CHƯƠNG 2

CÁC KỸ THUẬT TIỀN XỬ LÝ DỮ LIỆU(tt)

3. Biến đổi dữ liệu

3.Biến đổi dữ liệu – Data transformation

- Chuẩn hoá và tổng hợp dữ liệu

Data transformation $-2, 32, 100, 59, 48 \longrightarrow -0.02, 0.32, 1.00, 0.59, 0.48$

3. 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 (Feature Construction)

3. Biến đổi dữ liệu

- **Làm trơn dữ liệu (smoothing):** là quá trình loại bỏ nhiễu để làm nổi bật xu hướng chính trong tập dữ liệu. Phương pháp này thường được sử dụng để chuẩn bị dữ liệu cho các bước phân tích hoặc khia thác tri thức tiếp theo.

3. Biến đổi dữ liệu

Phương *pháp trung bình (Mean Smoothing)*

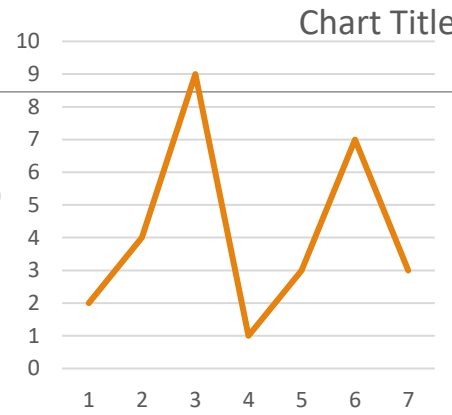
Sử dụng trung bình cộng của các giá trị lân cận thay thế giá trị nhiễu (k thường là số lẻ)

$$x_i' = \frac{x_{i-k} + x_{i-k+1} + \cdots + x_i + \cdots + x_{i+k}}{2k+1}$$

Ví dụ: Cho tập dữ liệu [2, 4, 9, 1 ,3 ,7, 3]

Hãy chuẩn hóa dữ liệu theo phương pháp trung bình với k = 3

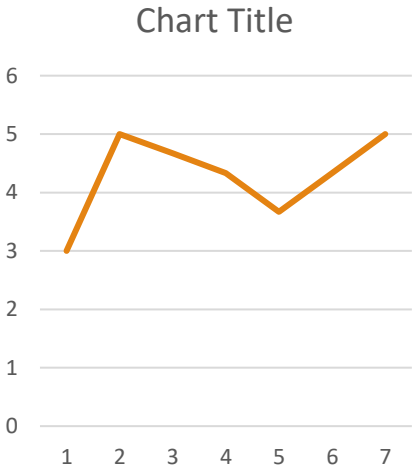
3. Biến đổi dữ liệu



Phương *pháp trung bình (Mean Smoothing)*

Ví dụ 1: Cho tập dữ liệu [2, 4, 9, 1 ,3 ,7, 3]

Sử dụng phương pháp trung bình để thay thế giá trị nhiễu với k = 3



Dãy mới [3, 5, 4.67, 4.33, 3.67, 4.33, 5]

3. Biến đổi dữ liệu

Ví dụ 2: Cho dữ liệu [5, 8, 12, 15, 18, 21, 25, 28, 30, 35]

Hãy chuẩn hóa dữ liệu theo phương pháp trung bình (với k =3).

3. Biến đổi dữ liệu

- Phương pháp chuẩn hóa
  - Một thuộc tính được chuẩn hóa bằng cách ánh xạ một cách có tỉ lệ dữ liệu về một khoảng xác định ví dụ như 0.0 đến 1.0.
  - Chúng ta sẽ xem xét ba phương pháp: min-max, z-score, và thay đổi số chữ số phần thập phân.

3. Biến đổi dữ liệu

- Chuẩn hóa min-max
  - Giả sử rằng $\min_A$ và $\max_A$ là giá trị tối thiểu và tối đa của thuộc tính A. Chuẩn hóa min-max sẽ ánh xạ giá trị v của thuộc tính A thành v' trong khoảng [new_min_A, new_max_A]
  - $v' = \frac{v - \min_A}{\max_A - \min_A}(new\_max_A - new\_min_A) + new\_min_A$

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

3. Biến đổi dữ liệu

Ví dụ: Giả sử giá trị nhỏ nhất và lớn nhất cho thuộc tính "thu nhập bình quân" là \$12.000 và \$98.000. Chúng ta muốn ánh xạ giá trị \$73.600 về khoảng [0.0, 1.0].

- $v' = \frac{73.600 - 12.000}{98.000 - 12.000}(1.0 - 0) + 0 = 0.716$

3. Biến đổi dữ liệu

- Chuẩn hóa z-score
  - Giá trị của một thuộc tính A được chuẩn hóa dựa vào độ lệch tiêu chuẩn và trung bình của A.
  - Một giá trị v của thuộc tính A được ánh xạ thành v’ như sau:

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

  - Trong đó: $\bar{A}$ là giá trị trung bình của A; σ_A là độ lêch chuẩn

3. Biến đổi dữ liệu

Ví dụ: Giả sử rằng giá trị trung bình và độ lệch chuẩn của thuộc tính "thu nhập" là 54000 và 16000. Với cách chuẩn hóa z-score, giá trị 73000 sẽ được chuyển thành $v' =$

$$\frac{v-\bar{A}}{\sigma_A} = \frac{73000-54000}{16000} = 1.1875$$

Bài tập: X = {−5.0, 1.11, 7.23, 17.6, 23.0} chuẩn hóa 7.23 theo min-max, z – score

3. Biến đổi dữ liệu

- Chuẩn hóa thay đổi số chữ số phần thập phân
  - Số chữ số phần thập phân được di chuyển phụ thuộc vào giá trị tuyệt đối lớn nhất có thể có của thuộc tính A.
  - Khi đó giá trị v sẽ được ánh xạ thành v’ bằng cách tính $\mathbf{v'} = \frac{\mathbf{v}}{\mathbf{10^j}}$
  - Trong đó j là giá trị nguyên nhỏ nhất để thỏa mãn max(|v’|) < 1

Ví dụ: Giả sử rằng các giá trị của thuộc tính A được ghi nhận nằm trong khoảng -986 đến 917. Hãy chuẩn hóa thay đổi số chữ số phần thập phân

3. Biến đổi dữ liệu

- Chuẩn hóa thay đổi số chữ số phần thập phân
  - Ví dụ: Giả sử rằng các giá trị của thuộc tính A được ghi nhận nằm trong khoảng -986 đến 917 → trị tuyệt đối lớn nhất là 986. Sau đó lấy các giá trị chia cho 1.000 (j = 3). Như vậy giá trị -986 sẽ chuyển thành -0.986 và 917 được chuyển thành 0.917.

3. Biến đổi dữ liệu

- ***Tạo thuộc tính mới từ dữ liệu hiện có*:** là quá trình biến đổi hoặc kết hợp các thuộc tính hiện có trong tập dữ liệu để tạo ra các thuộc tính mới có ý nghĩa hơn, hỗ trợ phân tích hoặc cải thiện hiệu quả của các mô hình học máy.

3. Biến đổi dữ liệu

Các phương pháp tạo thuộc tính mới:

+ Kết hợp các thuộc tính hiện có

+ Biến đổi hàm toán học

+ Tạo thuộc tính từ ngày tháng

+ Tạo thuộc tính nhị phân (Binning/Binary Feature)

3. Biến đổi dữ liệu

Ví dụ: Cho dữ liệu.

ID	Age	Salary/Week (triệu đồng)	Working Hours/Week	Work From
1	25	20	40	30/04/2023
2	30	30	45	02/06/2018
3	35	50	38	06/03/2015
4	40	40	50	07/05/2011
5	28	35	42	01/11/2020

-Tạo thuộc tính “Salary per hour” (Lương theo giờ)

-Tạo thuộc tính “Log Salary” (Lấy log cơ số 10 của Lương)

-Tạo thuộc tính “Year Experience” (Năm kinh nghiệm)

-Tạo thuộc tính “Incom Level” (=Low nếu Lương < 30, =Medium nếu Lương từ 30 đến 39, High nếu cao)

3. Biến đổi dữ liệu

Ví dụ: Cho tập dữ liệu

ID	smoke_fre	num_ciga
1	0	0
2	1	12
3	2	8
4	3	5
5	4	4
6	5	3
7	6	2
8	7	1
9	3	6
10	2	10

Trong đó, smoke_fre: tần số hút thuốc

- 0: không hút
- 1: hút hàng ngày
- 2: 1 tuần 3 lần
- 3: 1 tuần 2 lần
- 4: 1 tuần 1 lần
- 5: 1 tháng 2 lần
- 6: 1 tháng 1 lần
- 7: có hút nhưng ít hơn 1 tháng 1 lần

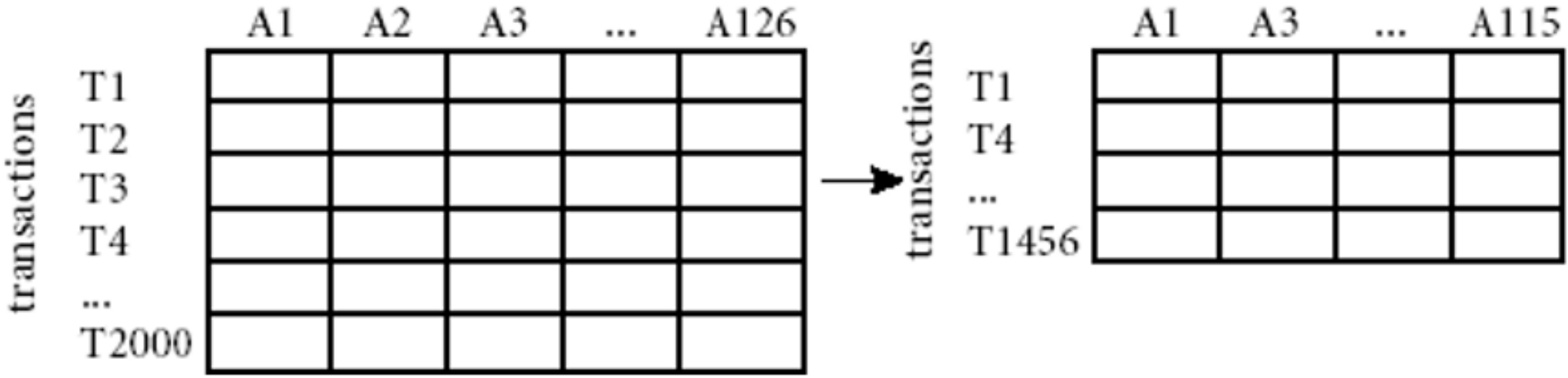
Num_ciga: số điếu thuốc hút 1 lần

Hãy tạo thuộc tính num_ciga_month: Thể hiện số điếu thuốc hút trong 1 tháng

4. Thu giảm dữ liệu

4.Thu giảm dữ liệu – Data reduction

- Giảm kích thước dữ liệu nhưng đảm bảo kết quả phân tích.
- Bằng cách kết hợp dữ liệu, loại bỏ các đặc điểm dư thừa, gom cụm dữ liệu.



4. 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.

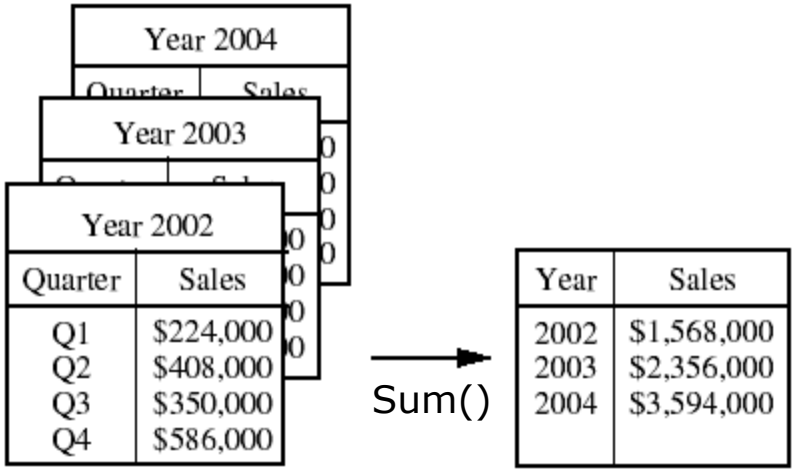
Các chiến lược thu giảm:

- Tổng hợp
- Giảm chiều dữ liệu
- Nén dữ liệu
- Giảm số lượng

4. Thu giảm dữ liệu

Tổng hợp

- Tổng hợp từ 2 thuộc tính dữ liệu trở lên thành một thuộc tính.



4. Thu giảm dữ liệu

Lựa chọn tập thuộc tính con

- Chỉ chọn những thuộc tính phù hợp cho bài toán phân tích cụ thể.
- Ví dụ nếu nhiệm vụ phân tích chỉ liên quan đến việc phân loại khách hàng xem họ có hoặc không muốn mua một đĩa nhạc mới hay không → thuộc tính điện thoại không quan trọng bằng thuộc tính tuổi tác

4. Thu giảm dữ liệu

Lựa chọn tập thuộc tính con

- Thông qua các phép kiểm thống kê để xác định thuộc tính nào là tốt (xấu).
- Kỹ thuật lựa chọn tăng dần: Xuất phát từ tập rỗng các thuộc tính, các thuộc tính tốt nhất mỗi khi xác định được thêm vào, lập lại khi không thêm được thuộc tính nào nữa.

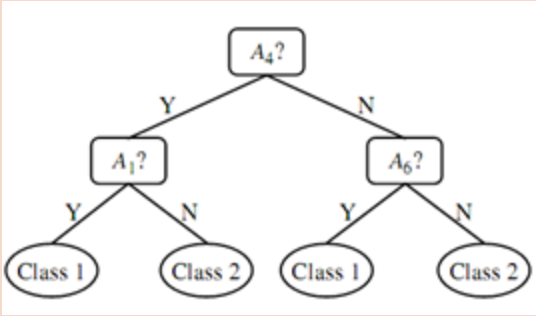
4. Thu giảm dữ liệu

Lựa chọn tập thuộc tính con

- Kỹ thuật loại bớt: Xuất phát từ tập đầy đủ các thuộc tính, ở mỗi bước loại ra thuộc tính tồi nhất.
- Kết hợp giữa phương pháp loại bớt và lựa chọn tăng dần bằng cách tại mỗi bước ngoài việc lựa chọn thêm các thuộc tính tốt nhất đưa vào tập thì cũng đồng thời loại bỏ đi các thuộc tính tồi nhất khỏi tập đang xét.
- Cây quyết định: Cây được xây dựng từ nguồn dữ liệu ban đầu. Tất cả thuộc tính không xuất hiện trên cây được coi là không hữu ích.

4. Thu giảm dữ liệu

Lựa chọn tập thuộc tính con

Lựa chọn tăng dần	Loại bớt	Cây quyết định
Tập thuộc tính ban đầu **{A1, A2, A3, A4, A5, A6}** **Tập rút gọn ban đầu** **{}** **=> {A1}** **=> {A1, A4}** **=> Kết quả {A1, A4, A6}**	Tậpthuộc tính ban đầu {A1, A2, A3, A4, A5, A6} => {A1, A3, A4, A5, A6} => {A1, A4, A5, A6} => Kết quả {A1, A4, A6}	Tậpthuộc tính ban đầu {A1, A2, A3, A4, A5, A6} => Kết quả {A1, A4, A6} 

Bảng 3.2. Ví dụ kỹ thuật rút gọn

5. Rời rạc hóa dữ liệu

Rời rạc hóa là quá trình chuyển đổi dữ liệu liên tục thành các khoảng hoặc các giá trị rời rạc

Các phương pháp rời rạc hóa

-Phân đoạn chiều rộng bằng nhau (equal Width Binning)

-Phân đoạn tần suất bằng nhau (Phân đoạn tần suất bằng nhau)

-Rời rạc hóa theo cụm (Rời rạc hóa theo cụm)

5. Rời rạc hóa dữ liệu

Ví dụ 1: Cho dữ liệu [3, 7, 5, 2, 9,1, 8]

Phân đoạn chiều rộng bằng nhau, bin = 3

R = max – min =9 – 1 = 8, W = 8/3 = 2.67

N1: [1, 3.67)

N2: [3.67, 6.34)

N3: [6.34, 9]

Dãy mới: [1, 3, 2, 1, 3, 1, 3]

5. Rời rạc hóa dữ liệu

Ví dụ 2: Cho dữ liệu [3, 7, 5, 2, 9, 1, 8]

Phân đoạn tần suất bằng nhau, bin = 3

7/3 = 2.3 gần bằng 2 => Mỗi khoảng là 2

1, 2 thuộc nhóm 1

3, 5 thuộc nhóm 2

7, 8, 9 thuộc nhóm 3

Dãy mới [2, 3, 2, 1, 3, 1, 3] (chia đều thành viên)

5. Rời rạc hóa dữ liệu

Ví dụ 3: Cho dữ liệu [5, 8, 12, 15, 18, 21, 25, 28, 30, 35]. Hãy rời rạc hóa bằng phương pháp Phân đoạn chiều rộng bằng nhau. Chia thành 4 bin

Bài tập chương 2

1. Cho mảng một chiều X = {−5.0, 1.11, 7.23, 17.6, 23.0}, hãy chuẩn hóa mảng sử dụng

   a/Phương pháp chuẩn hóa Min-Max: trong khoảng [−1, 1].

   b/ Phương pháp chuẩn hóa Min-Max : trong khoảng [0, 1].

   c/ Phương pháp chuẩn hóa Min-Max : trong khoảng [−1, 1].

   d/ Phương pháp chuẩn hóa z-core

   e/ Phương pháp chuẩn hóa thay đổi số chữ số phần thập phân

Bài tập chương 2

2. Làm mịn dữ liệu sử dụng kỹ thuật làm tròn cho tập sau:

Y = {1.17, 1.73, 2.59, 2.53, 2.67, 3.28, 3.38, 3.44, 4.23} hãy chuẩn hóa mảng sử dụng

a/Phương pháp chuẩn hóa Min-Max: trong khoảng [−1, 1].

b/ Phương pháp chuẩn hóa Min-Max : trong khoảng [0, 1].

c/ Phương pháp chuẩn hóa Min-Max : trong khoảng [−1, 1].

d/ Phương pháp chuẩn hóa z-core

e/ Phương pháp chuẩn hóa thay đổi số chữ số phần thập phân

f/ Phương pháp chia giỏ theo độ rộng với N = 4

g/ Phương pháp chia giỏ theo độ sâu với N = 3,

Hãy khử nhiễu theo giá trị biên, theo giá trị trung bình, theo giá trị trung vị.

Bài tập chương 2

3. Một công ty ghi nhận thông tin về khách hàng và giao dịch qua bảng dữ liệu sau:

ID	Age	Gender	Income ($)	Transactions	Spend ($)	Membership	JoinYear
C001	25	Male	50000	15	3000	Gold	2019
C002	34	Female	60000	20	4000	Gold	2018
C003	29	Male	40000	12	NaN	Silver	2020
C004	NaN	Female	55000	18	3700	Silver	2019
C005	45	Female	NaN	25	4500	Gold	2015
C006	32	Male	200000	50	20000	Platinum	2017
C007	-1	Male	42000	NaN	3800	Silver	2019
C008	30	Male	52000	14	3200	Silver	2019
C009	27	Other	60000	22	5000	Gold	2018
C010	35	Female	55000	18	-3000	Silver	2019

-Thực hiện xử lý các giá trị thiếu. Phát hiện và xử lý các giá trị mâu thuẫn.

-Làm trơn dữ liệu bằng phương pháp trung bình (k=3) với cột Spend và Transactions.

-Chuẩn hóa dữ liệu cho Spend và Income