KHAI PHÁ DỮ LIỆU
(DATAMINING)

TIỀN XỬ LÝ DỮ LIỆU

Mai Xuân Hùng

Nội dung

- Tổng quan
- Làm sạch dữ liệu
- Tích hợp dữ liệu
- Biến đổi dữ liệu
- Thu giảm dữ liệu
- Rời rạc hóa dữ liệu
- Tạo cây phân cấp ý niệm
- Tổng kết

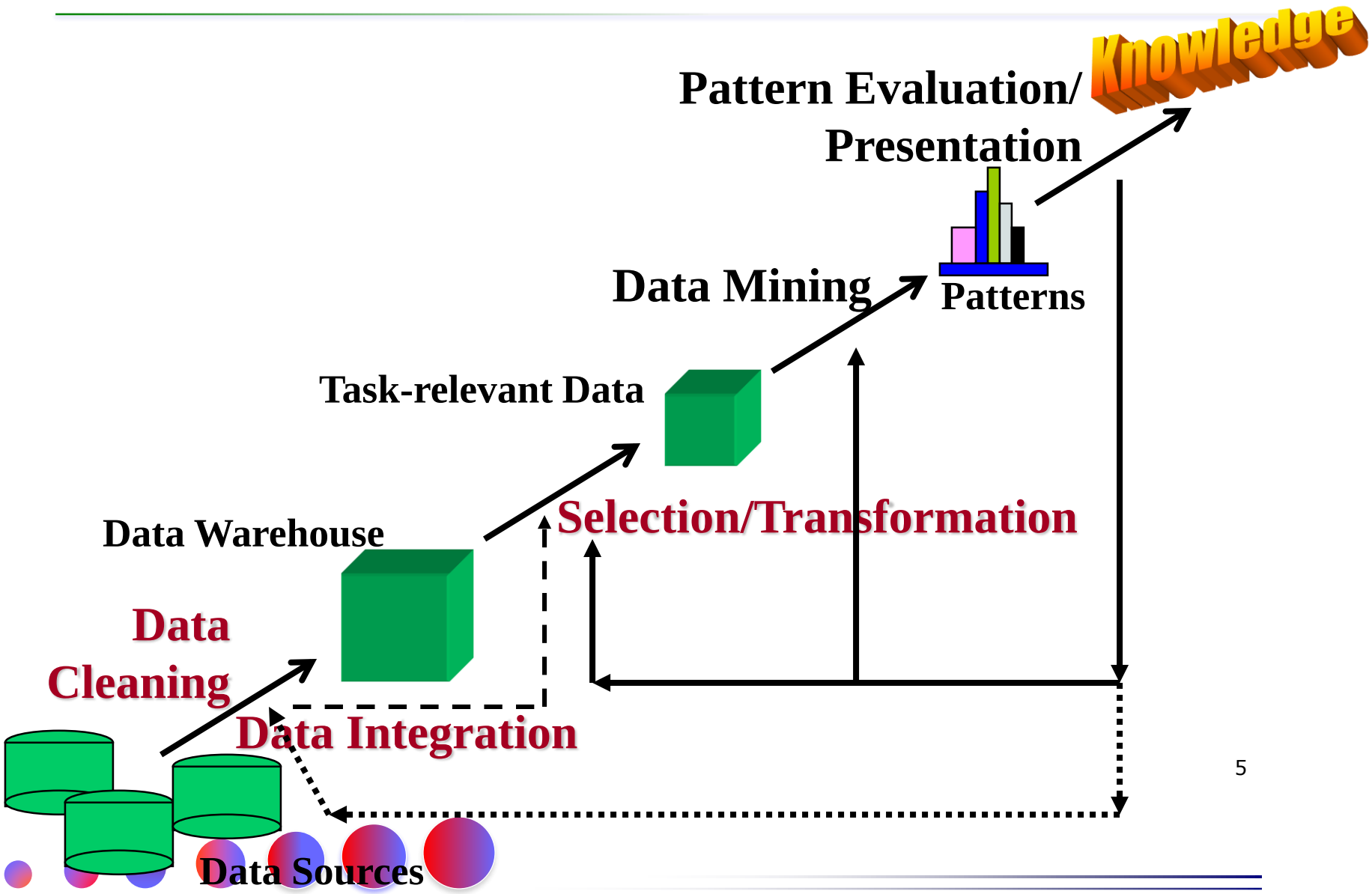
Tổng quan

- Giai đoạn tiền xử lý dữ liệu
  - Quá trình xử lý dữ liệu thô/gốc (raw/original data) nhằm cải thiện chất lượng dữ liệu (quality of the data) và do đó, cải thiện chất lượng của kết quả khai phá.
    - Dữ liệu thô/gốc
      - Có cấu trúc, bán cấu trúc, phi cấu trúc
      - Được đưa vào từ các nguồn dữ liệu trong các hệ thống xử lý tập tin (file processing systems) và/hay các hệ thống cơ sở dữ liệu (database systems)
    - Chất lượng dữ liệu (data quality): tính chính xác, tính hiện hành, tính toàn vẹn, tính nhất quán

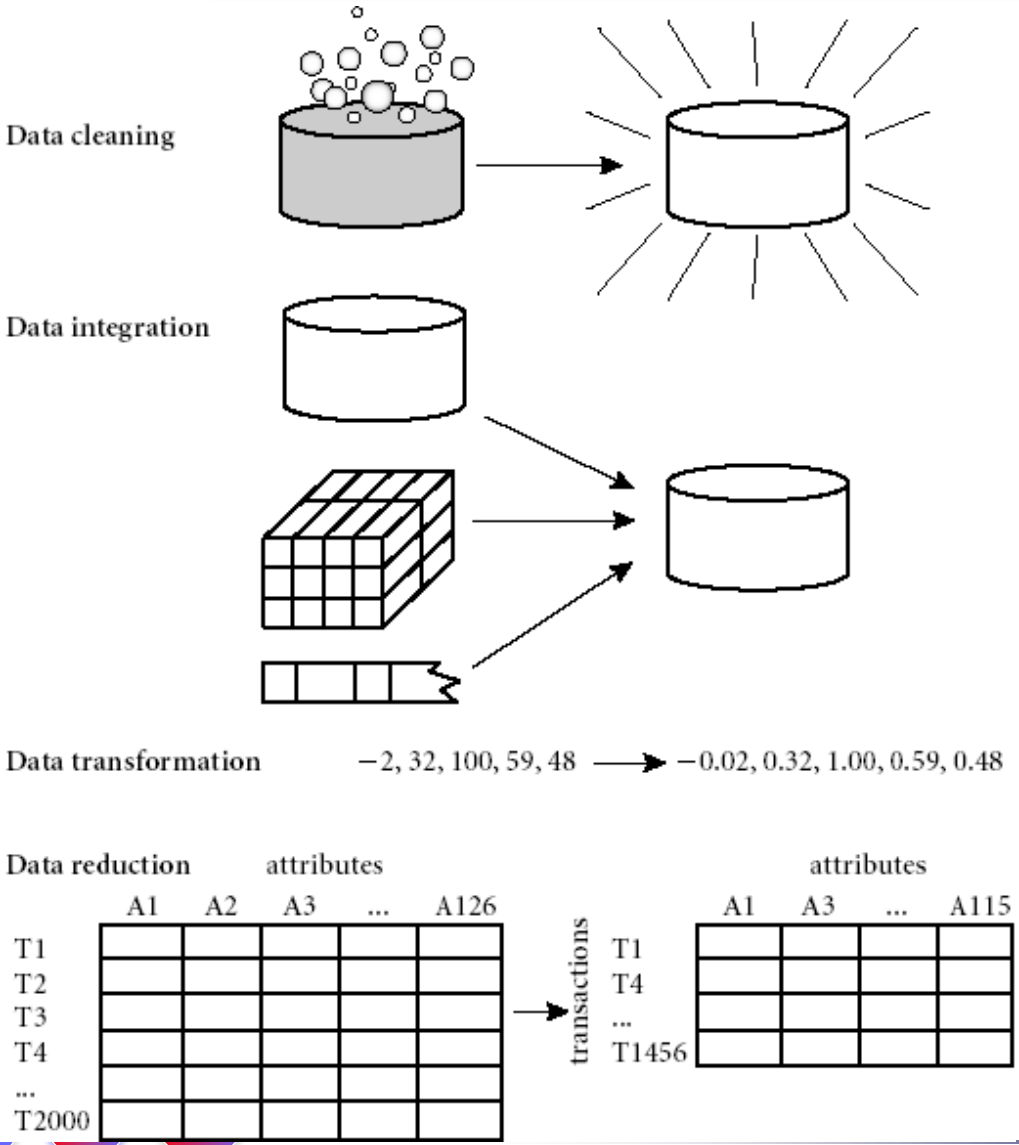
Tổng quan

- Chất lượng dữ liệu (data quality)
  - Tính chính xác (accuracy): giá trị được ghi nhận đúng với giá trị thực.
  - Tính hiện hành (currency/timeliness): giá trị được ghi nhận không bị lỗi thời.
  - Tính toàn vẹn (completeness): tất cả các giá trị dành cho một biến/thuộc tính đều được ghi nhận.
  - Tính nhất quán (consistency): tất cả giá trị dữ liệu đều được biểu diễn như nhau trong tất cả các trường hợp.

Tổng quan về giai đoạn tiền xử lý dữ liệu



Tổng quan về giai đoạn tiền xử lý dữ liệu



Tổng quan

- Các kỹ thuật tiền xử lý dữ liệu
  - Làm sạch dữ liệu (data cleaning/cleansing): loại bỏ nhiễu (remove noise), hiệu chỉnh những phần dữ liệu không nhất quán (correct data inconsistencies)
  - Tích hợp dữ liệu (data integration): trộn dữ liệu (merge data) từ nhiều nguồn khác nhau vào một kho dữ liệu
  - Biến đổi dữ liệu (data transformation): chuẩn hoá dữ liệu (data normalization)
  - Thu giảm dữ liệu (data reduction): thu giảm kích thước dữ liệu (nghĩa là giảm số phần tử) bằng kết hợp dữ liệu (data aggregation), loại bỏ các đặc điểm dư thừa (redundant features) (nghĩa là giảm số chiều/thuộc tính dữ liệu), gom cụm dữ liệu

Tổng quan

- Các kỹ thuật tiền xử lý dữ liệu
  - Làm sạch dữ liệu (data cleaning/cleansing)
    - Tóm tắt hoá dữ liệu: nhận diện đặc điểm chung của dữ liệu và sự hiện diện của nhiễu hoặc các phần tử kì dị (outliers)
    - Xử lý dữ liệu bị thiếu (missing data)
    - Xử lý dữ liệu bị nhiễu (noisy data)

Tổng quan

- Các kỹ thuật tiền xử lý dữ liệu
  - Tích hợp dữ liệu (data integration)
    - Tích hợp lược đồ (schema integration) và so trùng đối tượng (object matching)
    - Vấn đề dư thừa (redundancy)
    - Phát hiện và xử lý mâu thuẫn giá trị dữ liệu (detection and resolution of data value conflicts)

Tổng quan

- Các kỹ thuật tiền xử lý dữ liệu
  - Biến đổi dữ liệu (data transformation)
    - Làm trơn dữ liệu (smoothing)
    - Kết hợp dữ liệu (aggregation)
    - Tổng quát hóa dữ liệu (generalization)
    - Chuẩn hóa dữ liệu (normalization)
    - Xây dựng thuộc tích (attribute/feature construction)

Tổng quan

- Các kỹ thuật tiền xử lý dữ liệu
  - Thu giảm dữ liệu (data reduction)
    - Kết hợp khối dữ liệu (data cube aggregation)
    - Chọn tập con các thuộc tính (attribute subset selection)
    - Thu giảm chiều (dimensionality reduction)
    - Thu giảm lượng (numerosity reduction)
    - Tạo phân cấp ý niệm (concept hierarchy generation) và rời rạc hóa (discretization)

Làm sạch dữ liệu

- Xử lý dữ liệu bị thiếu (missing data)
- Nhận diện phần tử biên (outliers) và giảm thiểu nhiễu (noisy data)
- Xử lý dữ liệu không nhất quán (inconsistent data)

Làm sạch dữ liệu

- Xử lý dữ liệu bị thiếu (missing data)
  - Định nghĩa của dữ liệu bị thiếu
    - Dữ liệu không có sẵn khi cần được sử dụng
  - Nguyên nhân gây ra dữ liệu bị thiếu
    - Khách quan (không tồn tại lúc được nhập liệu, sự cố, …)
    - Chủ quan (tác nhân con người)
  - Giải pháp cho dữ liệu bị thiếu
    - Bỏ qua
    - Xử lý tay (không tự động, bán tự động)
    - Dùng giá trị thay thế (tự động): hằng số toàn cục, trị phổ biến nhất, trung bình toàn cục, trung bình cục bộ, trị dự đoán, …
    - Ngăn chặn dữ liệu bị thiếu: thiết kế tốt CSDL và các thủ tục nhập liệu

Làm sạch dữ liệu

- Nhận diện phần tử biên (outliers) và giảm thiểu nhiễu (noisy data)
  - Định nghĩa
    - Outliers: những dữ liệu (đối tượng) không tuân theo đặc tính/hành vi chung của tập dữ liệu (đối tượng).
    - Noisy data: outliers bị loại bỏ (rejected/discarded outliers) như là những trường hợp ngoại lệ (exceptions).
  - Nguyên nhân
    - Khách quan (công cụ thu thập dữ liệu, lỗi trên đường truyền, giới hạn công nghệ, …)
    - Chủ quan (tác nhân con người)

Làm sạch dữ liệu

- Nhận diện phần tử biên (outliers) và giảm thiểu nhiễu (noisy data)
  - Giải pháp nhận diện phần tử biên
    - Dựa trên phân bố thống kê (statistical distribution-based)
    - Dựa trên khoảng cách (distance-based)
    - Dựa trên mật độ (density-based)
    - Dựa trên độ lệch (deviation-based)
  - Giải pháp giảm thiểu nhiễu
    - Binning
    - Hồi quy (regression)
    - Phân tích cụm (cluster analysis)

Làm sạch dữ liệu

- Giải pháp giảm thiểu nhiễu
  - Binning (by bin means, bin median, bin boundaries)
    - Dữ liệu có thứ tự
    - Phân bố dữ liệu vào các bins (buckets)
    - Bin boundaries: trị min và trị max

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:

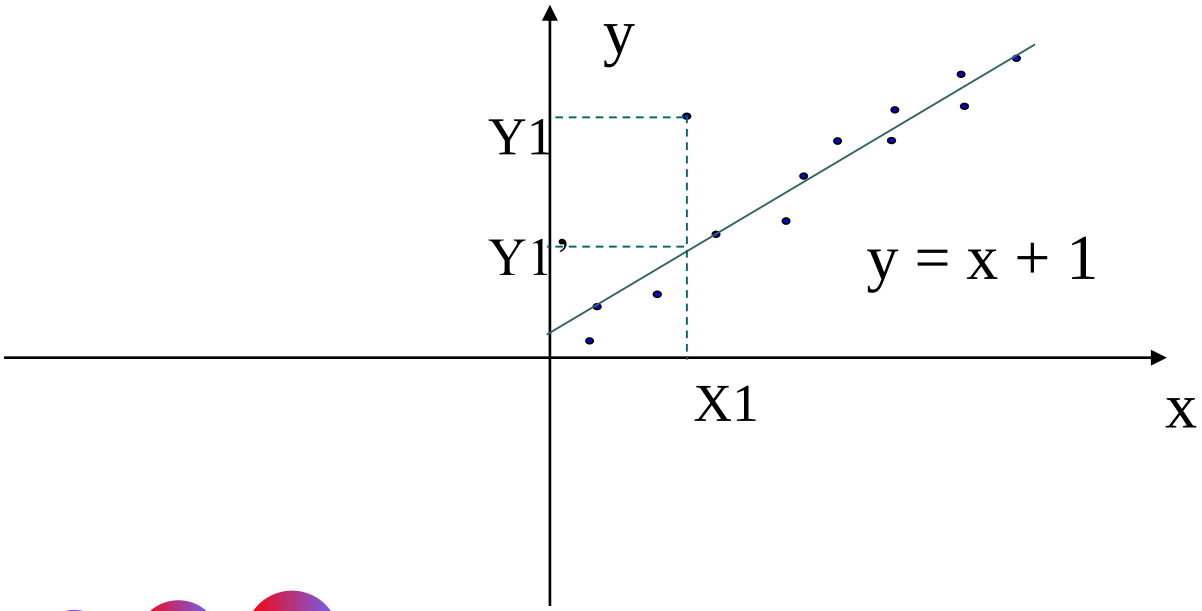
Bin 1: 9, 9, 9
Bin 2: 22, 22, 22
Bin 3: 29, 29, 29

Smoothing by bin boundaries:

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

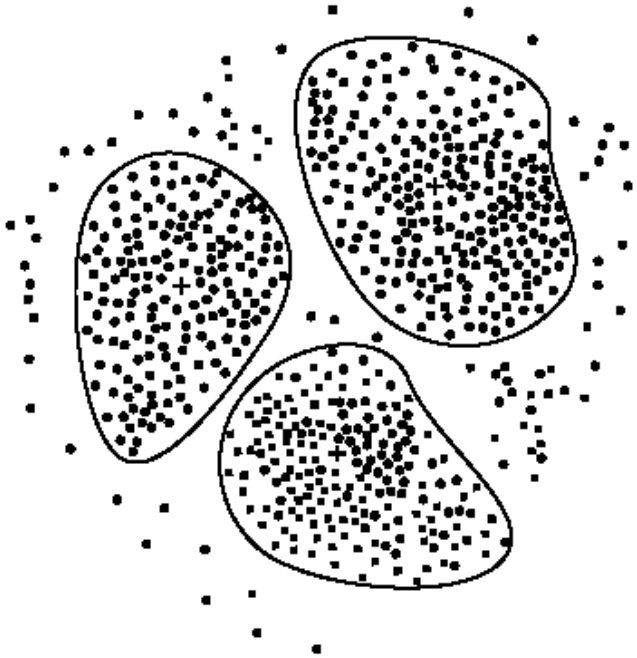
Làm sạch dữ liệu

- Nhận diện phần tử biên (outliers) và giảm thiểu nhiễu (noisy data)
  - Giải pháp giảm thiểu nhiễu
    - Hồi quy (regression)



Làm sạch dữ liệu

- Nhận diện phần tử biên (outliers) và giảm thiểu nhiễu (noisy data)
  - Giải pháp giảm thiểu nhiễu
    - Phân tích cụm (cluster analysis)



Làm sạch dữ liệu

- Xử lý dữ liệu không nhất quán
  - Định nghĩa của dữ liệu không nhất quán
    - Dữ liệu được ghi nhận khác nhau cho cùng một đối tượng/thực thể → discrepancies from inconsistent data representations
      - 2004/12/25 và 25/12/2004
    - Dữ liệu được ghi nhận không phản ánh đúng ngữ nghĩa cho các đối tượng/thực thể
      - Ràng buộc khóa ngoại
  - Nguyên nhân
    - Sự không nhất quán trong các qui ước đặt tên hay mã dữ liệu
    - Định dạng không nhất quán của các vùng nhập liệu
    - Thiết bị ghi nhận dữ liệu, …

Làm sạch dữ liệu

- Xử lý dữ liệu không nhất quán (inconsistent data)
  - Giải pháp
    - Tận dụng siêu dữ liệu, ràng buộc dữ liệu, sự kiểm tra của nhà phân tích dữ liệu cho việc nhận diện
    - Điều chỉnh dữ liệu không nhất quán bằng tay
    - Các giải pháp biến đổi/chuẩn hóa dữ liệu tự động

Tích hợp dữ liệu

- Tích hợp dữ liệu: quá trình trộn dữ liệu từ các nguồn khác nhau vào một kho dữ liệu sẵn sàng cho quá trình khai phá dữ liệu
  - Vấn đề nhận dạng thực thể (entity identification problem)
    - Tích hợp lược đồ (schema integration)
    - So trùng đối tượng (object matching)
  - Vấn đề dư thừa (redundancy)
  - Vấn đề mâu thuẫn giá trị dữ liệu (data value conflicts)
- → Liên quan đến cấu trúc và tính không thuần nhất (heterogeneity) về ngữ nghĩa (semantics) của dữ liệu
- → Hỗ trợ việc giảm và tránh dư thừa và không nhất quan về dữ liệu → cải thiện tính chính xác và tốc độ quá trình khai phá dữ liệu

Tích hợp dữ liệu

- Vấn đề nhận dạng thực thể
  - Các thực thể (object/entity/attribute) đến từ nhiều nguồn dữ liệu.
  - Hai hay nhiều thực thể khác nhau diễn tả cùng một thực thể thực.
  - Ví dụ ở mức lược đồ (schema): customer_id trong nguồn S1 và cust_number trong nguồn S2.
  - Ví dụ ở mức thể hiện (instance): “R & D” trong nguồn S1 và “Research & Development” trong nguồn S2. “Male” và “Female” trong nguồn S1 và “Nam” và “Nữ” trong nguồn S2.

  → Vai trò của siêu dữ liệu (metadata)

Tích hợp dữ liệu

- Vấn đề dư thừa
  - Hiện tượng: giá trị của một thuộc tính có thể được dẫn ra/tính từ một/nhiều thuộc tính khác, vấn đề trùng lắp dữ liệu (duplication).
  - Nguyên nhân: tổ chức dữ liệu kém, không nhất quán trong việc đặt tên chiều/thuộc tính.
  - Phát hiện dư thừa: phân tích tương quan (correlation analysis)
    - Dựa trên dữ liệu hiện có, kiểm tra khả năng dẫn ra một thuộc tính B từ thuộc tính A.
    - Đối với các thuộc tính số (numerical attributes), đánh giá tương quan giữa hai thuộc tính với các hệ số tương quan (correlation coefficient, aka Pearson's product moment coefficient).
    - Đối với các thuộc tính rời rạc (categorical/discrete attributes), đánh giá tương quan giữa hai thuộc tính với phép kiểm thử chi-square (χ^2).

Tích hợp dữ liệu

- Phân tích tương quan giữa hai thuộc tính số x và y

$$r = \frac{\overline{xy} - \overline{x}\,\overline{y}}{\sigma_x \sigma_y} = b_1 \frac{\sigma_x}{\sigma_y}$$

  - $r \in [-1, 1]$
  - r > 0: x và y tương quan thuận với nhau, trị số của x tăng khi trị số của y tăng, r càng lớn thì mức độ tương quan càng cao, x hoặc y có thể được loại bỏ vì dư thừa.
  - r = 0: x và y không tương quan với nhau (độc lập).
  - r < 0: x và y tương quan nghịch với nhau, x và y loại trừ lẫn nhau.

Phân tích hệ số tương quan

HSTQ	Ý nghĩa
±0.01 .. ±0.1	Mối tương quan quá thấp
±0.2 .. ±0.3	Mối tương quan thấp
±0.4 .. ±0.5	Mối tương quan trung bình
±0.6 .. ±0.7	Mối tương quan cao
±0.8 trở lên	Mối tương quan rất cao

Các tham số trong công thức

$$b_1 = \frac{\overline{xy} - \overline{x}\,\overline{y}}{\sigma_x^2}$$

$$b_0 = \overline{y} - b_1 \overline{x}$$

Trong đó: $\overline{x} = \frac{\sum x_i}{n}$ $\overline{y} = \frac{\sum y_i}{n}$

$$\overline{xy} = \frac{\sum xy}{n}$$

$$\sigma_x^2 = \frac{\sum x^2}{n} - \left(\frac{\sum x}{n}\right)^2 = \overline{x^2} - (\overline{x})^2$$

Ví dụ

Stt	Số năm sử dụng (năm) x	Giá bán (triệu đồng) y	xy	x^2	y^2
1	5	8,5	42,5	25,0	72,25
2	4	10,3	41,2	16,0	106,09
3	6	7,0	42,0	36,0	49,00
4	5	8,2	41,0	25,0	67,24
5	5	8,9	44,5	25,0	79,21
6	5	9,8	49,0	25,0	96,04
7	6	6,6	39,6	36,0	43,56
8	6	9,5	57,0	36,0	90,25
9	2	16,9	33,8	4,0	285,61
10	7	7,0	49,0	49,0	49,00
11	7	4,8	33,6	49,0	23,04
Tổng	58	97,5	473,2	326,0	961,29
Trung bình	5,273	8,864	43,018	29,636	87,390

Ví dụ

$$\sigma_x^2 = x^2 - (\bar{x})^2 = 29,636 - 5,273^2 = 1,831$$

$$b_1 = \frac{\overline{xy} - \bar{x}\bar{y}}{\sigma_x^2} = \frac{43,018 - 5,273 \times 8,864}{1,831} = -2,03 < 0$$

$$b_0 = \bar{y} - b_1 \bar{x} = 8,864 - (-2,03 \times 5,273) = 19,57$$

Ví dụ

$$\sigma_y = \sqrt{\sigma_y^2} = \sqrt{\overline{y^2} - \overline{y}^2} = \sqrt{87,390 - 8,864^2} = 2,97$$

$$\rightarrow r = (-2,03) \times \frac{1,353}{2,97} = -0,925$$

Tích hợp dữ liệu

- Vấn đề mâu thuẫn giá trị dữ liệu
  - Cho cùng một thực thể thật, các giá trị thuộc tính đến từ các nguồn dữ liệu khác nhau có thể khác nhau về cách biểu diễn (representation), đo lường (scaling), và mã hóa (encoding).
    - Representation: “2004/12/25” với “25/12/2004”.
    - Scaling: thuộc tính *weight* trong các hệ thống đo khác nhau với các đơn vị đo khác nhau, thuộc tính *price* trong các hệ thống tiền tệ khác nhau với các đơn vị tiền tệ khác nhau.
    - Encoding: “yes” và “no” với “1” và “0”.

Biến đổi dữ liệu

- Biến đổi dữ liệu: quá trình biến đổi hay kết hợp dữ liệu vào những dạng thích hợp cho quá trình khai phá dữ liệu
  - Làm trơn dữ liệu (smoothing)
  - Kết hợp dữ liệu (aggregation)
  - Tổng quát hoá (generalization)
  - Chuẩn hoá (normalization)
  - Xây dựng thuộc tính/đặc tính (attribute/feature construction)

Biến đổi dữ liệu

- Làm trơn dữ liệu (smoothing)
  - Các phương pháp binning (bin means, bin medians, bin boundaries)
  - Hồi quy
  - Các kỹ thuật gom cụm (phân tích phần tử biên)
  - Các phương pháp rời rạc hóa dữ liệu (các phân cấp ý niệm)

  → Loại bỏ/giảm thiểu nhiễu khỏi dữ liệu.

Biến đổi dữ liệu

- Kết hợp dữ liệu (aggregation)
  - Các tác vụ kết hợp/tóm tắt dữ liệu
  - Chuyển dữ liệu ở mức chi tiết này sang dữ liệu ở mức kém chi tiết hơn
  - Hỗ trợ việc phân tích dữ liệu ở nhiều độ mịn thời gian khác nhau

→ Thu giảm dữ liệu (data reduction)

Biến đổi dữ liệu

- Tổng quát hóa (generalization)
  - Chuyển đổi dữ liệu cấp thấp/nguyên tố/thô sang các khái niệm ở mức cao hơn thông qua các phân cấp ý niệm

  → Thu giảm dữ liệu (data reduction)

Biến đổi dữ liệu

- Chuẩn hóa (normalization)
  - min-max normalization
  - z-score normalization
  - → Các giá trị thuộc tính được chuyển đổi vào một miền trị nhất định được định nghĩa trước.

Biến đổi dữ liệu

- Chuẩn hóa (normalization)
  - min-max normalization
    - Giá trị cũ: v ∈ [minA, maxA]
    - Giá trị mới: v' ∈ [new_minA, new_maxA]

    → Ví dụ: chuẩn hóa điểm số từ 0-4.0 sang 0-10.0.

    → Đặc điểm của phép chuẩn hóa min-max?

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

Biến đổi dữ liệu

- Chuẩn hóa (normalization)
  - z-score normalization
    - Giá trị cũ: v tương ứng với mean Ā và độ lệch tiêu chuẩn (standard deviation) $б_A$
    - Giá trị mới: v’

    → Đặc điểm của chuẩn hóa z-score?

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

Biến đổi dữ liệu

- Xây dựng thuộc tính/đặc tính (attribute/feature construction)
  - Các thuộc tính mới được xây dựng và thêm vào từ tập các thuộc tính sẵn có.
  - Hỗ trợ kiểm tra tính chính xác và giúp hiểu cấu trúc của dữ liệu nhiều chiều.
  - Hỗ trợ phát hiện thông tin thiếu sót về các mối quan hệ giữa các thuộc tính dữ liệu.

  → Các thuộc tính dẫn xuất

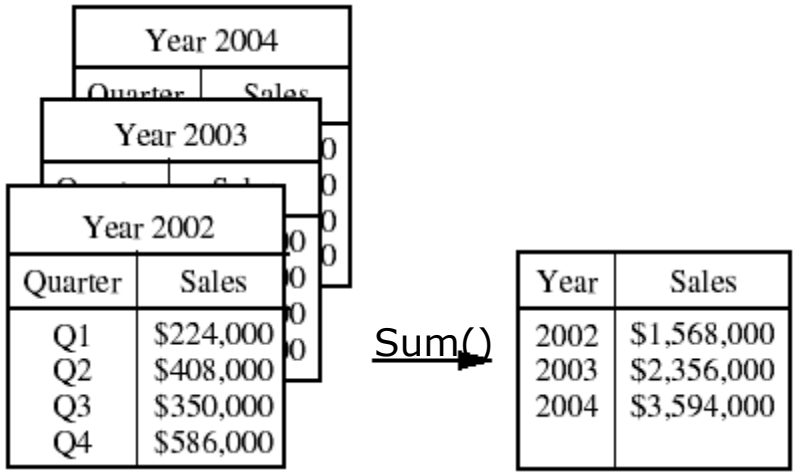
Thu giảm dữ liệu

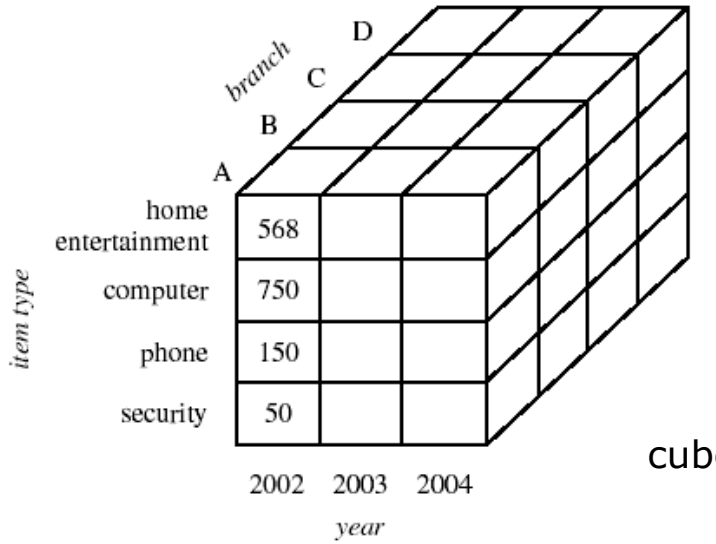
- Tập dữ liệu được biến đổi đảm bảo các toàn vẹn, nhưng nhỏ/ít hơn nhiều về số lượng so với ban đầu.
- Các chiến lược thu giảm
  - Kết hợp khối dữ liệu (data cube aggregation)
  - Chọn một số thuộc tính (attribute subset selection)
  - Thu giảm chiều (dimensionality reduction)
  - Thu giảm lượng (numerosity reduction)
  - Rời rạc hóa (discretization)
  - Tạo phân cấp ý niệm (concept hierarchy generation)

  → Thu giảm dữ liệu: lossless và lossy

Thu giảm dữ liệu

- Kết hợp khối dữ liệu (data cube aggregation)
  - Dạng dữ liệu: additive, semi-additive (numerical)
  - Kết hợp dữ liệu bằng các hàm nhóm: average, min, max, sum, count,..
  - → Dữ liệu ở các mức trừu tượng khác nhau.
  - → Mức trừu tượng càng cao giúp thu giảm lượng dữ liệu càng nhiều.





cube: Sale

Thu giảm dữ liệu

- Chọn một số thuộc tính (attribute subset selection)
  - Giảm kích thước tập dữ liệu bằng việc loại bỏ những thuộc tính/chiều/đặc trưng (attribute/dimension/feature) dư thừa/không thích hợp (redundant/irrelevant)
  - Mục tiêu: tập ít các thuộc tính nhất vẫn đảm bảo phân bố xác suất (probability distribution) của các lớp dữ liệu đạt được gần với phân bố xác suất ban đầu với tất cả các thuộc tính

Thu giảm dữ liệu

- Thu giảm chiều (dimensionality reduction)
  - Biến đổi wavelet (wavelet transforms)
  - Phân tích nhân tố chính (principal component analysis)

  → đặc điểm và ứng dụng?

Thu giảm dữ liệu

- Thu giảm lượng (numerosity reduction)
  - Các kỹ thuật giảm lượng dữ liệu bằng các dạng biểu diễn dữ liệu thay thế.
  - Các phương pháp có thông số (parametric): mô hình ước lượng dữ liệu → các thông số được lưu trữ thay cho dữ liệu thật
    - Hồi quy
  - Các phương pháp phi thông số (nonparametric): lưu trữ các biểu diễn thu giảm của dữ liệu
    - Histogram, Clustering, Sampling

Rời rạc hóa dữ liệu

- Giảm số lượng giá trị của một thuộc tính liên tục (continuous attribute) bằng các chia miền trị thuộc tính thành các khoảng (intervals)
- Các nhãn (labels) được gán cho các khoảng (intervals) này và được dùng thay giá trị thực của thuộc tính
- Các trị thuộc tính có thể được phân hoạch theo một phân cấp (hierarchical) hay ở nhiều mức phân giải khác nhau (multiresolution)

Tạo cây phân cấp ý niệm

- Hỗ trợ khai thác dữ liệu ở mức trừu tượng
- Rời rạc hóa dữ liệu hữu dụng cho việc tạo cây phân cấp ý niệm

Tổng kết

- Dữ liệu thực tế: không đầy đủ (incomplete/missing), nhiễu (noisy), không nhất quán (inconsistent)
- Quá trình tiền xử lý dữ liệu
  - làm sạch dữ liệu: xử lý dữ liệu bị thiếu, làm trơn dữ liệu nhiễu, nhận dạng các phần tử biên, hiệu chỉnh dữ liệu không nhất quán
  - tích hợp dữ liệu: vấn đề nhận dạng thực thể, vấn đề dư thừa, vấn đề mâu thuẫn giá trị dữ liệu
  - biến đổi dữ liệu: làm trơn dữ liệu, kết hợp dữ liệu, tổng quát hóa, chuẩn hóa, xây dựng thuộc tính/đặc tính
  - thu giảm dữ liệu: kết hợp khối dữ liệu, chọn một số thuộc tính, thu giảm chiều, rời rạc hóa và tạo phân cấp ý niệm