CHƯƠNG 1: TIỀN XỬ LÝ DỮ LIỆU

1.1. KHÁI NIỆM VỀ TIỀN XỬ LÝ DỮ LIỆU?

1.2. TÓM TẮT MÔ TẢ DỮ LIỆU

1.3. LÀM SẠCH DỮ LIỆU

1.4. TÍCH HỢP VÀ CHUYỂN DẠNG DỮ LIỆU

1.5. RÚT GỌN DỮ LIỆU

1.1. KHÁI NIỆM VỀ TIỀN XỬ LÝ DỮ LIỆU

1.1.1. Tại sao phải tiền xử lý dữ liệu?

Dữ liệu trong thế giới thực (mà chúng ta muốn phân tích bằng cách áp dụng các kỹ thuật khai phá dữ liệu) thường:

- **Không hoàn chỉnh** (incomplete): thiếu vắng các giá trị hoặc các thuộc tính đáng quan tâm, hoặc chỉ chứa các dữ liệu gộp nhóm.
- **Chứa đựng các giá trị nhiễu** (noisy): bao gồm các lỗi hoặc các giá trị lệch quá xa ra ngoài phạm vi mong đợi.
- **Không nhất quán** (inconsistent).

Lý do:

- Kích thước dữ liệu quá lớn.
- Được thu thập từ nhiều nguồn khác nhau.

$\Longrightarrow$ Chất lượng dữ liệu thấp sẽ dẫn tới những kết quả khai phá tồi.

Tiền xử lý dữ liệu là quá trình áp dụng các kỹ thuật nhằm nâng cao chất lượng dữ liệu và từ đó giúp nâng cao chất lượng kết quả khai phá.

1.1.2. Những nguyên nhân ảnh hưởng đến chất lượng dữ liệu

A. Nguyên nhân khiến dữ liệu không hoàn chỉnh (incomplete):

- ✓ Giá trị tương ứng không thể chấp nhận vào thời điểm thu thập.
- ✓ Sự khác biệt về quan điểm giữa thời điểm thu thập và thời điểm phân tích.
- ✓ Các lỗi gây ra bởi con người (nhập liệu sót) hoặc bởi hệ thống (phần cứng/phần mềm).

B. Nguyên nhân gây ra các giá trị nhiễu (noisy):

- ✓ Lỗi của các thiết bị thu thập dữ liệu.
- ✓ Lỗi nhập dữ liệu sai (gây ra bởi con người hay máy tính).
- ✓ Lỗi trong quá trình truyền dữ liệu.

C. Nguyên nhân gây ra tính không nhất quán (inconsistent):

- ✓ Dữ liệu đến từ các nguồn khác nhau.
- ✓ Sự vi phạm các phụ thuộc hàm.

D. Sự xuất hiện các bản ghi trùng lặp.

1.1.3. Các kỹ thuật tiền xử lý dữ liệu

***A. Tích hợp dữ liệu (Data Integration)*:** kết hợp dữ liệu từ nhiều nguồn khác nhau thành một kho dữ liệu thống nhất.

⟹ Có thể gây ra: - Sự không nhất quán (inconsistencies).

- Dư thừa dữ liệu (redundancies).

***B. Làm sạch dữ liệu (Data Cleaning)*:** kỹ thuật này được thực hiện thông qua việc bổ sung các *giá trị thiếu (missing values)*, loại bỏ các *dữ liệu nhiễu (noisy data)*, xác định và loại bỏ những *giá trị lệch quá xa so với mong đợi (outliers)*, giải quyết vấn đề *không nhất quán trong dữ liệu (inconsistencies)*.

- Nếu người dùng thấy rằng dữ liệu là không "sạch", họ sẽ không mấy tin tưởng vào kết quả khai phá trên dữ liệu đó.
- Dữ liệu không "sạch" có thể gây ra những nhiễu loạn cho các thủ tục khai phá dữ liệu và dẫn tới những kết quả không đáng tin cậy.
- Dù trong hầu hết các thủ tục khai phá dữ liệu đều cài đặt những cơ chế nhằm xử lý các vấn đề về thiếu vắng giá trị hay nhiễu nhưng chúng không phải lúc nào cũng đáng tin cậy.

⟹ **Làm sạch dữ liệu là bước tiền xử lý cực kỳ quan trọng.**

***C. Chuyển dạng dữ liệu (Data Transformation)*:** bao gồm các thao tác như là *chuẩn hóa (normalization)* và gộp nhóm (*aggregation*). Đây là kỹ thuật bổ sung góp phần vào thành công của tiến trình khai phá dữ liệu.

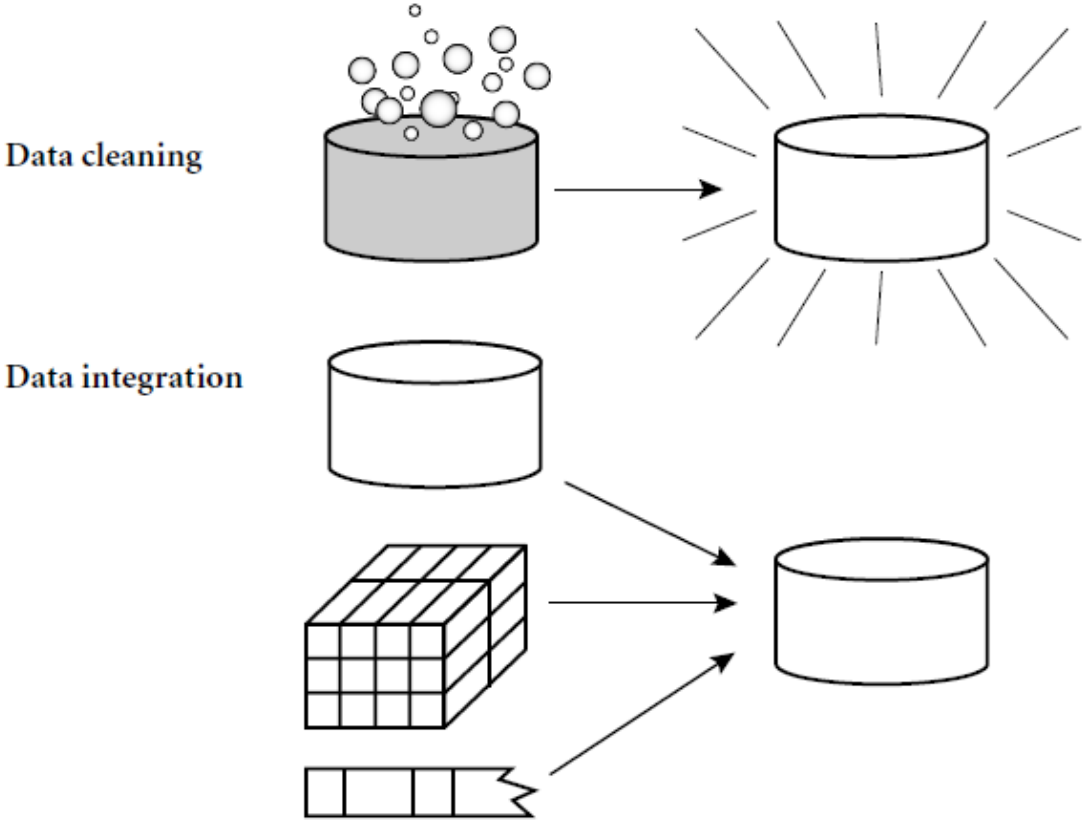
***D. Rút gọn dữ liệu (Data Reduction)*:**

Tập dữ liệu quá lớn (huge) sẽ làm tiến trình khai phá trở nên chậm chạp

⟹ ***Nhu cầu: Giảm kích thước tập dữ liệu mà không ảnh hưởng đến kết quả khai phá.***

Kỹ thuật rút gọn dữ liệu cho phép biểu diễn tập dữ liệu dưới dạng rút gọn tức là nhỏ hơn rất nhiều về mặt kích thước/dung lượng (volume) nhưng vẫn cho kết quả khai phá/phân tích chính xác. Các chiến lược:

- ***Gộp nhóm dữ liệu (data aggregation)***: vd: xây dựng một data cube.
- ***Lựa chọn tập thuộc tính (attribute subset selection)***: vd: loại bỏ các thuộc tính không thích hợp thông qua phân tích tương quan (correlation analysis).
- ***Giảm số chiều dữ liệu (dimensionality reduction)***: giảm số lượng các biến ngẫu nhiên hoặc thuộc tính. Vd: sử dụng các lược đồ mã hóa với chiều dài mã tối thiểu hoặc sử dụng biến đổi wavelet.
- ***Giảm biểu diễn số lớn (numerosity reduction)***: thay dữ liệu đã có bằng các cách biểu diễn thay thế gọn hơn như là sử dụng biểu diễn cụm (cluster) hoặc mô hình tham số (parametric model).
- ***Sử dụng lược đồ phân cấp khái niệm***: khái niệm mức thấp (low-level) được thay thế bằng các khai niệm ở mức cao hơn (higher-level).


Data cleaning
Data integration
Data transformation
−2, 32, 100, 59, 48
−0.02, 0.32, 1.00, 0.59, 0.48
Data reduction
attributes
A1 A2 A3 ... A126
transactions
T1
T2
T3
T4
...
T2000
attributes
A1 A3 ... A115
transactions
T1
T4
...
T1456

1.2. TÓM TẮT MÔ TẢ DỮ LIỆU

Để có thể khai phá dữ liệu thành công, cần có cái nhìn toàn thể về bức tranh dữ liệu muốn khai phá.

1.2.1. Khái niệm về tóm tắt mô tả dữ liệu

Tóm tắt mô tả dữ liệu (descriptive data summarization) là kỹ thuật được sử dụng nhằm xác định những đặc trưng điển hình và những đặc điểm nổi bật (hightlight) của dữ liệu (những giá trị được xem là nhiễu (noise) hoặc vượt ngoài phạm vi mong đợi (outliers)).

Khi nghiên cứu các đặc trưng của dữ liệu, người ta quan tâm tới:

1. ***Xu hướng tập trung của dữ liệu*** (central tendency): đặc trưng bởi các đại lượng thống kê: trung bình, trung vị, mode, midrange.
2. ***Sự phân ly của dữ liệu*** (dispersion): đặc trưng bởi các các đại lượng như: tứ phân vị (quartile),khoảng tứ phân vị (interquartile range – IRQ), phương sai (variance).

1.2.2. Đánh giá xu hướng tập trung của dữ liệu

1.2.2.1. Giá trị trung bình (Mean)

Xét dãy gồm N giá trị $\{x_1, x_2, \ldots, x_N\}$. *Giá trị trung bình* (mean) được xác định bởi công thức:

$$\overline{X} = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{x_1 + x_2 + \ldots + x_N}{N}$$

Nếu mỗi giá trị x_i có một trọng số w_i đi kèm thì giá trị trung bình gọi là *trung bình dựa trên trọng số* (weighted average) và được xác định bởi:

$$\overline{X} = \frac{\sum_{i=1}^{N} x_i w_i}{\sum_{i=1}^{N} w_i} = \frac{x_1 w_1 + x_2 w_2 + \ldots + x_N w_N}{w_1 + w_2 + \ldots + w_N}$$

Trị trung bình xác định giá trị "trung tâm" (center) của tập dữ liệu.

1.2.2.2. Trung vị (Median)

Xét dãy gồm N giá trị được ***sắp có thứ tự*** $\{x_1, x_2, \ldots, x_N\}$. Nếu N là số nguyên lẻ (N=2K+1) thì trung vị Med = $x_{[N/2]+1}$ (phần tử chính giữa dãy). Nếu N là số nguyên chẵn (N=2K) thì trung vị Med = $(X_{N/2} + X_{N/2+1})/2$ (trung bình cộng của hai phần tử chính giữa dãy).

Tính xấp xỉ giá trị của trung vị

- Dữ liệu được nhóm thành từng **đoạn** (intervals) tùy thuộc vào các giá trị dữ liệu x_i.
- **Tần suất xuất hiện** (frequency) ứng với mỗi đoạn(*thường được xác định bằng số giá trị có trong mỗi đoạn*) đều đã biết.
- Đoạn có tần suất xuất hiện là trung vị của các tần suất gọi là **đoạn trung vị** (median interval).

Trung vị của toàn tập dữ liệu có thể tính xấp xỉ bởi:

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

N: số giá trị có trong toàn bộ tập dữ liệu

L_1: biên dưới của đoạn trung vị

$(\sum freq)_l$: tổng tần suất của các đoạn nhỏ hơn đoạn trung vị

$freq_{median}$: tần suất của đoạn trung vị

$width$: độ rộng của đoạn trung vị

1.2.2.3. Giá trị mode

Mode là giá trị có tần suất xuất hiện lớn nhất trong tập dữ liệu đang xét. Giả sử tập dữ liệu đang xét chứa N giá trị khác nhau $x_1, x_2, \ldots, x_N$. Gọi tần suất xuất hiện của giá trị x_i là $f(x_i)$. Khi đó:

$$f(\text{mode}) = \max_{1 \le i \le n} \{f(x_i)\}$$

Một tập dữ liệu có thể có nhiều giá trị mode.

1.2.2.4. Khoảng trung bình (midrange)

Khoảng trung bình cũng có thể được sử dụng để xác định độ tập trung của dữ liệu. Khoảng trung bình được xác định là trung bình cộng của các giá trị lớn nhất và nhỏ nhất trong tập dữ liệu.

$$midrange = \frac{\max + \min}{2}$$

1.2.3. Đánh giá sự phân ly của dữ liệu

1.2.3.1. K-thập phân vị và tứ phân vị

K-thập phân vị (k^{th} percentile) của của một tập dữ liệu có thứ tự là một giá trị x_i có tính chất: K% các mục dữ liệu trong tập dữ liệu có giá trị bằng hoặc nhỏ hơn x_i.

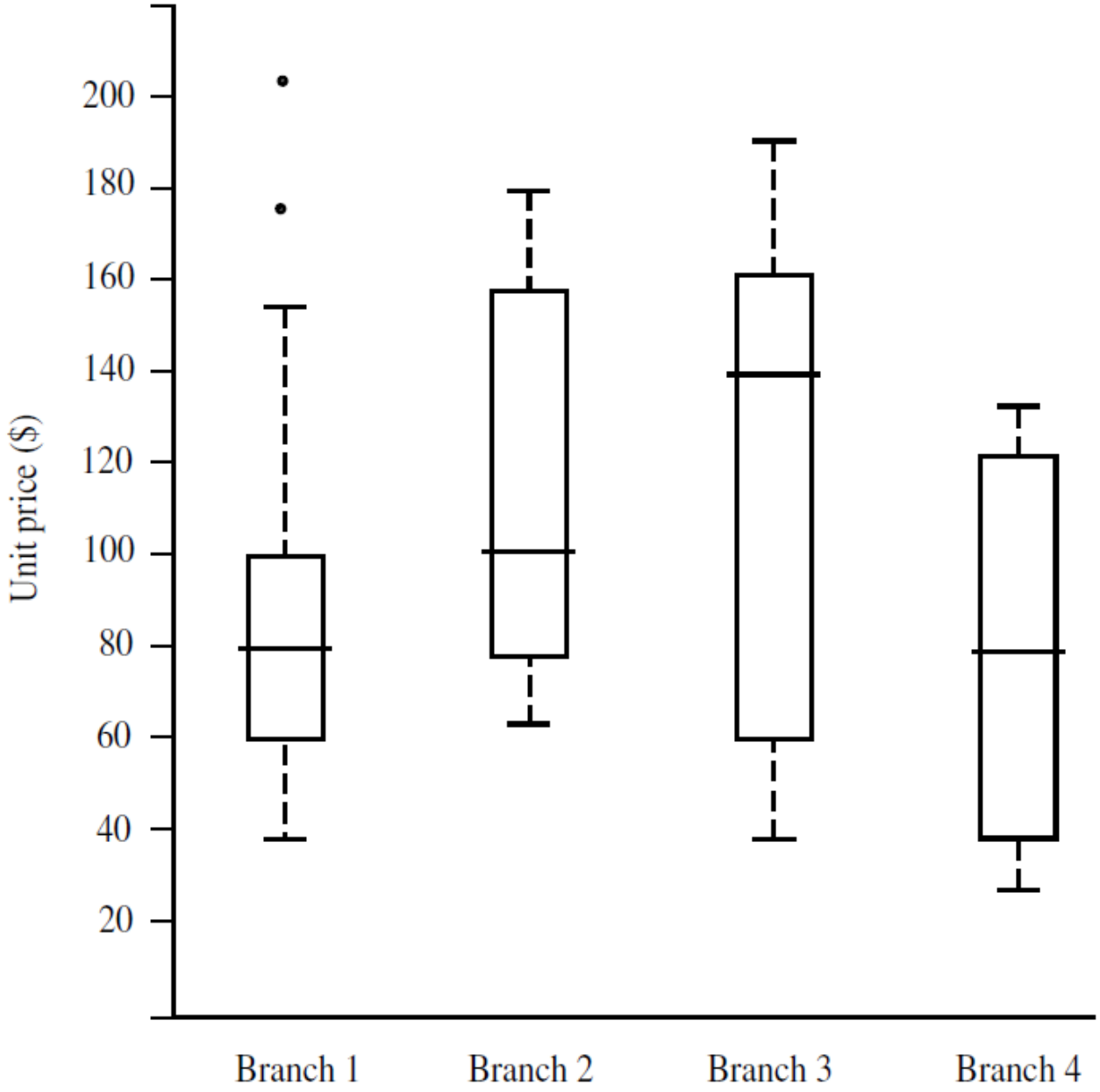
- **Nhất-tứ phân vị (first quartile) là 25-thập phân vị (Q_1)**
- **Tam-tứ phân vị (third quartile) là 75-thập phân vị (Q_3)**
- **Khoảng liên tứ phân vị (interquartile range - IQR):**

$$IQR = Q_3 - Q_1$$

$\Longrightarrow$ Có 5 giá trị biểu diễn tóm tắt dữ liệu: Min, Q_1, Median, Q_2, Max.

Biểu diễn phân bố bằng biểu đồ cột (boxplots):

- *Cuối của mỗi cột biểu diễn là giá trị tứ phân vị và chiều dài của mỗi cột là khoảng liên tứ phân vị.*
- *Trung vị được ký hiệu bằng một đường gạch ngang giữa cột biểu diễn.*
- *Hai đường thẳng bên ngoài cột mở rộng tới vị trí biểu diễn cho giá trị lớn nhất và nhỏ nhất của dãy.*


Unit price ($)
200
180
160
140
120
100
80
60
40
20
Branch 1
Branch 2
Branch 3
Branch 4

1.2.3.2. Phương sai và độ lệch chuẩn

Phương sai (variance) của N giá trị x_1, x_2,…, x_N được xác định bằng công thức:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 = \frac{1}{N}\left[\sum x_i^2 - \frac{1}{N}(\sum x_i)^2\right]$$

$\bar{x}$: giá trị trung bình của N giá trị.

Độ lệch chuẩn (standard deviation) σ được xác định bằng căn bậc 2 của phương sai.

Lưu ý:

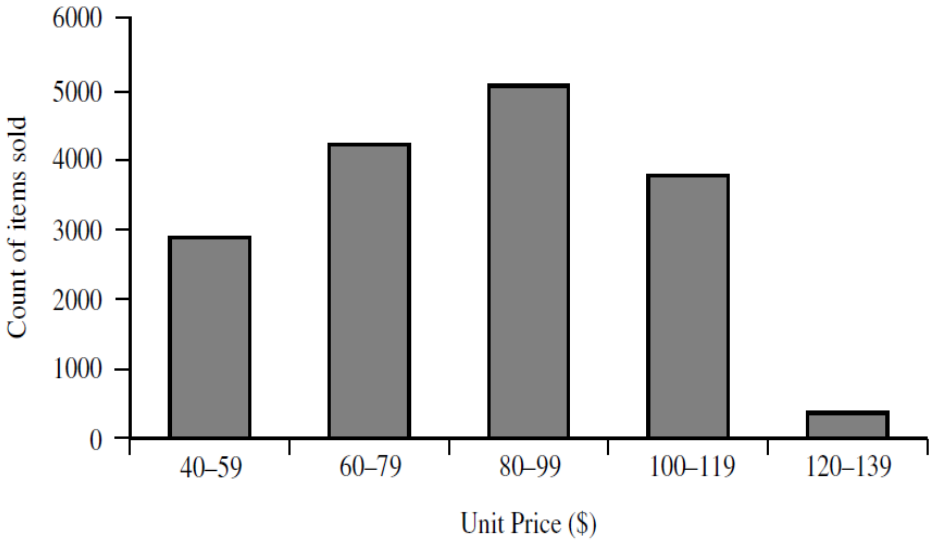
- *Độ lệch chuẩn phân bổ xung quanh giá trị trung bình và chỉ được sử dụng khi giá trị trung bình được chọn làm giá trị đặc trưng cho trung tâm của dãy.*
- *σ = 0 có nghĩa là không có sự phân bố phương sai, tất cả các giá trị đều bằng nhau.*

1.2.4. Biểu diễn tóm tắt mô tả dữ liệu dưới dạng đồ thị

1.2.4.1. Biểu đồ tần suất (frequency histograms)

- Là phương pháp biểu diễn tóm tắt sự phân bố của một thuộc tính cho trước nào đó dưới dạng trực quan.
- Biểu đồ tần suất ứng với một thuộc tính A nào đó sẽ chia sự phân bố dữ liệu của A thành các tập không giao nhau gọi là bucket (thường thì độ rộng của các bucket là bằng nhau).
- Mỗi bucket được biểu diễn bằng một hình chữ nhật có chiều cao tương ứng là số lượng hay tần suất của các giá trị có trong bucket.

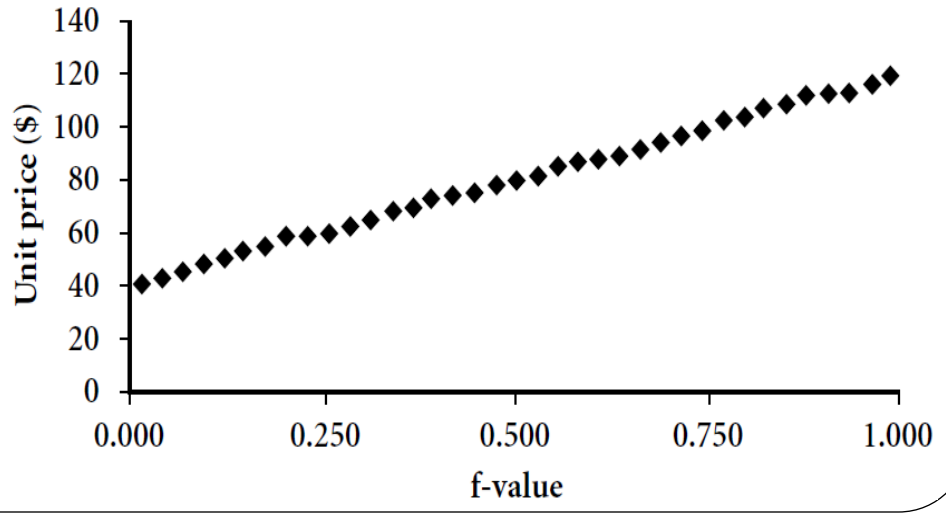
Unit price ($)	*Count of items sold*
40	275
43	300
47	250
..	..
74	360
75	515
78	540
..	..
115	320
117	270
120	350



1.2.4.2. Đồ thị phân vị (quantile plot):

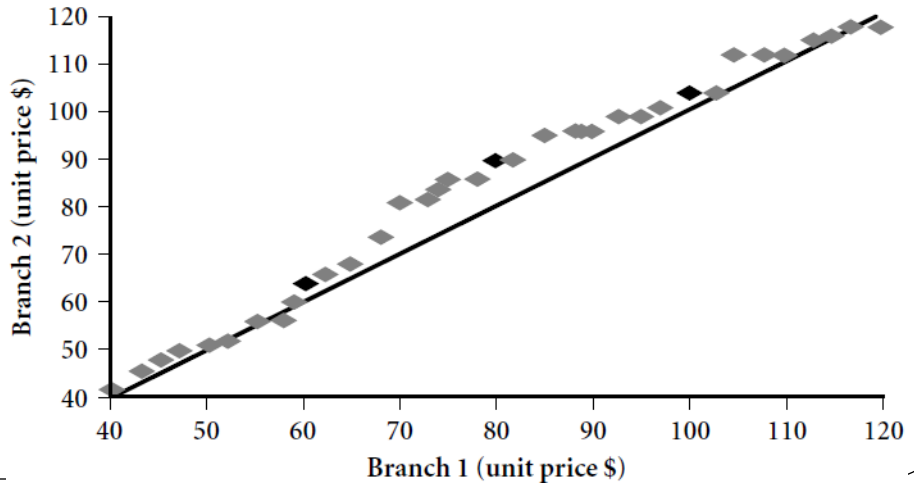
- Là cách thức đơn giản và hiệu quả để cho ta một cái nhìn về sự phân bố của dữ liệu đơn biến.
- Cho phép biểu diễn toàn bộ dữ liêu ứng với thuộc tính cho trước.
- Biểu diễn đồ thị thông tin phân vị (quantile information).
- Kỹ thuật biểu diễn:
  - Dãy giá trị x_i sẽ được sắp tăng dần từ x_1 tới x_N. Mỗi giá trị x_i sẽ được đi kèm với một giá trị f_i là tỷ lệ phần trăm các giá trị dữ liệu trong dãy nhỏ hơn hoặc bằng x_i.
  - Giá trị f_i có thể tính bởi công thức: $f_i = \frac{i - 0.5}{N}$
  - Trên đồ thị, x_i được biểu diễn theo f_i.

Unit price ($)	Count of items sold
40	275
43	300
47	250
..	..
74	360
75	515
78	540
..	..
115	320
117	270
120	350



1.2.4.3. Đồ thị song phân vị (quantile-quantile plot):

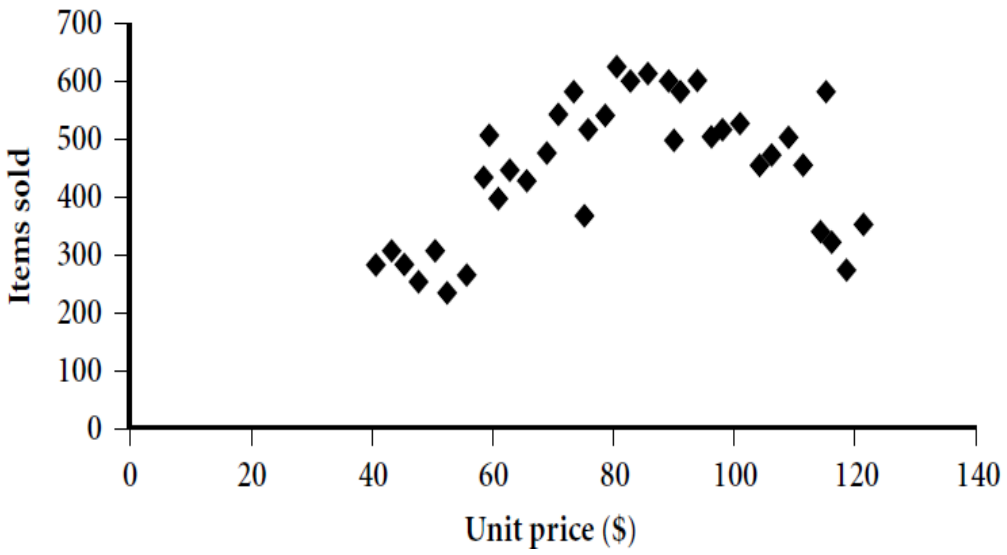
- Biểu diễn mối liên hệ giữa phân vị của một phân bố đơn biến này với phân vị của một phân bố đơn biến khác.
- Đây là công cụ trực quan mạnh mẽ cho phép quan sát sự thay đổi khi chuyển từ phân bố này sang một phân bố khác.
- Kỹ thuật biểu diễn:
  - Giả sử chúng ta có hai dãy giá trị của cùng một biến ngẫu nhiên được thu thập độc lập nhau: dãy $x = \{x_1, x_2, \ldots, x_N\}$ và dãy $y = \{y_1, y_2, \ldots, y_M\}$
  - Nếu N = M: biểu diễn Y_i theo X_i trong đó X_i ,Y_i tương ứng là các phân vị của dãy x và dãy y xác định theo công thức (i – 0.5)/N.
  - Nếu M < N: biểu diễn Y_i theo X_i và chỉ có M điểm biểu diễn trên đồ thị. Trong đó X_i ,Y_i tương ứng là các phân vị của dãy x và dãy y xác định theo công thức (i – 0.5)/M.



1.2.4.4. Đồ thị phân tán (scatter plot):

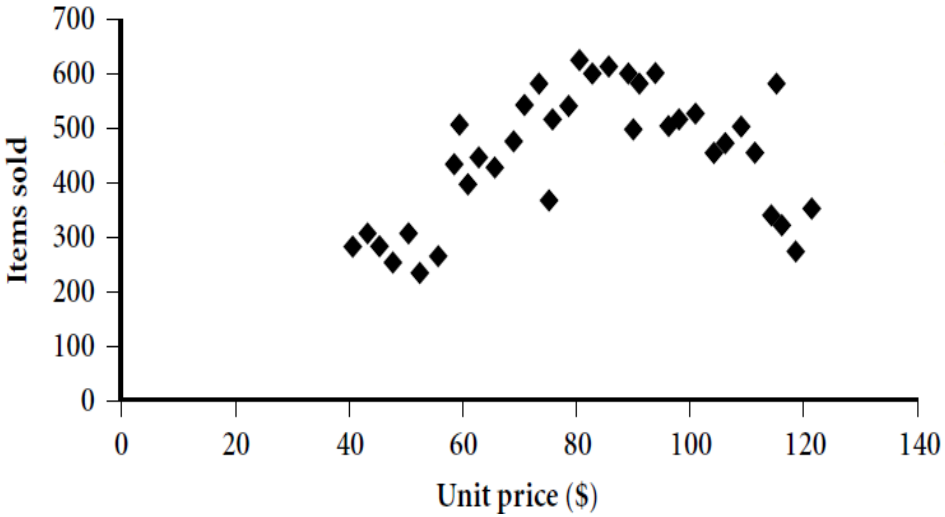
- Là phương pháp hiệu quả để xác định xem liệu có xuất hiện mối quan hệ, các mẫu hay xu hướng giữa 02 thuộc tính mang giá trị số hay không.
- Mỗi cặp giá trị được biểu diễn bằng một cặp tọa độ (tương ứng với một điểm trên mặt phẳng tọa độ).
- Cung cấp một cái nhìn sơ bộ về dữ liệu để thấy được các cụm điểm và các giá trị kỳ dị (outliers) cũng như phát hiện khả năng tồn tại của các mỗi liên hệ phụ thuộc.

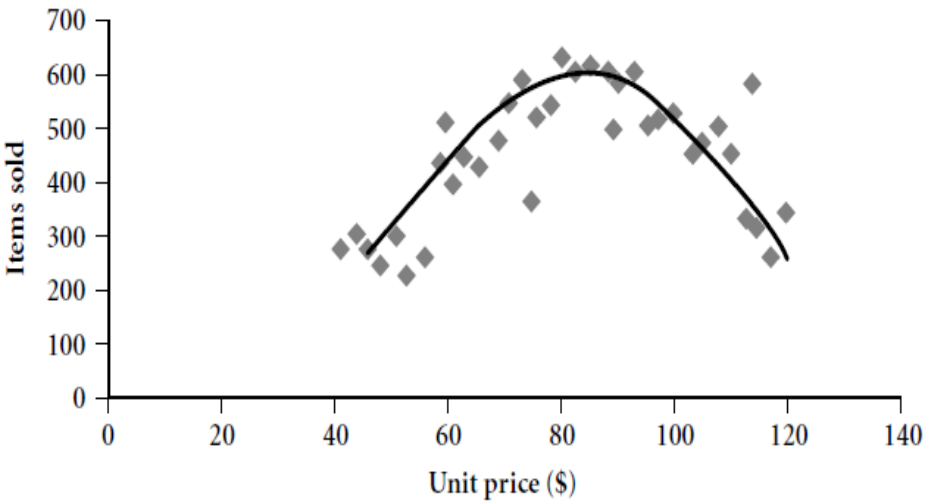
Unit price ($)	*Count of items sold*
40	275
43	300
47	250
..	..
74	360
75	515
78	540
..	..
115	320
117	270
120	350



1.2.4.5. Đường loess

- Là công cụ biểu diễn đồ thị quan trọng cho phép bổ sung một đường cong "trơn" vào đồ thị phân tán nhằm cung cấp một sụ hình dung tốt hơn về mẫu độc lập (loess = local regression: hồi quy cục bộ).
- Để khớp với đường cong hồi quy, các giá trị cần được thiết lập với 02 tham số là α-tham số độ trơn và λ-bậc của đa thức hồi quy.
- Cần chọn α để tạo ra một đường cong "trơn" nhất có thể nhưng không làm biến dạng mẫu dữ liệu được phản ánh.





1.3. LÀM SẠCH DỮ LIỆU

Làm sạch dữ liệu (data cleaning) là kỹ thuật giúp xử lý sự thiếu vắng giá trị, loại bỏ nhiễu và các giá trị không mong muốn cũng như giải quyết vấn đề không nhất quán dữ liệu.

1.3.1. Xử lý sự thiếu vắng giá trị (missing values)

A. ***Bỏ qua các bản ghi***: vd: thiếu vắng nhãn phân lớp. Phương pháp này thực sự không hiệu quả trừ phi trong 1 bản ghi có sự thiếu vắng giá trị ở một vài thuộc tính.

B. ***Điền các giá trị thiếu một cách thủ công***: Phương pháp này tiêu tốn nhiều thời gian và không khả thi với các tập dữ liệu lớn có nhiều giá trị thiếu vắng.

C. ***Sử dụng các giá trị (hằng) quy ước để thay cho các giá trị thiếu***: Thay thế các giá trị thiếu bằng các giá trị (hằng) quy ước giống nhau (vd: “unknown”). Cách này có thể gây hiểu lầm cho hệ thống KPDL khi nghĩ rằng “unknown” là một giá trị đáng quan tâm.

D. ***Sử dụng giá trị trung bình để thay cho các giá trị thiếu***: Sử dụng giá trị trung bình của một thuộc tính để thay thế cho các giá trị thiếu trên thuộc tính đó.

D. ***Sử dụng giá trị trung bình trên phân lớp để thay thế cho giá thị thiếu trong phân lớp*****:** thay thế giá trị bị thiếu bằng trị trung bình của các giá trị tương ứng trong cùng phân lớp.

E. ***Sử dụng giá trị có xác suất cao nhất (most probable) để thay thế cho giá trị thiếu*****:** Giá trị này có thể xác định thông qua hồi quy, các công cụ suy diễn dựa trên chuẩn hóa Bayes hoặc suy luận nhờ cây quyết định.

1.3.2. Xử lý dữ liệu nhiễu (noisy data)

Nhiễu (noise) là những lỗi ngẫu nhiên hoặc những giá trị "lệch chuẩn".

⟹ ***Làm thế nào để làm "mượt" (smooth) dữ liệu và loại bỏ nhiễu?***

A. <u>"Đóng thùng" (binning)</u>:

- Là phương pháp làm "trơn" một giá trị dữ liệu đã được sắp xếp dựa trên các giá trị xung quanh (làm "trơn" cục bộ).
- Các giá trị dữ liệu đã được sắp xếp sẽ được phân chia vào các "thùng chứa" (gọi là bin/bucket) có kích thước bằng nhau. Có 2 kiểu phân chia:
  - ❖ **Equal-frequency**: Các "thùng chứa" chứa số giá trị như nhau.
  - ❖ **Equal-width**: Các "thùng chứa" có khoảng giá trị biến động (từ giá trị min đến giá trị max của thùng) là như nhau.
- Có 2 kỹ thuật phổ biến:
  - ❖ **Làm trơn trung bình/trung vị (smoothing by bin means/median)**: mỗi giá tri trong "thùng chứa" sẽ được thay thế bằng trung bình cộng (hoặc trung vị) của toàn bộ các giá trị ban đầu có trong "thùng chứa" đó.
  - ❖ **Làm trơn dựa trên biên (smoothing by boundaries)**: giá trị lớn nhất hoặc nhỏ nhất trong "thùng chứa" sẽ được chọn làm biên. Mỗi giá trị trong thùng chứa sẽ được thay thế bằng giá trị biên gần nhất.

Sorted data for *price* (in dollars): 4, 8, 15, 21, 21, 24, 25, 28, 34

Partition into (equal-frequency) bins:

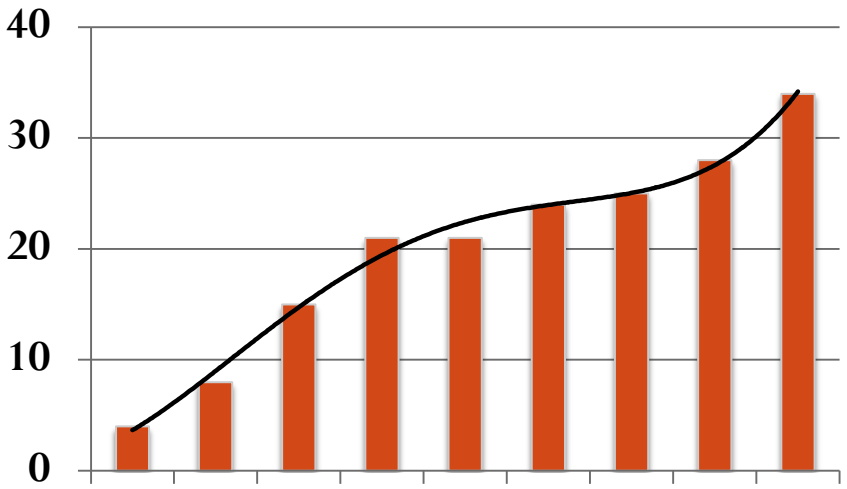
Bin 1: 4, 8, 15
Bin 2: 21, 21, 24
Bin 3: 25, 28, 34

Smoothing by bin means:

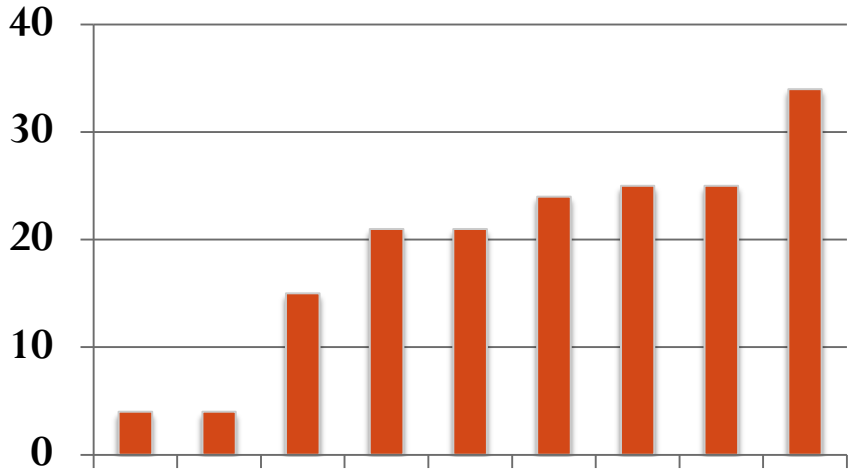
Bin 1: 9, 9, 9
Bin 2: 22, 22, 22
Bin 3: 29, 29, 29

Smoothing by bin boundaries:

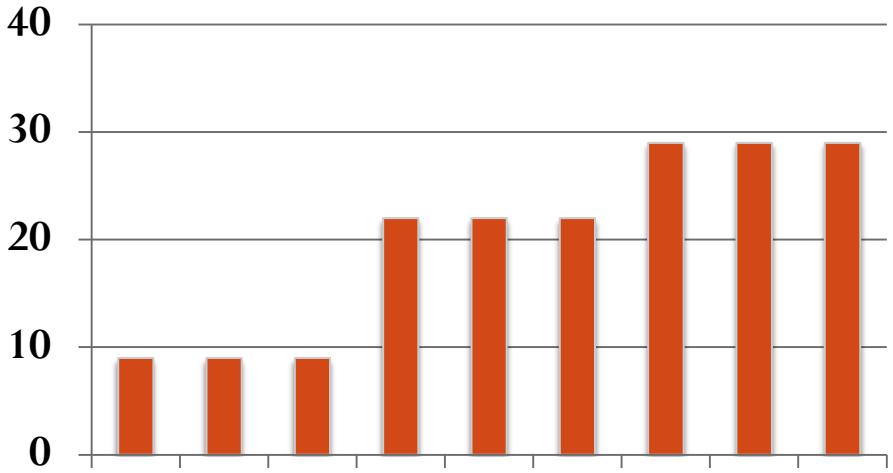
Bin 1: 4, 4, 15
Bin 2: 21, 21, 24
Bin 3: 25, 25, 34



Dữ liệu được sắp xếp



Làm trơn dựa trên biên



Làm trơn trung bình

B. *Hồi quy (regression):*

- Dữ liệu có thể được làm trơn bằng cách khớp dữ liệu với một hàm hồi quy.
- Hồi quy tuyến tính đòi hỏi phải tìm ra đường thẳng tối ưu khớp với 2 biến (thuộc tính). Từ đó, một thuộc tính có thể được sử dụng để dự đoán thuộc tính còn lại.
- Hồi quy tuyến tính kép là sự mở rộng của hồi quy tuyến tính khi mà có nhiều hơn 02 biến (thuộc tính) và dữ liệu sẽ được khớp với đồ thị không gian là một mặt đa chiều.

C. *Phân cụm (clustering):*

- Giá trị bất thường (outliers) có thể được phát hiện bằng kỹ thuật phân cụm khi mà các giá trị tương tự nhau được đưa vào cùng nhóm (cụm). Các giá trị không thuộc về một cụm nào cả có thể xem là bất thường.

1.4. TÍCH HỢP VÀ CHUYỂN DẠNG DỮ LIỆU

1.4.1. Tích hợp dữ liệu (Data Integration)

- Kết hợp dữ liệu từ nhiều nguồn khác nhau thành một kho dữ liệu thống nhất.
- Các nguồn dữ liệu khác nhau: cơ sở dữ liệu, data cube, tập tin phẳng,…
- Các vấn đề phải đối mặt:
  - ❖ **Tích hợp lược đồ (shema integration) và khớp các đối tượng (object matching)**: cùng một thực thể trong thế giới thực có thể được phản ánh trong dữ liệu từ các nguồn khác nhau ⟹ cần phải khớp lại các đối tượng này. VD: Vấn đề về định danh thực thể
  - ❖ **Sự dư thừa (redundancy):**
    - ✓ Một thuộc tính có thể dư thừa nếu có thể được suy diễn từ một hay một tập các thuộc tính khác.
    - ✓ Sự không nhất quán trong thuộc tính hay do cách đặt tên có thể gây ra sự dư thừa trong tập dữ liệu kết quả.
    - ✓ Dư thừa dữ liệu có thể được phát hiện thông qua phân tích tương quan (correlation analysis).

Phân tích dựa trên hệ số tương quan

- Dựa trên các dữ liệu đã có, phân tích tương quan có thể cho thấy mức độ mà một thuộc tính có thể được suy diễn hoặc được quyết định bởi một thuộc tính khác.
- **Hệ số tương quan:** dùng để đánh giá độ tương quan giữa 02 thuộc tính. Cụ thể, hệ số tương quan giữa 02 thuộc tính A và B được xác định:

$$r_{A,B} = \frac{\sum_{i=1}^{N}(a_i - \bar{A})(b_i - \bar{B})}{N\sigma_A\sigma_B} = \frac{\sum_{i=1}^{N}(a_i b_i) - N\bar{A}\bar{B}}{N\sigma_A\sigma_B}$$

Trong đó:

- ✓ N: số bộ dữ liệu.
- ✓ a_i, b_i là các giá trị tương ứng với 02 thuộc tính A và B trong bộ i.
- ✓ $\bar{A}, \bar{B}$ tương ứng là các giá trị trung bình trên A và B.
- ✓ σ_A, σ_B tương ứng là độ lệch chuẩn của A và B.

Ta luôn có $-1 \le r_{A,B} \le 1$ và:

- **Nếu $r_{A,B} > 0$**: A, B có mối tương quan dương (giá trị ứng với A tăng thì giá trị ứng với B cũng tăng). Giá trị $r_{A,B}$ càng lớn thể hiện tính tương quan giữa 02 thuộc tính càng mạnh $\Longrightarrow$ Có thể loại bỏ một trong 02 thuộc tính (A hoặc B) vì nó là dư thừa.
- **Nếu $r_{A,B} = 0$**: Không tồn tại mối liên hệ tương quan. A và B là 02 thuộc tính hoàn toàn độc lập.
- **Nếu $r_{A,B} < 0$**: A, B có mối tương quan âm (giá trị ứng với A tăng thì giá trị ứng với B giảm và ngược lại) $\Longrightarrow$ A và B là 02 thuộc tính trái ngược nhau.

Phân tích tương quan đối với dữ liệu rời rạc

Mối quan hệ tương quan giữa 02 thuộc tính A và B có thể được đặc trưng bởi phép đo Khi – Bình phương (Chi-square) χ^2

- **Giả sử thuộc tính A có c giá trị khác nhau a_1, a_2,…, a_c và B có r giá trị khác nhau b_1, b_2,…,b_r.**
- **Các bộ dữ liệu đặc trưng bởi A, B được biểu diễn dưới dạng một bảng ngẫu nhiên (contingency table) với các cột là c giá trị khác nhau của A và các dòng là r giá trị khác nhau của B.**
- **Ký hiệu (A_i, B_j) là sự kiện thuộc tính A nhận giá trị a_i và thuộc tính B nhận giá trị b_j. Mỗi sự kiện (A_i, B_j) có thể có sẽ chiếm trọn một ô trong bảng.**
- **Giá trị Khi – Bình phương χ^2 có thể được xác định qua công thức:**

$$\chi^2 = \sum_{i=1}^{c} \sum_{j=1}^{r} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

Trong đó:

- **o_{ij} là tần suất quan sát được hay tần suất biểu kiến (observed frequency) của sự kiện (A_i, B_j)**
- **e_{ij} là tần xuất kỳ vọng (expected frequency) của sự kiện (A_i, B_j) .**

Tần xuất kỳ vọng (expected frequency) của sự kiện (A_i, B_j) có thể tính bởi công thức:

$$e_{ij} = \frac{count(A = a_i) \times count(B = b_j)}{N}$$

Trong đó:

N: số lượng các bộ dữ liệu.

count($A=a_i$): số lượng các bộ có thuộc tính A nhận giá trị a_i.

count($B=b_j$): số lượng các bộ có thuộc tính B nhận giá trị b_j.

Chú ý:

Độ đo Khi – Bình phương dùng để kiểm tra giả thiết về tính độc lập của 02 thuộc tính A và B. Việc kiểm tra này dựa trên mức độ chú ý (significance level) với (r-1)(c-1) bậc tự do.

	male	*female*	Total
fiction	250 (90)	200 (360)	450
non_fiction	50 (210)	1000 (840)	1050
Total	300	1200	1500

$$e_{11} = \frac{count(male) \times count(fiction)}{N} = \frac{300 \times 450}{1500} = 90$$

$$\begin{aligned} \chi^2 &= \frac{(250-90)^2}{90} + \frac{(50-210)^2}{210} + \frac{(200-360)^2}{360} + \frac{(1000-840)^2}{840} \\ &= 284.44 + 121.90 + 71.11 + 30.48 = 507.93. \end{aligned}$$

Với số bậc tự do là (2-1)(2-1) = 1, mức độ chú ý là 0.001 thì để đảm bảo 02 thuộc tính A, B là độc lập, giá trị χ^2 = 10.828 (*đề nghị SV tham khảo thêm các giáo trình về xác suất thống kê*)

$\Longrightarrow$ Giá trị tính được là 507.93 > 10.828 nên A và B là 02 thuộc tính phụ thuộc chặt chẽ.

1.4.2. Chuyển dạng dữ liệu (Data Transformation)

Dữ liệu được chuyển đổi hoặc hợp nhất thành các dạng phù hợp cho việc khai phá. Chuyển dạng dữ liệu liên quan tới các vấn đề sau đây:

- **Làm trơn (Smoothing)**: Loại bỏ các nhiễu (noisy) khỏi dữ liệu. Các kỹ thuật được sử dụng bao gồm: đóng thùng (binning), hồi quy (regression), phân cụm (clustering).
- **Gộp nhóm (Aggregation)**: các thao tác tóm tắt hay gộp nhóm được áp dụng với dữ liệu. Bước này thường được sử dụng để xây dựng data cube cho phân tích dữ liệu từ nhiều nguồn.
- **Khởi tạo dữ liệu (Generalization of the data)**: dữ liệu thô được thay thế bởi các các khái niệm ở mức cao hơn thông qua việc sử dụng lược đồ khái niệm.
- **Xây dựng thuộc tính (Attribute construction)**: các thuộc tính mới được xây dựng và thêm vào từ tập thuộc tính đã có để hỗ trợ quá trình khai phá (tăng độ chính xác và sự dễ hiểu của cấu trúc trong dữ liệu nhiều chiều (high-dimensional data)). Bằng cách kết hợp các thuộc tính $\Longrightarrow$ phát hiện ra các thông tin bị thiếu liên quan đến mối quan hệ giữa các thuộc tính (hữu ích cho quá trình khai phá).

- **Chuẩn hóa (Normalization)**: Dữ liệu thuộc tính được chuyển đổi tương ứng với các phạm vi biểu diễn nhỏ hơn như [-1,1] hoặc [0,1].

***Chuẩn hóa min-max**: thực hiện việc chuyển đổi tuyến tính dựa trên dữ liệu gốc. Gọi min_A, max_A là giá trị lớn nhất và nhỏ nhất của thuộc tính A. Chuẩn hóa min-max sẽ ánh xạ một giá trị v của A tương ứng với một giá trị v' trong khoảng [new_min_A , new_max_A] thông qua công thức:*

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

***Ví dụ**: Giả sử giá trị lớn nhất và nhỏ nhất của thuộc tính income là \$12,000 và \$98,000. Người ta định ánh xạ miền giá trị của thuộc tính income tương ứng với khoảng [0.0,1.0]. Hỏi giá trị v = \$73,000 của income sẽ tương ứng với giá trị ánh xạ v' bằng bao nhiêu trong khoảng [0.0,1.0]?*

min_A = \$12,000
max_A = \$98,000
new_min_A = 0.0
new_max_A = 0.1
v = \$73,000

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

$$= \frac{73{,}600 - 12{,}000}{98{,}000 - 12{,}000}(1.0 - 0) + 0 = 0.716$$

***Chuẩn hóa z-score**: các giá trị ứng với thuộc tính A được chuẩn hóa dựa trên giá trị trung bình và độ lệch chuẩn của A. Một giá trị v của A sẽ được chuẩn hóa tương ứng với một giá trị v' thông qua công thức:*

$$v' = \frac{v - \bar{A}}{\sigma_A}$$

Chuẩn hóa z-score rất hữu dụng khi:

- *Không biết giá trị lớn nhất và nhỏ nhất thực tế của thuộc tính A.*
- *Các giá trị kỳ dị (outliers) chi phối chuẩn hóa min-max*

***Ví dụ**: Giả sử rằng giá trị trung bình và độ lêch chuẩn của thuộc tính income tương ứng là $54,000 và $16,000. Một giá trị v = $73,600 của income sẽ được chuẩn hóa tương ứng với giá trị v' bằng bao nhiêu?*

$$v' = \frac{v - \bar{A}}{\sigma_A} = \frac{73{,}600 - 54{,}000}{16{,}000} = 1.225$$

Chuẩn hóa thập phân (decimal scaling): *dịch chuyển dấu phẩy thập phân của các giá trị ứng với thuộc tính A. Số vị trí di chuyển phụ thuộc vào giá trị tuyệt đối lớn nhất của A. Một giá trị v của A được chuẩn hóa thập phân tương ứng với một giá trị v' theo công thức:*

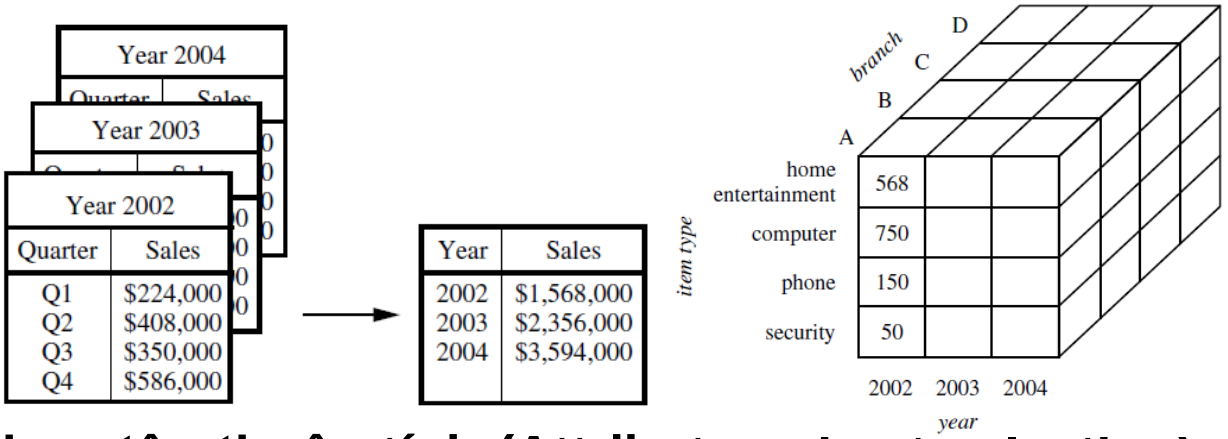
$$v' = \frac{v}{10^j}$$

(*j là số nguyên nhỏ nhất sao cho* $Max(|v'|) < 1$)

Ví dụ: *Giả sử thuộc tính A có miền giá trị là [-986,917]. Giá trị tuyệt đối lớn nhất của A là 986. Như vậy, ta chọn j = 3. Khi đó thì một giá trị v = 817 sẽ được chuẩn hóa thành v' = 0.817*

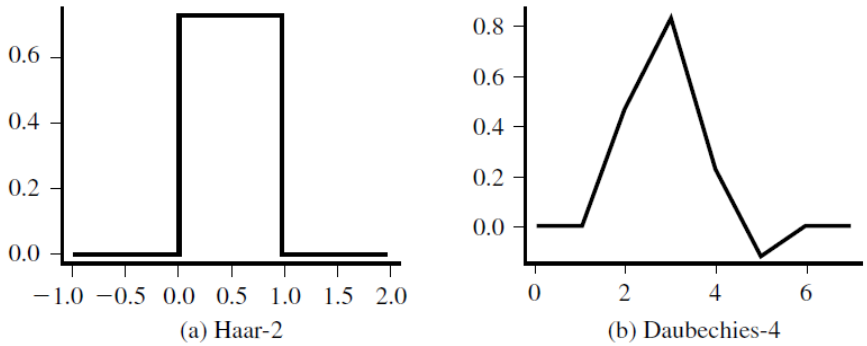
1.5. RÚT GỌN DỮ LIỆU

1.5.1. Gộp nhóm dữ liệu dưới dạng data cube: Các thao tác gộp nhóm sẽ được áp dụng trên dữ liệu để tạo ra một data cube.



1.5.2. Lựa chọn tập thuộc tính (Attribute subset selection): Các thuộc tính thừa hoặc không thích hợp sẽ được phát hiện và loại bỏ.

1.5.3. Giảm số chiều dữ liệu (Dimensionality reduction): Các cơ chế mã hóa (encoding) sẽ được áp dụng để làm giảm kích thước dữ liệu.



1.5.4. Giảm biểu diễn số lớn (Numerosity reduction): Dữ liệu sẽ được thay thế hoặc tính toán thông qua những cách thức biểu diễn dữ liệu khác gọn hơn, ví dụ như các mô hình tham số (parametric models) hoặc các phương pháp không tham số (nonparametric methods) như phân cụm, lấy mẫu, sử dụng histogram.

1.5.5. Rời rạc hóa dữ liệu (discretization) và tạo lược đồ khái niệm (concept hierarchy generation):

- Các giá trị dữ liệu thô ứng với các thuộc tính được thay thế bằng các khoảng (range) hoặc các mức khái niệm (conceptual levels) cao hơn.
- Rời rạc hóa dữ liệu được xem là một dạng thức của việc giảm biểu diễn số lớn và rất hữu dụng trong việc tạo lược đồ khái niệm.
- Rời rạc hóa dữ liệu và tạo lược đồ khái niệm được xem là những công cụ mạnh mẽ cho khai phá dữ liệu. Chúng cho phép thực hiện công việc khai phá ở những cấp độ trừu tượng khác nhau.

Q & A