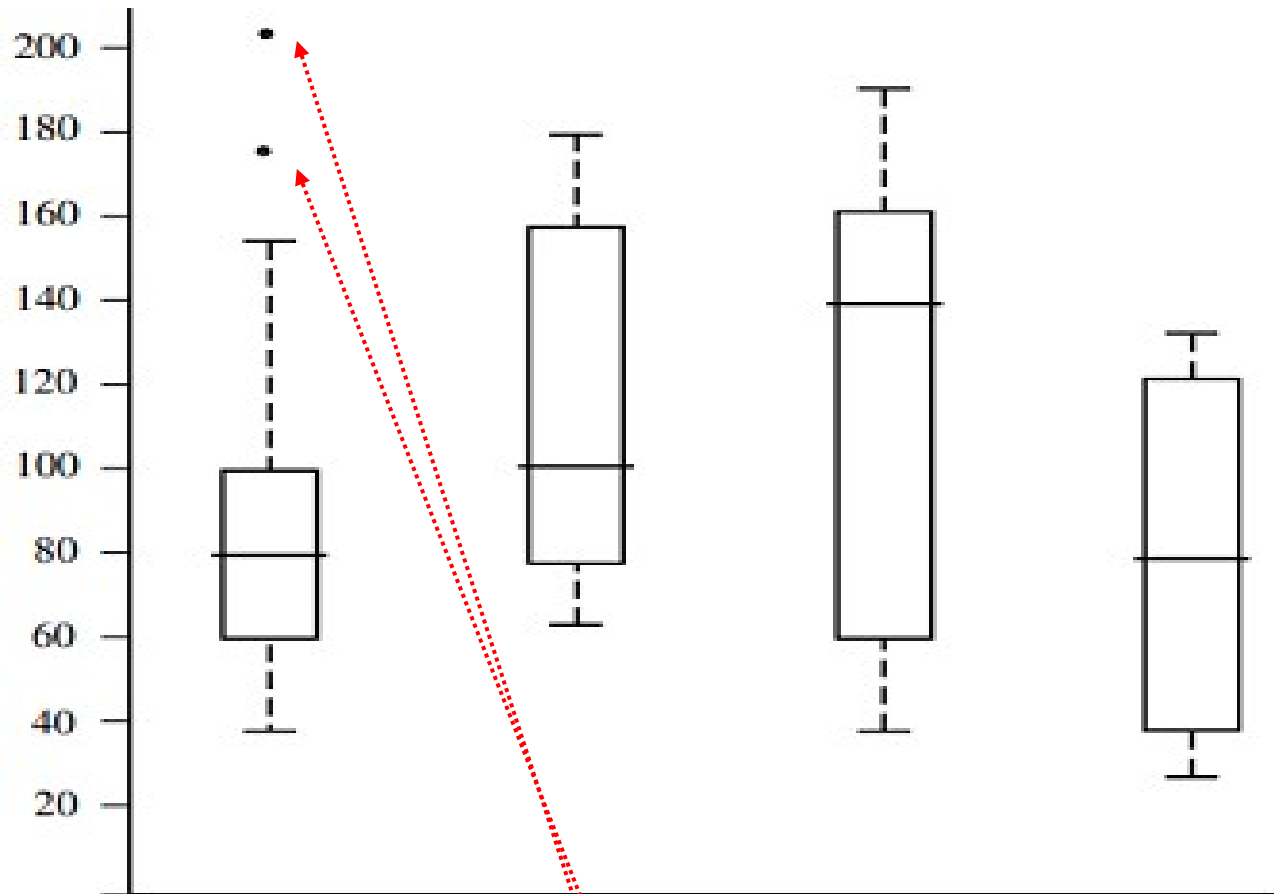
(b) positively skewed data



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 $|y - x|/|y - \text{Min}| = k\%$
 - $Q1=25\%$, $Q2=50\%$, $Q3=75\%$
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



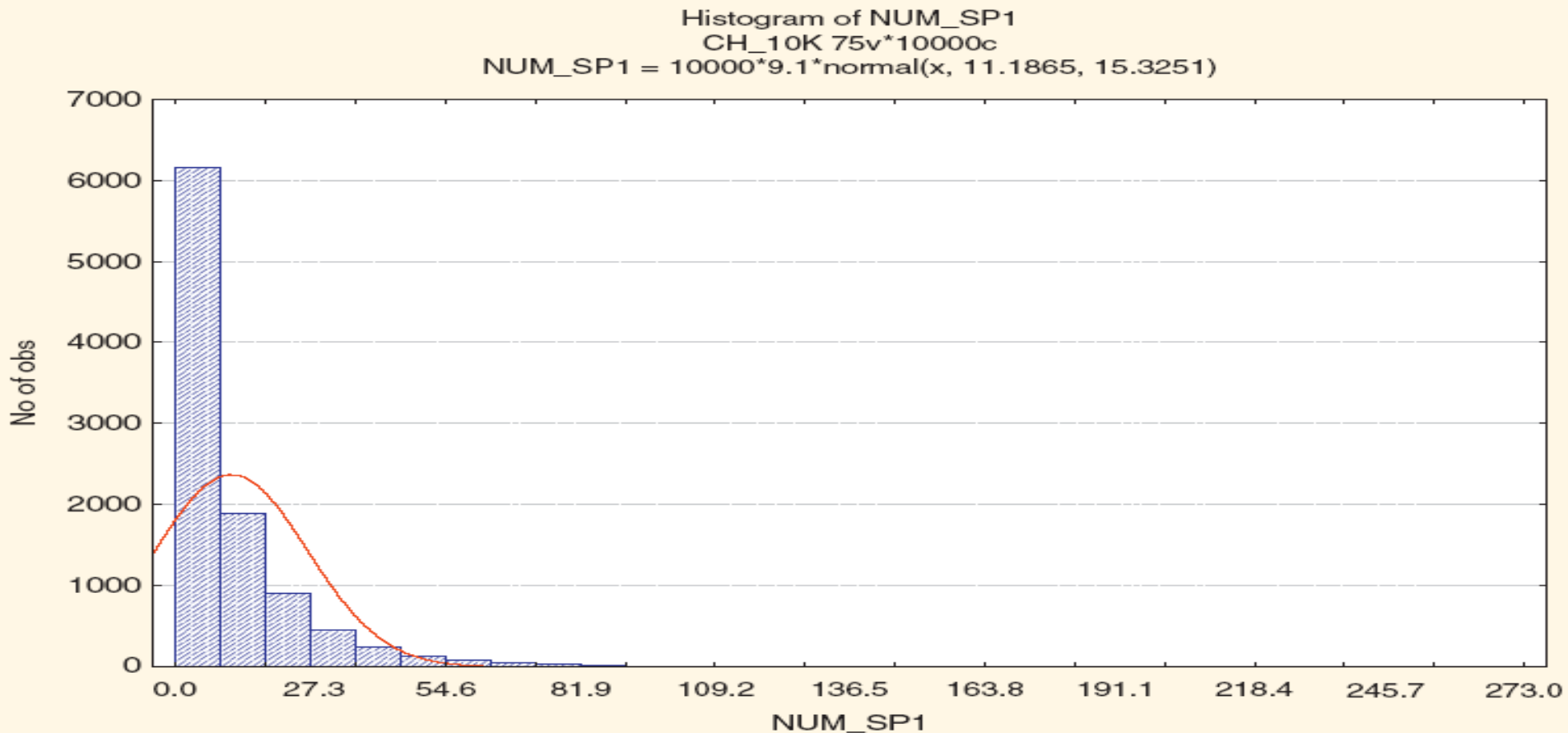
Min, Q1, Median, Q3, Max

$Q1 - 1.5 \cdot IQR$, 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>	<i>Mean</i>	<i>Min</i>	<i>Max</i>	<i>StDev</i>
10000	9.769700	0.00	454.0000	15.10153



Đánh giá và lập hồ sơ dữ liệu

- **Đánh giá dữ liệu**
 - Định vị một vấn đề trong dữ liệu cần giải quyết: Tìm ra và quyết định cách nắm bắt vấn đề
 - Mô tả dữ liệu sẽ làm hiện rõ một số vấn đề
 - Kiểm toán dữ liệu: lập hồ sơ dữ liệu và phân tích ảnh hưởng của dữ liệu chất lượng kém.
- **Lập hồ sơ dữ liệu (cơ sở căn cứ: phân bố dữ liệu)**
 - Tâm của dữ liệu
 - Các ngoại lai tiềm năng bất kỳ
 - Số lượng và phân bố các khoảng trong mọi trường hợp
 - Bất cứ dữ liệu đáng ngờ, như mã thiếu (miscodes), dữ liệu học, dữ liệu test, hoặc chỉ đơn giản dữ liệu rác
 - Những phát hiện nên được trình bày dưới dạng các báo cáo và liệt kê như các mốc quan trọng của kế hoạch



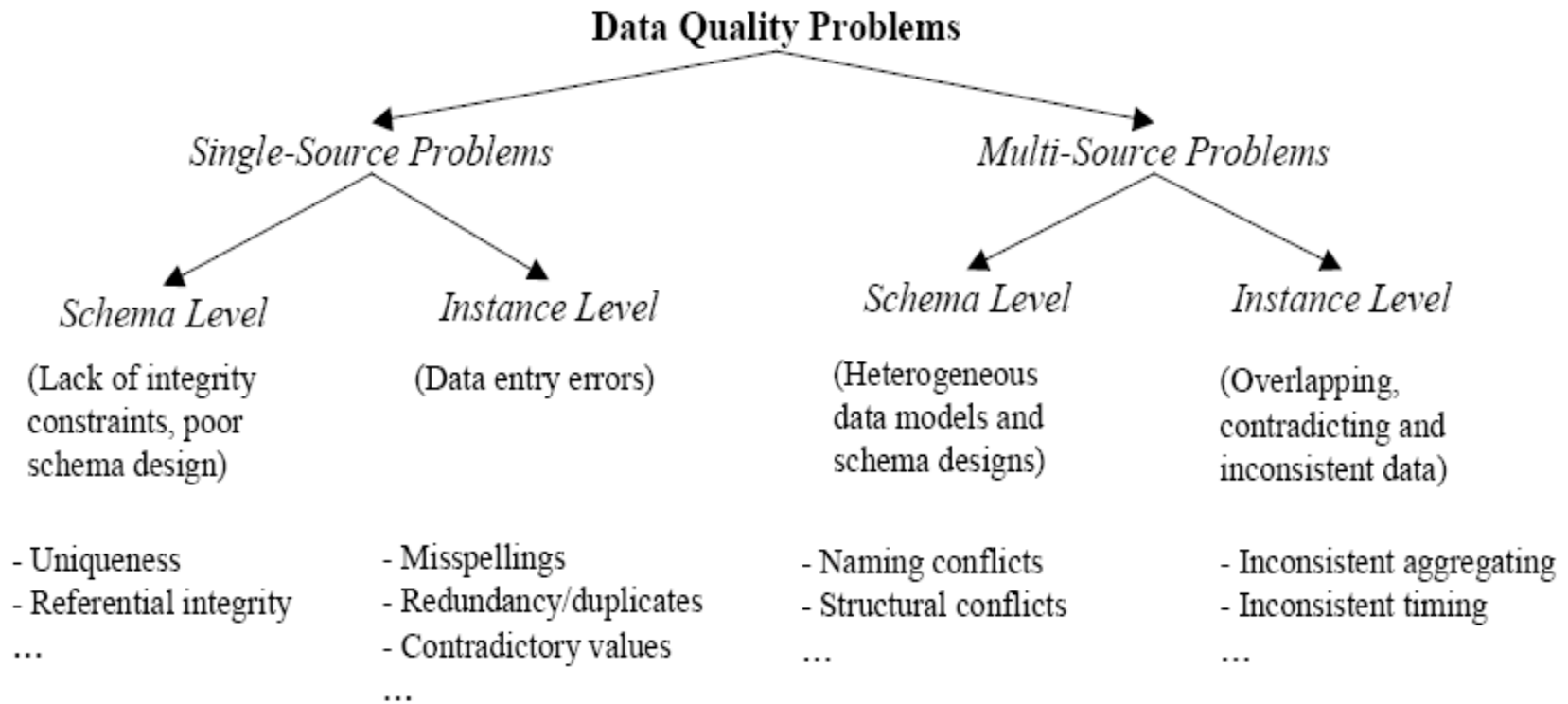
2. Tiền xử lý dữ liệu

- **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,



- ▮ (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.



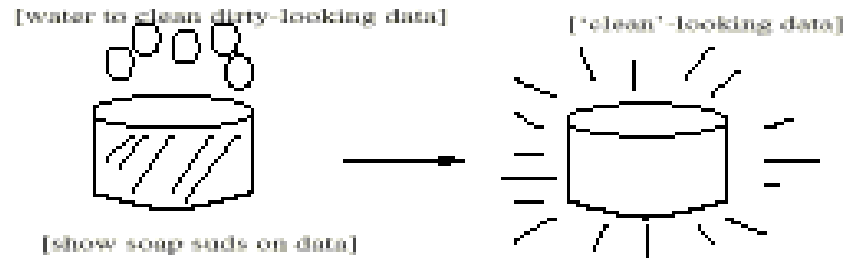
Độ đo đa chiều chất lượng dữ liệu

- 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).

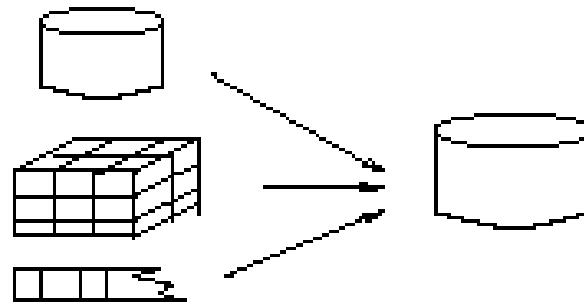
- **Làm sạch dữ liệu**
 - Điền giá trị thiếu, làm trơn dữ liệu nhiễu, định danh hoặc xóa ngoại lai, và khử tính không nhất quán
- **Tích hợp dữ liệu**
 - Tích hợp CSDL, khối dữ liệu hoặc tập tin phức
- **Chuyển dạng dữ liệu**
 - Chuẩn hóa và tổng hợp
- **Rút gọn dữ liệu**
 - Thu được trình bày thu gọn 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

Các thành phần của tiền xử lý dữ liệu

Data Cleaning



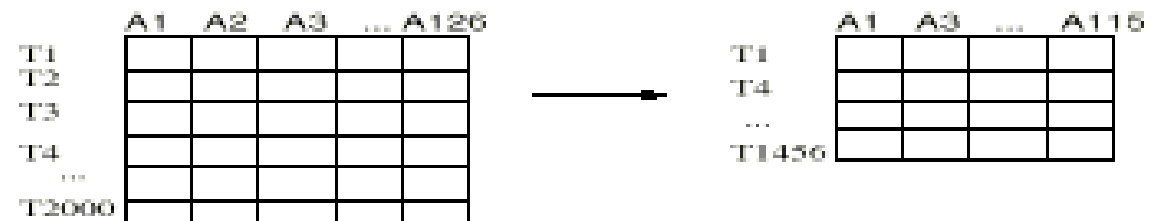
Data Integration



Data Transformation

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

Data Reduction





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

- Là quá trình
 - xác định tính không chính xác, không đầy đủ/tính bất hợp lý của dữ liệu
 - chỉnh sửa các sai sót và thiếu sót được phát hiện
 - nâng cao chất lượng dữ liệu.
- Quá trình bao gồm
 - kiểm tra định dạng, tính đầy đủ, tính hợp lý, miền giới hạn,
 - xem xét dữ liệu để xác định ngoại lai (địa lý, thống kê, thời gian hay môi trường) hoặc các lỗi khác,
 - đánh giá dữ liệu của các chuyên gia miền chủ đề.
- Quá trình thường dẫn đến
 - loại bỏ, lập tài liệu và kiểm tra liên tiếp và hiệu chỉnh đúng bản ghi nghi ngờ.
 - Kiểm tra xác nhận có thể được tiến hành nhằm đạt tính phù hợp với các chuẩn áp dụng, các quy luật, và quy tắc.



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 (bán 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**

- Nhiều:
 - 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



Phương pháp phân hoạch đơn giản xếp thùng

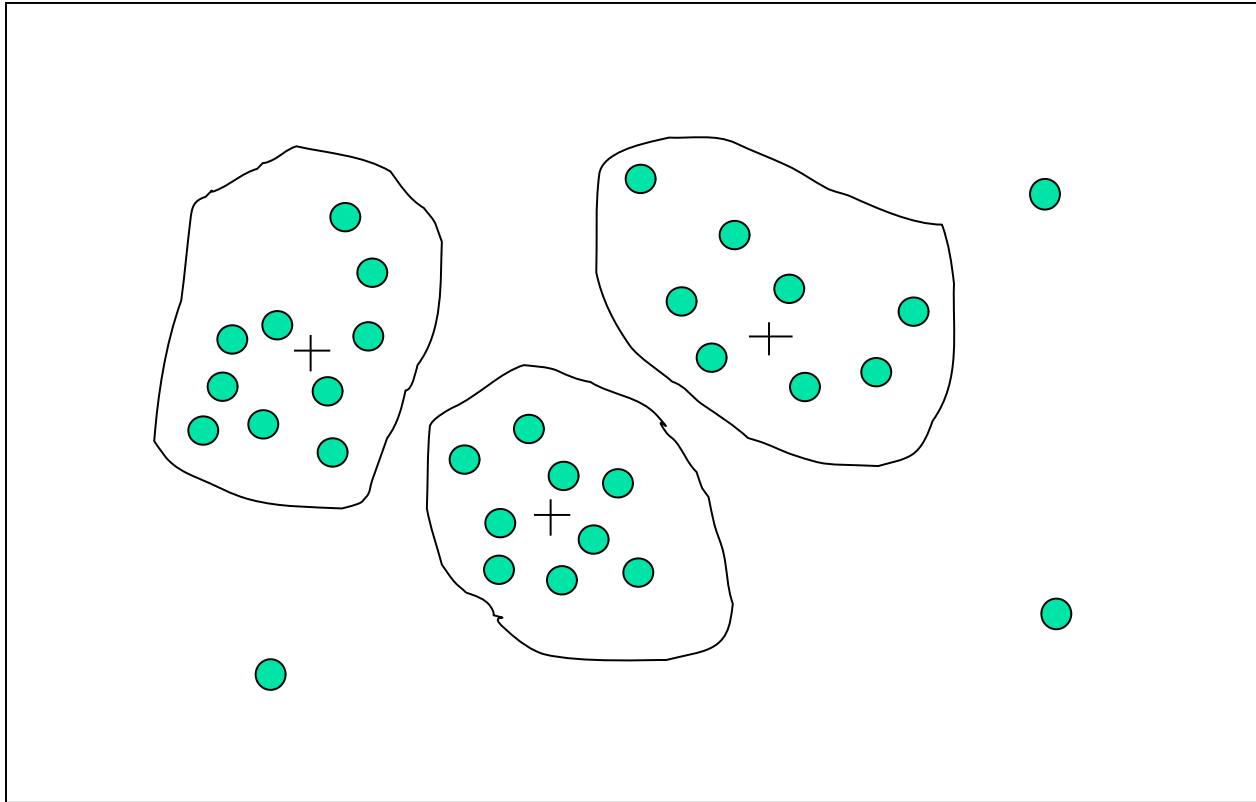
- **Binning**
- **Phân hoạch cân bằng bề rộng Equal-width** (distance) partitioning:
 - Chia miền giá trị: N đoạn dài như nhau: uniform grid
 - Miền giá trị từ A (nhỏ nhất) tới B (lớn nhất) $\rightarrow W = (B - A)/N$.
 - Đơn giản nhất song bị định hướng theo ngoại lai.
 - Không xử lý tốt khi dữ liệu không cân bằng (đều).
- **Phân hoạch cân bằng theo chiều sâu Equal-depth** (frequency) partitioning:
 - Chia miền xác định thành N đoạn “đều nhau về số lượng”, các đoạn có xấp xỉ số ví dụ mẫu.
 - Khả cỡ dữ liệu: tốt.
 - Việc quản lý các thuộc tính lớp: có thể “khôn



P/pháp xếp thùng làm trơn dữ liệu

- * 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
- * Làm trơn thùng theo trung bình:
 - Bin 1: 9, 9, 9, 9
 - Bin 2: 23, 23, 23, 23
 - Bin 3: 29, 29, 29, 29
- * Làm trơn thùng theo biên:
 - Bin 1: 4, 4, 4, 15
 - Bin 2: 21, 21, 25, 25
 - Bin 3: 26, 26, 26, 34

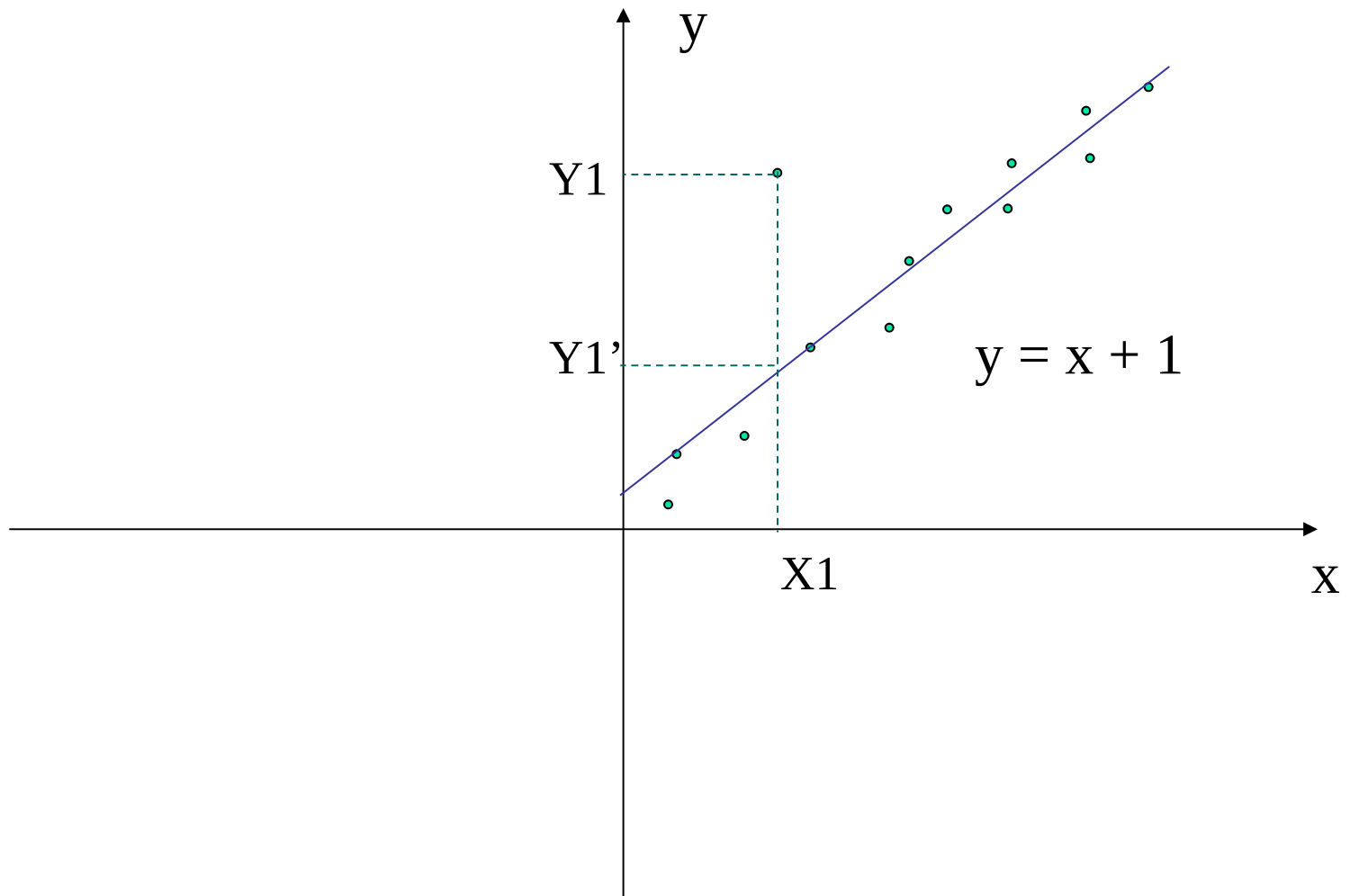
Analysis)



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 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

Nguồn dữ liệu đơn: mức sơ đồ

Ví dụ

Scope/Problem		Dirty Data	Reasons/Remarks
Attribute	Illegal values	bdate=30.13.70	values outside of domain range
Record	Violated attribute dependencies	age=22, bdate=12.02.70	age = (current date – birth date) should hold
Record type	Uniqueness violation	emp ₁ =(name="John Smith", SSN="123456") emp ₂ =(name="Peter Miller", SSN="123456")	uniqueness for SSN (social security number) violated
Source	Referential integrity violation	emp=(name="John Smith", deptno=127)	referenced department (127) not defined

Nguồn dữ liệu đơn: mức thể hiện

Scope/Problem		Dirty Data	Reasons/Remarks
Attribute	Missing values	phone=9999-999999	unavailable values during data entry (dummy values or null)
	Misspellings	city="Liipzig"	usually typos, phonetic errors
	Cryptic values, Abbreviations	experience="B"; occupation="DB Prog."	
	Embedded values	name="J. Smith 12.02.70 New York"	multiple values entered in one attribute (e.g. in a free-form field)
	Misfielded values	city="Germany"	
Record	Violated attribute dependencies	city="Redmond", zip=77777	city and zip code should correspond
Record type	Word transpositions	name ₁ ="J. Smith", name ₂ ="Miller P."	usually in a free-form field
	Duplicated records	emp ₁ =(name="John Smith",...); emp ₂ =(name="J. Smith",...)	same employee represented twice due to some data entry errors
	Contradicting records	emp ₁ =(name="John Smith", bdate=12.02.70); emp ₂ =(name="John Smith", bdate=12.12.70)	the same real world entity is described by different values
Source	Wrong references	emp=(name="John Smith", deptno=17)	referenced department (17) is defined but wrong

Nguồn dữ liệu phức: sơ đồ/thể hiện

Customer (source 1)

<i>CID</i>	<i>Name</i>	<i>Street</i>	<i>City</i>	<i>Sex</i>
11	Kristen Smith	2 Hurley Pl	South Fork, MN 48503	0
24	Christian Smith	Hurley St 2	S Fork MN	1

Client (source 2)

<i>Cno</i>	<i>LastName</i>	<i>FirstName</i>	<i>Gender</i>	<i>Address</i>	<i>Phone/Fax</i>
24	Smith	Christoph	M	23 Harley St, Chicago IL, 60633-2394	333-222-6542 / 333-222-6599
493	Smith	Kris L.	F	2 Hurley Place, South Fork MN, 48503-5998	444-555-6666

Customers (integrated target with cleaned data)

<i>No</i>	<i>LName</i>	<i>FName</i>	<i>Gender</i>	<i>Street</i>	<i>City</i>	<i>State</i>	<i>ZIP</i>	<i>Phone</i>	<i>Fax</i>	<i>CID</i>	<i>Cno</i>
1	Smith	Kristen L.	F	2 Hurley Place	South Fork	MN	48503-5998	444-555-6666		11	493
2	Smith	Christian	M	2 Hurley Place	South Fork	MN	48503-5998			24	
3	Smith	Christoph	M	23 Harley Street	Chicago	IL	60633-2394	333-222-6542	333-222-6599		24

- 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



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



Kết hợp khối dữ liệu

- **DataCube Aggregation**
- Mức thấp nhất của khối dữ liệu
 - Tổng hợp dữ liệu thành một cá thể quan tâm
 - Chẳng hạn, một khách hàng trong kho dữ liệu cuộc gọi điện thoại.
- Các mức phức hợp của tích hợp thành khối dữ liệu
 - Giảm thêm kích thước dữ liệu
- Tham khảo mức thích hợp
 - Sử dụng trình diễn nhỏ nhất đủ để giải bài toán
- Nên sử dụng dữ liệu khối lập phương khi trả lời câu hỏi tổng hợp thông tin

Rút gọn chiều

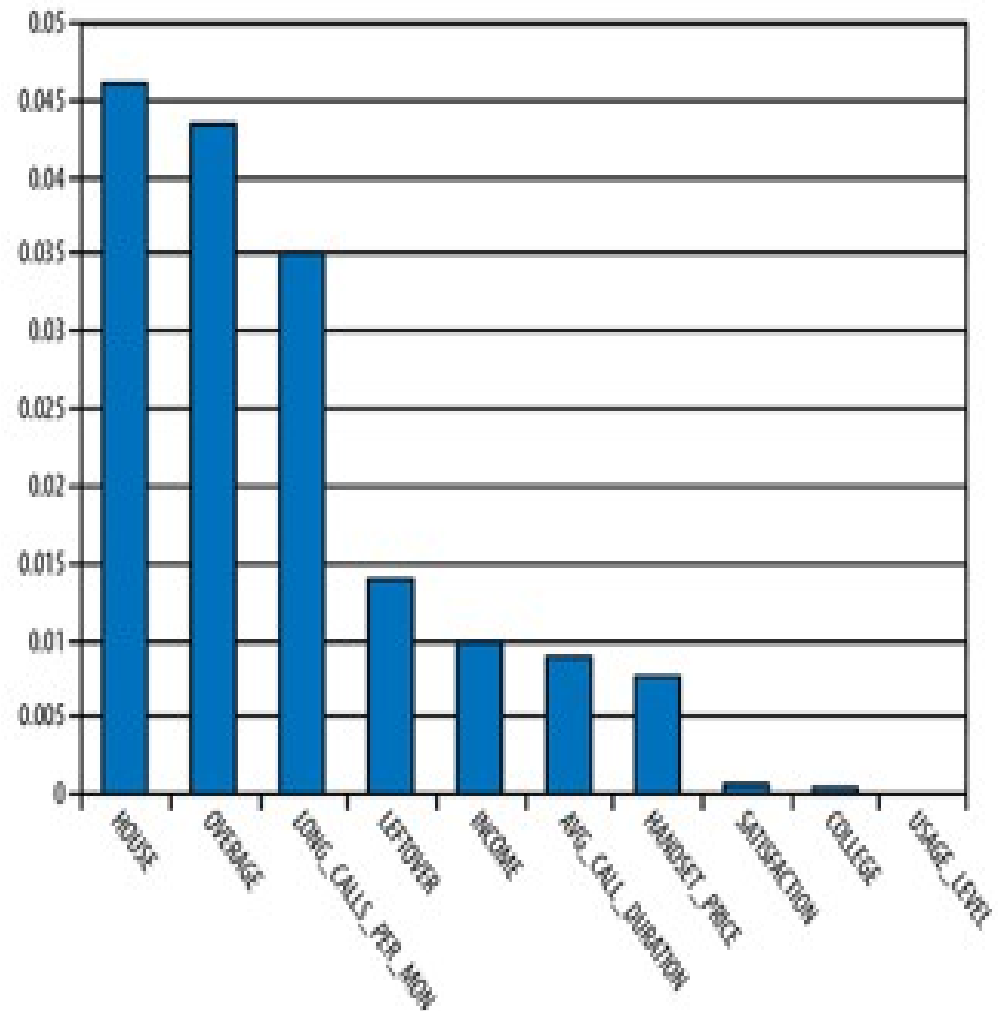
- Rút gọn đặc trưng (như., lựa chọn tập con thuộc tính):
 - Lựa chọn tập nhỏ nhất các đặc trưng mà phân bố xác suất của các lớp khác nhau cho giá trị khi cho giá trị của các lớp này gần như phân bố vốn có đã cho giá trị của các đặc trưng
 - Rút gọn # của các mẫu trong tập mẫu dễ dàng hơn để hiểu dữ liệu
- 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

<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

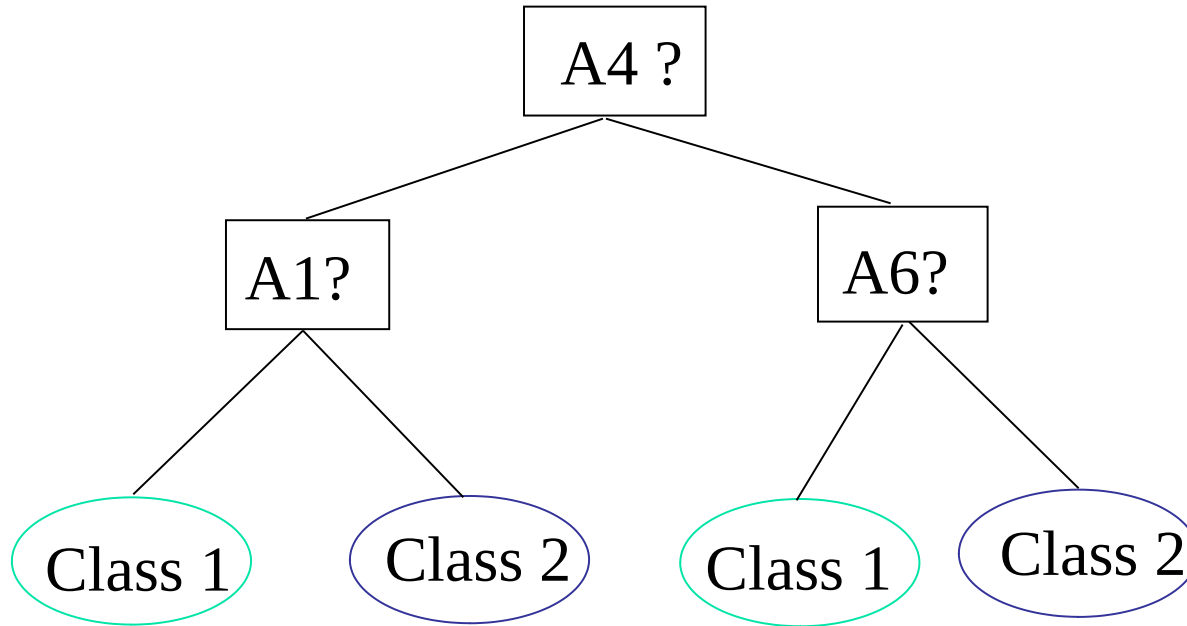


Độ quan trọng các thuộc tính: Tiến hành lại sau mỗi bước

Ví dụ rút gọn cây quyết định

Tập thuộc tính khởi tạo:

$\{A1, A2, A3, A4, A5, A6\}$



-----> Tập thuộc tính rút gọn: $\{A1, A4, A6\}$

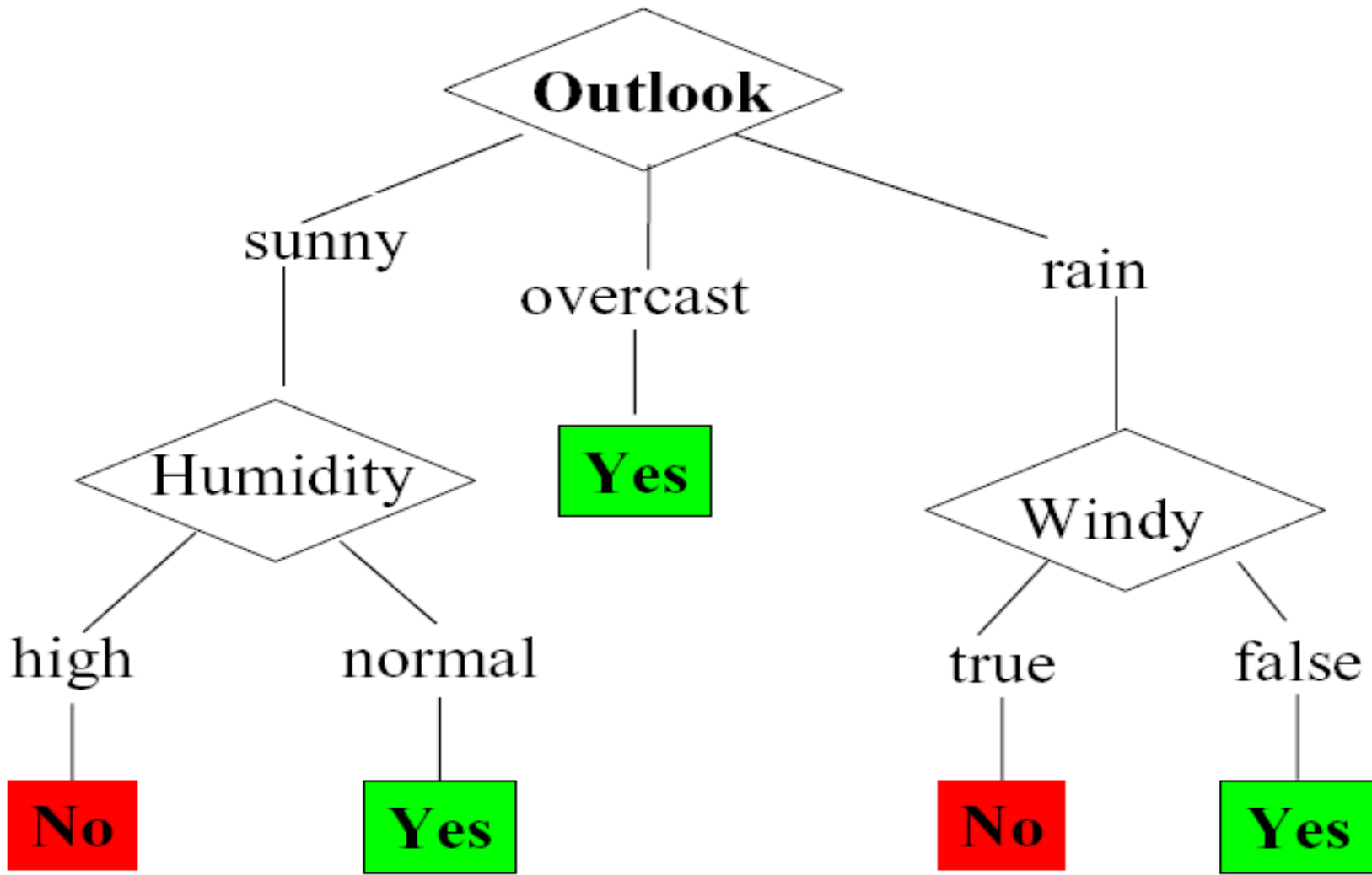


Phân lớp cây quyết định

- Đồ thị dạng cây
- Đỉnh trong là một hàm test
- Các nhánh tương ứng với kết quả kiểm tra tại đỉnh trong
- Các lá là các nhãn, hoặc các lớp.
- Xem Chương 5



Phân lớp cây quyết định





Phân lớp cây quyết định

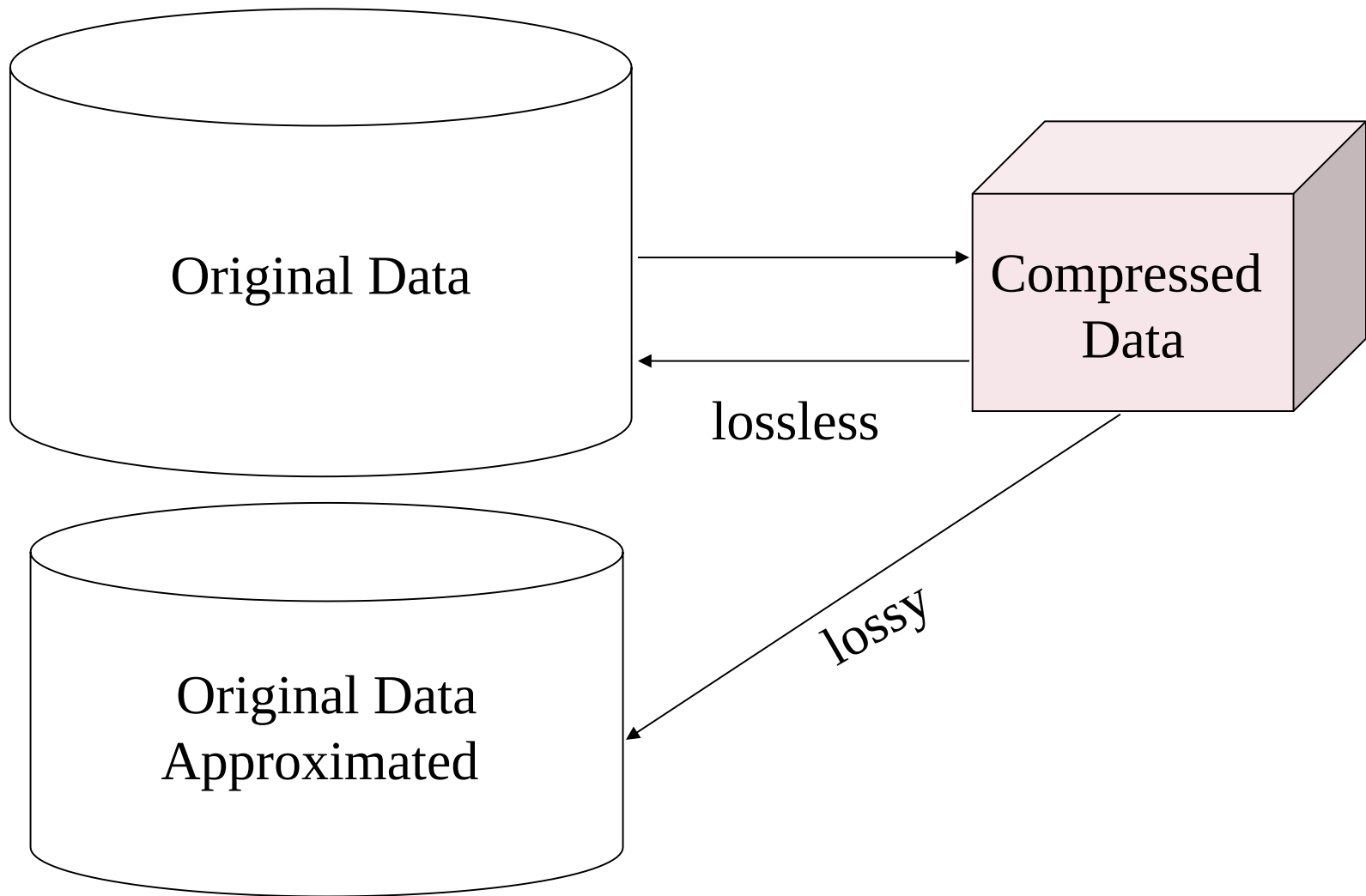
- Xây dựng cây quyết định:
 - Xây dựng cây quyết định
 - Phương pháp top-down
 - Cắt tỉa cây (pruning)
 - Phương pháp bottom-up: xác định và loại bỏ những nhánh rườm rà tăng độ chính xác khi phân lớp những đối tượng mới
- Sử dụng cây quyết định: phân lớp các đối tượng chưa được gán nhãn



Compression)

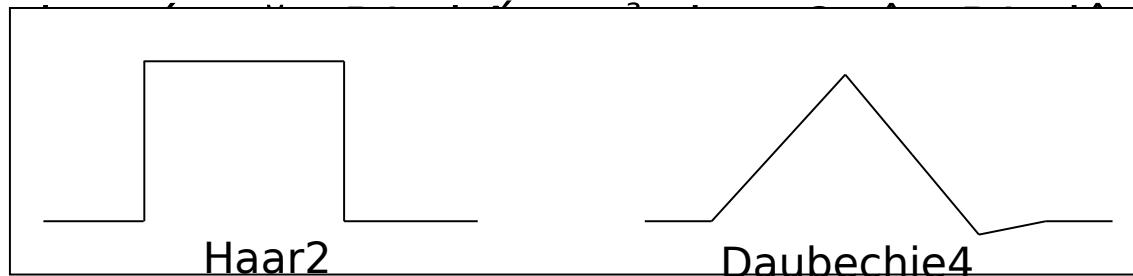
- Nén sâu văn bản
 - Tồn tại lý thuyết phong phú và thuật toán điển hình
 - **Mạnh:** Không tổn thất điển hình
 - **Yếu:** chỉ các thao tác hạn hẹp mà không mở rộng
- Nén Audio/video
 - Nén tổn thất điển hình, với tinh lọc cải tiến
 - Vài trường hợp mảnh tín hiệu nhỏ được tái hợp không cần dựng toàn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

Compression)

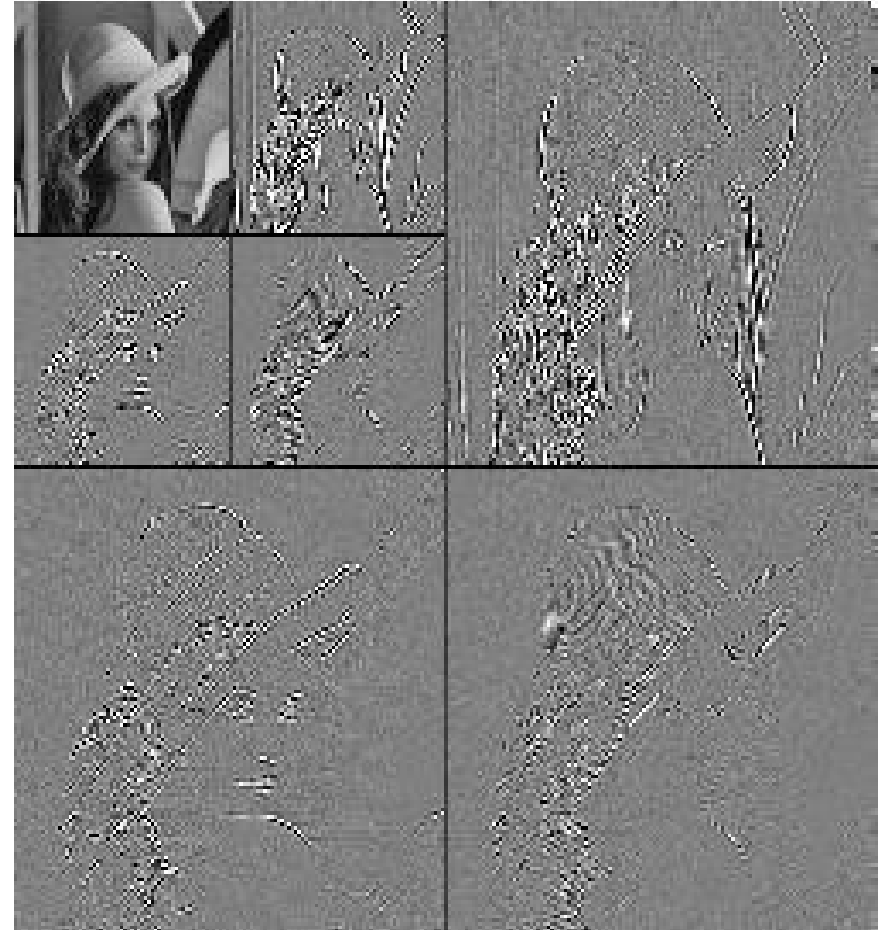
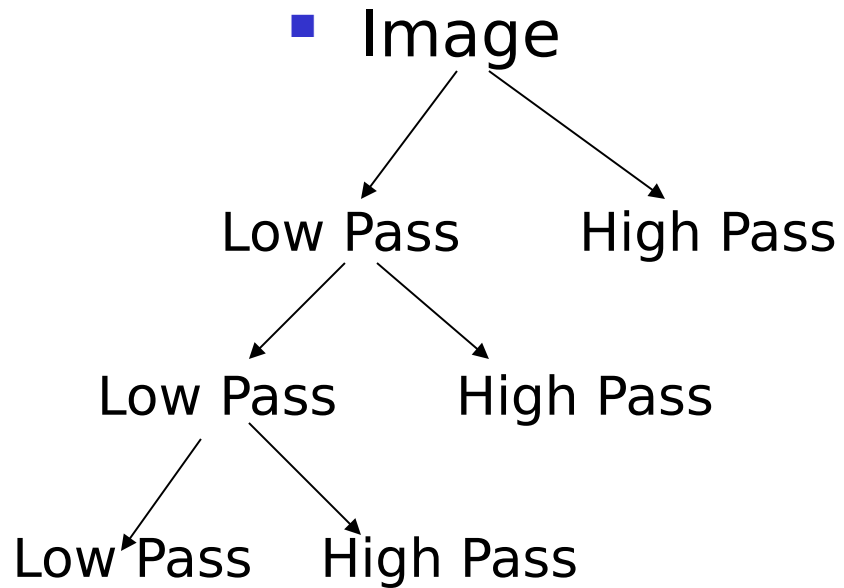


Chuyển dạng sóng

- Wavelet Transformation
- Biến dạng sóng rời rạc (Discrete wavelet transform:DWT): XL tín hiệu tuyến tính, phân tích đa giải pháp
- Xấp xỉ nén: chỉ lưu một mảnh nhỏ các hệ số sóng lớn nhất
- Tương tự như biến đổi rời rạc Fourier (DFT), nhưng nén tổn thất tốt hơn, bản địa hóa trong không gian
- Phương pháp:
 - Độ dài, L , buộc là số nguyên lũy thừa 2 (đệm thêm các chữ số 0, khi cần)
 - Mỗi phép biến đổi có 2 chức năng: làm mịn, phân tách
 - Áp dụng
 - Áp dụng



DWT cho nén ảnh

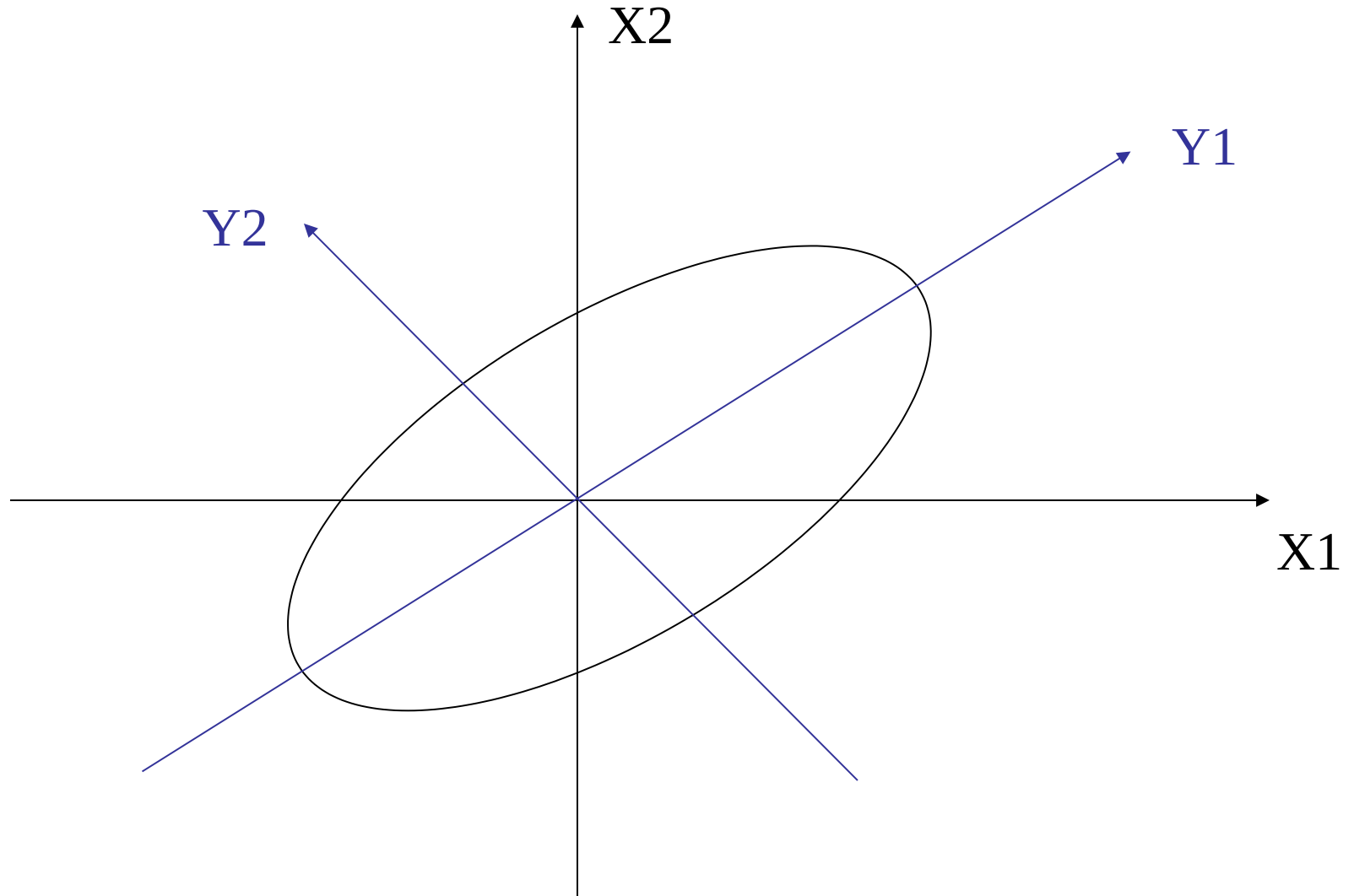




Phân tích thành phần chính PCA

- **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 tích thành phần chính



Rút gọn kích thước số

- Phương pháp tham số
 - Giả sử dữ liệu phù hợp với mô hình nào đó, ước lượng tham số mô hình, lưu chỉ các tham số, và không lưu dữ liệu (ngoại trừ các ngoại lai có thể có)
 - Mô hình tuyến tính loga (Log-linear models): lấy giá trị tại một điểm trong không gian M-chiều như là tích của các không gian con thích hợp
- Phương pháp không tham số
 - Không giả thiết mô hình
 - Tập hợp chính: biểu đồ (histograms), phân cụm (clustering), lấy mẫu (sampling)

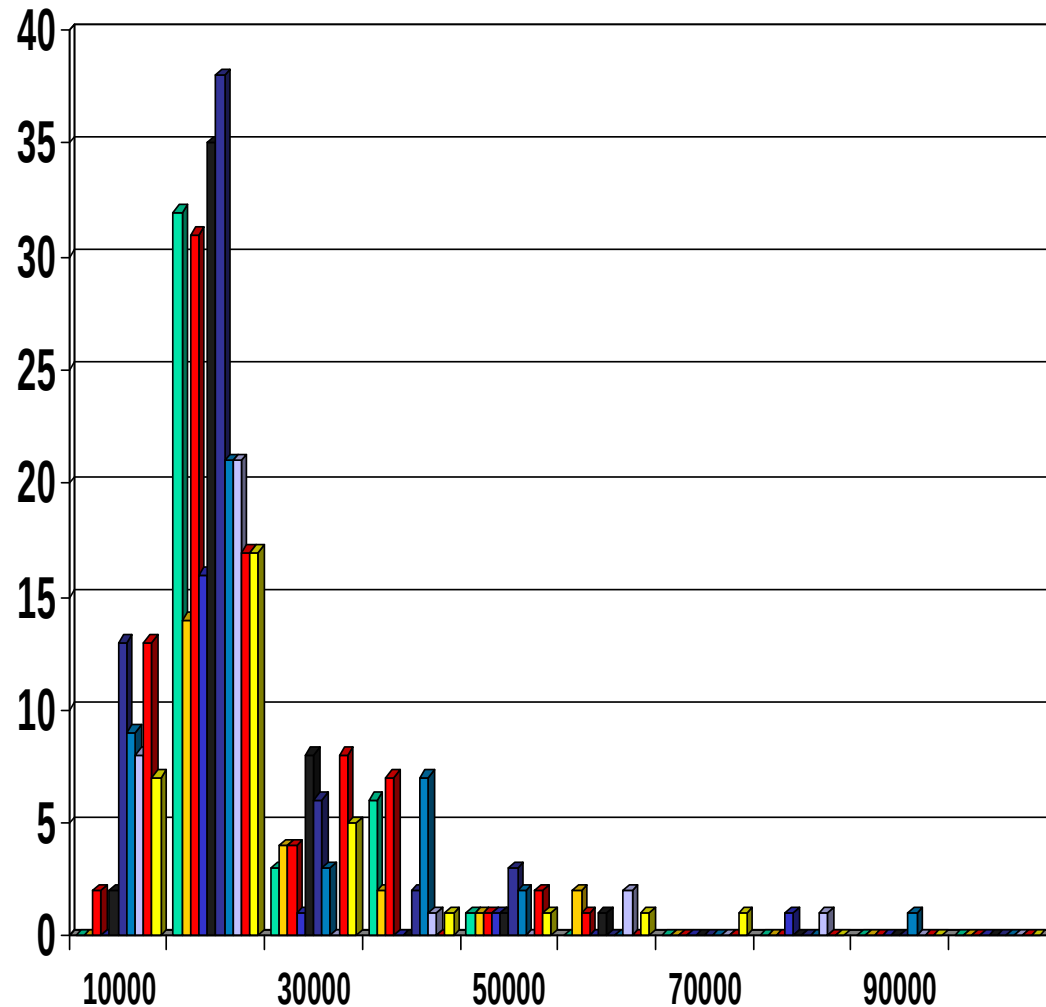
- Hồ quy tuyến tính: DL được mô hình hóa phù hợp với 1 đường thẳng
 - Thường dùng phương pháp bình phương tối thiểu để khớp với đường
- Hồ quy đa chiều: Cho một biến đích Y được mô hình hóa như ột hàm tuyến tính của vector đặc trưng đa chiều
- Mô hình tuyến tính loga: rời rạc hóa xấp xỉ các phân bố xác suất đa chiều



Phân tích MH hồi quy tuyến tính và logarit

- Hồi quy tuyến tính: $Y = \beta_0 + \beta_1 X$
 - Hai tham số, β_0 và β_1 đặc trưng cho đường và được xấp xỉ qua dữ liệu đã nắm bắt được.
 - Sử dụng chiến lược BP tối thiểu tới các giá trị đã biết $Y_1, Y_2, \dots, X_1, X_2, \dots$
- Hồi quy đa chiều: $Y = b_0 + b_1 X_1 + b_2 X_2$.
 - Nhiều hàm không tuyến tính được chuyển dạng như trên.
- Mô hình tuyến tính loga:
 - Bảng đa chiều của xác suất tích nối được xấp xỉ bởi tích của các bảng bậc thấp hơn
 - Xác suất: $p(a, b, c, d) = p(a, b) p(a, c) p(a, d) p(b, c, d)$

- Histograms
- Kỹ thuật rút gọn dữ liệu phổ biến
- Phân dữ liệu vào các thùng và giữ trung bình (tổng) của mỗi thùng
- Có thể được dựng tối ưu hóa theo 1 chiều khi dùng quy hoạch động
- Có quan hệ tới bài toán lượng tử hóa.





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

- 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 tốt 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.

Rút gọn mẫu

- Simple Random Sampling (SRS)
- SRS with replacement (SRSWR)
 - Chọn một phần tử dữ liệu đưa vào mẫu
 - Loại bỏ phần tử dữ liệu đó ra khỏi tập dữ liệu
 - Lặp tiếp cho đến khi có n phần tử dữ liệu
 - Các phần tử dữ liệu giống nhau có thể được chọn nhiều lần
- SRS without replacement (SRSWOR)
 - Chọn một phần tử và không bị loại bỏ. Các mẫu DL phân biệt
- Ví dụ: Chọn mẫu 2 (n) phần tử từ tập 4 dữ liệu

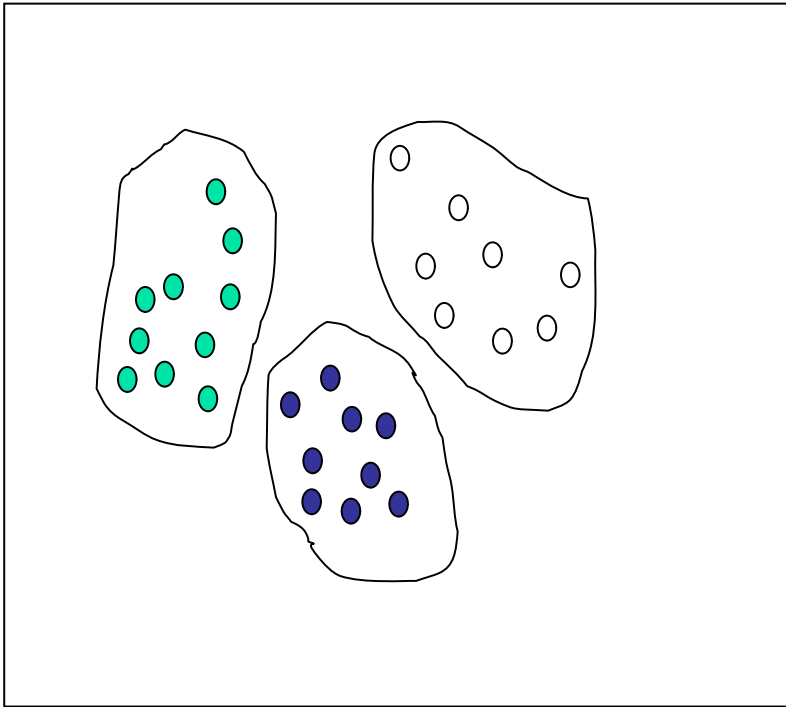
- With SRSWR: $4^2 = 16$
- AA, AB, AC, AD,
BA, BB, BC, BD,
CA, CB, CC, CD,
DA, DB, DC, DD = 16 samples

$$\frac{N!}{(N-n)!n!} = \frac{4!}{2!2!} = 6$$

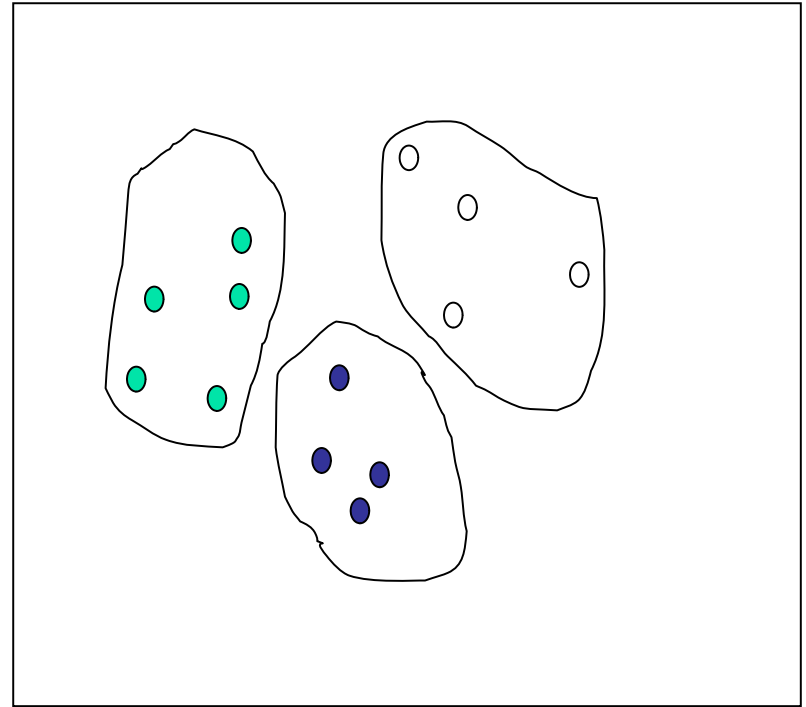
AB, AC, AD,
~~BA~~, BC, BD,
~~CA~~, ~~CB~~, CD,
~~DA~~, ~~DB~~, ~~DC~~

Rút gọn mẫu

Raw Data



Mẫu cụm/phân tầng





Rút gọn phân cấp

- Dùng cấu trúc đa phân giải với các mức độ khác nhau của rút gọn
- Phân cụm phân cấp thường được thi hành song có khuynh hướng xác định phân vùng DL hơn là “phân cụm”
- Phương pháp tham số thường không tuân theo trình bày phân cấp
- Tích hợp phân cấp
 - Một cây chỉ số được chia phân cấp một tập DL thành các vùng bởi miền giá trị của một vài thuộc tính
 - Mỗi vùng được coi như một thùng
 - Như vậy, cây chỉ số với tích hợp lưu trữ mỗi nút là một sơ đồ phân cấp



Rút gọn đặc trưng

■ Giới thiệu chung

- “Tối ưu hóa” chọn tập đặc trưng
 - Số lượng đặc trưng nhỏ hơn
 - Hy vọng tăng tốc độ thi hành
 - Tăng cường chất lượng khai phá văn bản. ? Giảm đặc trưng đi là tăng chất lượng: có các đặc trưng “nhiều”
 - Hoặc cả hai mục tiêu trên

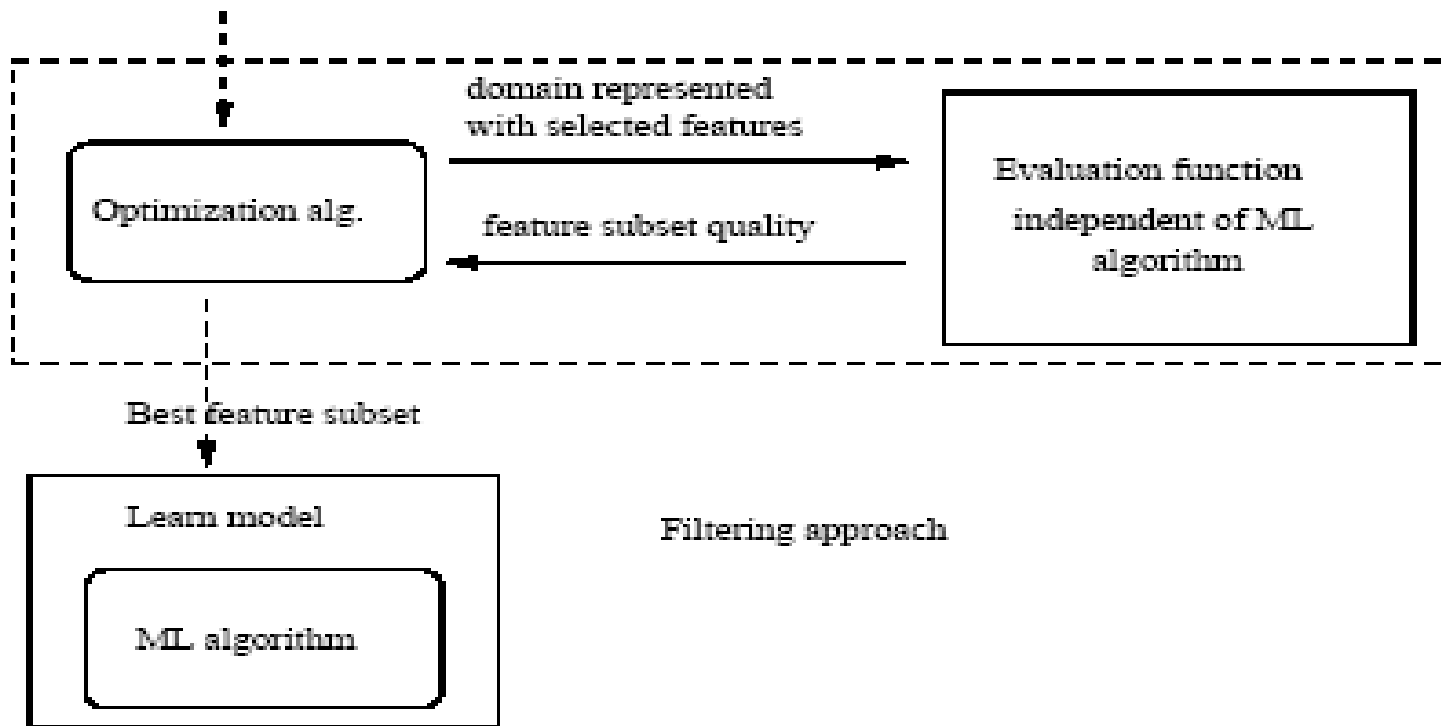
■ Hai tiếp cận điển hình

- Tiếp cận lọc
- Tiếp cận bao gói

■ Với dữ liệu văn bản

- Tập đặc trưng: thường theo mô hình vector
- Tính giá trị của từng đặc trưng giữ lại các đặc trưng được coi là “tốt”.

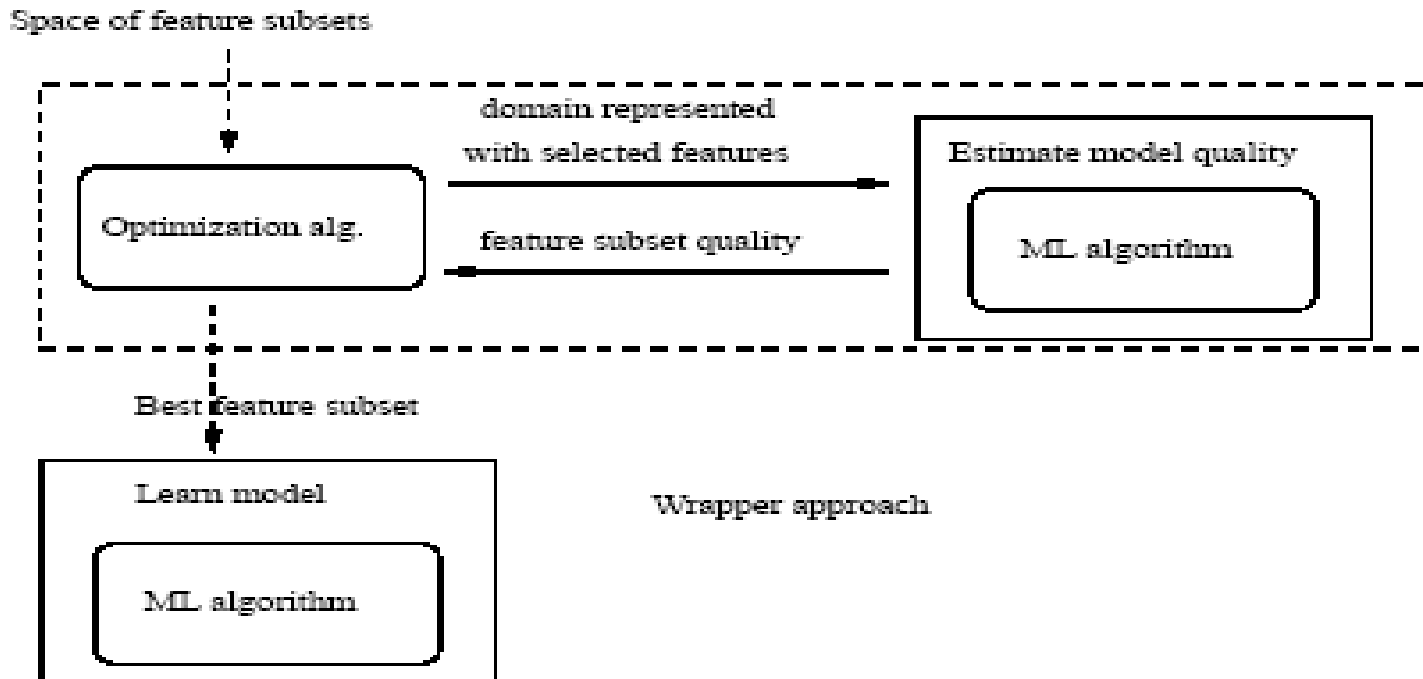
Space of feature subsets



■ Tiếp cận lọc

- Đầu vào: Không gian tập các tập đặc trưng
- Đầu ra: Tập con đặc trưng tốt nhất
- Phương pháp
 - Dò tìm “cải tiến” bộ đặc trưng: Thuật toán tối ưu hóa
 - Đánh giá chất lượng mô hình: độc lập với thuật toán học máy

Tiếp cận bao gói tổng quát



■ Tiếp cận bao gói

- Đầu vào: Không gian tập các tập đặc trưng
- Đầu ra: Tập con đặc trưng tốt nhất
- Phương pháp
 - Dò tìm “cải tiến” bộ đặc trưng: Thuật toán tối ưu hóa
 - Đánh giá chất lượng mô hình: Dùng chính thuật toán học để đánh giá



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

Rời rạc hóa dựa trên Entropy

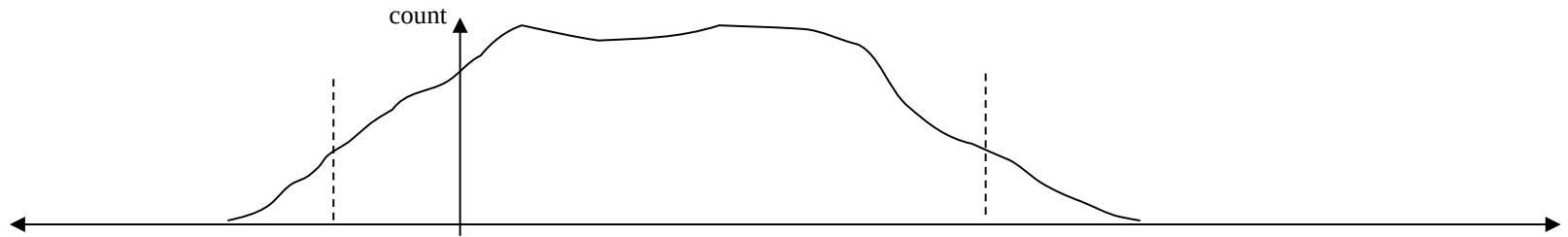
- Cho tập ví dụ S , nếu S được chia thành 2 đoạn S_1 và S_2 dùng biên T , thì entropy sau khi phân đoạn là
$$E(S, T) = \frac{|S_1|}{|S|} Ent(S_1) + \frac{|S_2|}{|S|} Ent(S_2)$$
- Biên làm cực tiểu hàm entropy trên tất cả các biên được chọn như một rời rạc hóa nhị phân.
- Quá trình đệ quy tới các vùng cho tới khi đạt điều kiện dừng nào đó, như
$$Ent(S) - E(T, S) > \delta$$
- Thực nghiệm chỉ ra rằng cho phép rút gọn cỡ DL và tăng độ chính xác phân lớp



Phân đoạn bằng phân hoạch tự nhiên

- Quy tắc đơn giản 3-4-5 được dùng để phân đoạn dữ liệu số thành các đoạn tương đối thống nhất, “tự nhiên”.
 - Hướng tới số giá trị khác biệt ở vùng quan trọng nhất
 - Nếu 3, 6, 7 hoặc 9 giá trị khác biệt thì chia miền thành 3 đoạn tương đương.
 - Nếu phủ 2, 4, hoặc 8 giá trị phân biệt thì chia thành 4.
 - Nếu phủ 1, 5, hoặc 10 giá trị phân biệt thì chia thành 5.


Ví dụ luật 3-4-5



Step 1:	-\$351	-\$159	profit	\$1,838	\$4,700	
	Min	Low (i.e, 5%-tile)			High(i.e, 95%-0 tile)	Max

Step 2: msd=1,000 Low=-\$1,000 High=\$2,000

Step 3:



```
graph TD; A["(-$1,000 - $2,000)"] --- B["(-$1,000 - 0)"]; A --- C["(0 -$ 1,000)"]; A --- D["($1,000 - $2,000)"]
```

Step 4:

The diagram shows a decision tree for Step 4. The root node is labeled $(-\$4000 - \$5,000)$. It branches into four nodes: $(-\$400 - 0)$, $(0 - \$1,000)$, $(\$1,000 - \$2,000)$, and $(\$2,000 - \$5,000)$. Each of these nodes further branches into three more nodes, representing a total of 16 terminal nodes. The terminal nodes are labeled with pairs of values, such as $(-\$400 - \$300)$, $(0 - \$200)$, $(\$1,000 - \$1,200)$, and $(\$2,000 - \$3,000)$.



Sinh kiến trúc khái niệm dữ liệu phân loại

- Đặc tả một thứ tự bộ phận giá trị thuộc tính theo mức sơ đồ do người dùng hoặc chuyên gia
 - $\text{street} < \text{city} < \text{state} < \text{country}$
- Đặc tả thành cấu trúc phân cấp nhờ nhóm dữ liệu
 - $\{\text{Urbana, Champaign, Chicago}\} < \text{Illinois}$
- Đặc tả theo tập các thuộc tính.
 - Tự động sắp xếp một phần bằng cách phân tích số lượng các giá trị khác biệt
 - Như, $\text{street} < \text{city} < \text{state} < \text{country}$
- Đặc tả một phần thứ tự bộ phận
 - Như, chỉ $\text{street} < \text{city}$ mà không có cái khác



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

