TRƯỜNG ĐẠI HỌC VINH

DỮ LIỆU VÀ TIỀN XỬ LÝ DỮ LIỆU

Phan Anh Phong, PhD.
Vinh University

1

Nội dung

- Dữ liệu
  - Khái niệm
  - Thuộc tính, giá trị, kiểu thuộc tính
  - Tập dữ liệu
- Chất lượng dữ liệu
  - Dữ liệu bị nhiễu, ngoại lai
  - Dữ liệu bị thiếu
- Tiền xử lý dữ liệu
  - Một số độ đo thống kê
  - Làm sạch dữ liệu
  - Biến đổi dữ liệu
  - Thu gọn dữ liệu

2

Dữ liệu

- Tập hợp các đối tượng dữ liệu và các thuộc tính của chúng
- Một thuộc tính là 1 tính chất/ đặc điểm của đối tượng
  - Ví dụ: màu mắt của một người, nhiệt độ…
  - Thuộc tính cũng được hiệu như 1 biến, một đặc trưng
- Một tập các thuộc tính mô tả một đối tượng / bản ghi
  - Đối tượng cũng được hiểu như là bản ghi, điểm, thực thể, thể hiện, trwongf hợp,

Các thuộc tính

Các đối tượng

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Giá trị của thuộc tính

- Giá trị của thuộc tính là các số hoặc các ký hiệu được gán với thuộc tính đó
- Phân biệt giữa thuộc tính với giá trị thuộc tính
  - Cùng thuộc tính nhưng có thể được ánh xạ thành các giá trị thuộc tính khác nhau
    - Ví dụ: chiều cao có thể được đo bằng mét hoặc bằng feet

Các loại thuộc tính

- Thuộc tính phân loại - nominal
  - Giá trị dữ liệu của thuộc tính nominal chỉ có thể liệt kê và có tính phân biệt, chứ không có tính thứ tự
  - Phép toán trên thuộc tính này là: = và ≠
  - Vi dụ: màu mắt, mã vùng số điện thoại…
- Thuộc tính phân loại có thứ tự - ordinal
  - Tương tự thuộc tính phân loại nhưng giá trị dữ liệu có tính thứ tự, tức là có thể sắp xếp theo mức độ
  - Phép toán trên thuộc tính ordinal: =, ≠, <, >, ≥, ≤
  - Ví dụ: độ ngon của Bimbim {được phân bậc từ 1 đến 5}, xếp loại học tập {A, B+, B…}, Chiều cao {cao, vừa, thấp}

5

5

Các loại thuộc tính

- Thuộc tính số
  - Giá trị dữ liệu là các số (nguyên, thực) - kết quả của sự đo, đếm theo số lượng, khối lượng…
  - Phép toán: +, -, x, /, =, ≠, <, >, ≤, ≥.
  - Ví dụ: độ dài, thời gian,

6

6

Thuộc tính rời rạc, liên tục

- Thuộc tính rời rạc
  - Chỉ có hữu hạn hoặc vô hạn đếm được các giá trị
  - Ví dụ: mã vùng số điện thoại, số lượng, tập hợp các từ có trong một tài liệu
  - Thường được biểu diễn như 1 biến nguyên
  - Chú ý: thuộc tính nhị phân (có 2 giá trị) là một trường hợp đặc biệt của thuộc tính rời rạc
- Thuộc tính liên tục
  - Có số thực giá trị các thuộc tính
  - Ví dụ: nhiệt độ, chiều cao, trọng lượng...
  - Trong máy tính tập số thực cũng hữu hạn đếm được!
  - Thuộc tính liên tục được biểu diễn như các biến thực dấu phẩy động

7

7

Tập dữ liệu – các dạng

- Dữ liệu dạng bản ghi - Record
  - Dữ liệu giao dịch
  - Ma trận dữ liệu
  - Ma trận khoảng cách
  - Dữ liệu văn bản
- Tập dữ liệu dạng đồ thị - Graph
  - World Wide Web
  - Cấu trúc phân tử
- Tập dữ liệu có thứ tự - Ordered
  - Dữ liệu không gian - Spatial Data
  - Dữ liệu thời gian - Temporal Data
  - Dữ liệu chuỗi - Sequential Data
  - ...

8

8

Dữ liệu dạng bản ghi - Record

- Tập dữ liệu là tập hợp các bản ghi có số thuộc tính cố định.
- Mỗi bản ghi mô tả một đối tượng

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

9

9

Dữ liệu giao dịch - Transaction Data

- Là một dạng dữ liệu đặc biệt của dữ liệu bản ghi, trong đó:
  - Mỗi bản ghi (transaction) liên quan đến 1 tập hợp các các mặt hàng (phần tử).
  - Ví dụ: Tập hợp các sản phẩm được mua bởi khách hàng trong một cửa hàng

TID	*Items*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

10

10

Ma trận dữ liệu

- Nếu các đối tượng dữ liệu có cố định 1 số thuộc tính kiểu số, khi đó các đối tượng dữ liệu có thể xem là 1 điểm trong không gian nhiều chiều, trong đó mỗi chiều biểu diễn 1 thuộc tính cụ thể
- Những tập dữ liệu như thế có thể biểu diễn bằng 1 ma trận m dòng n cột, mỗi dòng là 1 đối tượng, mỗi cột là một thuộc tính

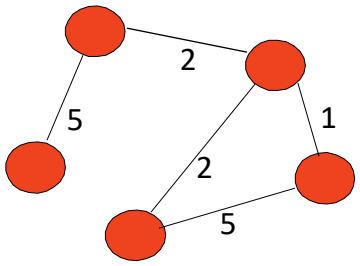
Ví dụ về ma trận dữ liệu các văn bản

- Mỗi tài liệu được biểu diễn như 1 véc tơ các hạng từ (term – thuật ngữ),
  - Mỗi term là một thuộc tính (thành phần) của véc tơ,
  - Giá trị của mỗi thành phần là 1 số tương ứng với số lần xuất hiện của term đó trong tài liệu

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

Dữ liệu dạng đồ thị - Graph Data

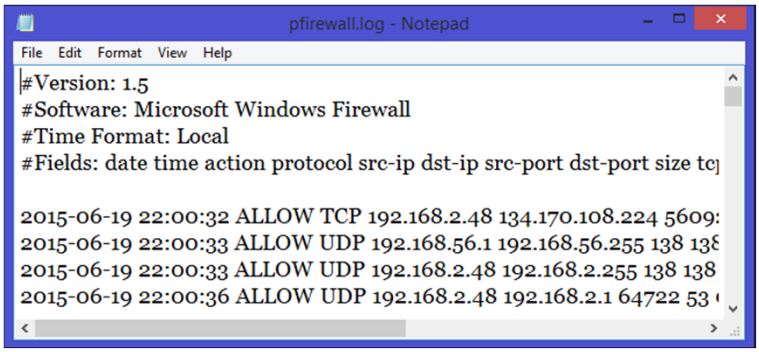
- Dữ liệu được biểu diễn bằng đồ thị
- Ví dụ: Mạng xã hội, liên kết trang web… có thể tạo ra bằng cách sử dụng "khoảng cách" giữa các nút – ma trận khoảng cách các nút



13

Dữ liệu có trình tự - Sequence Data

- Dữ liệu có trình tự theo thời gian
- Ví dụ: log entries, spatio-temporal data



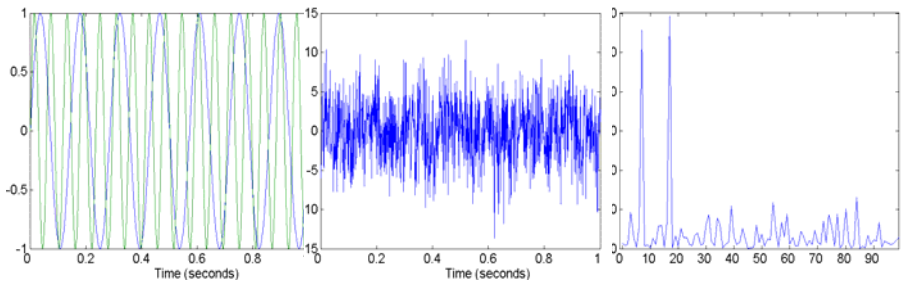
14

Chất lượng dữ liệu

- Dữ liệu thu được phần lớn là chứa LỖI – chất lượng chưa tốt
- Ví dụ về chất lượng dữ liệu chưa tốt:
  - Nhiễu và ngoại lai (bất thường)
  - Thiếu giá trị (missing values)
  - Dữ liệu trùng lặp (duplicate data)
  - …

Dữ liệu bị nhiễu

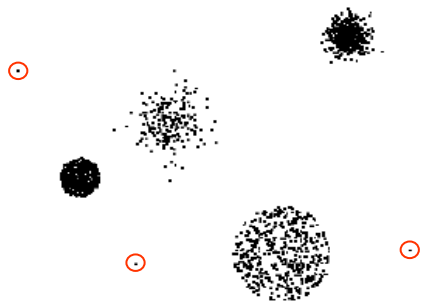
- Nhiễu dữ liệu – thay đổi giá trị gốc của dữ liệu
- Ví dụ: tiếng nói của 1 người khi nói qua điện thoại có sóng mang kém



Two Sine Waves | Two Sine Waves + Noise | Frequency Plot (FFT)

Dữ liệu ngoại lai – Outliers

- Dữ liệu ngoại lai là những đối tượng dữ liệu có các đặc trưng khác biệt hẳn so với hầu hết các đối tượng dữ liệu còn lại trong tập dữ liệu



17

17

Khuyết giá trị - Missing Values

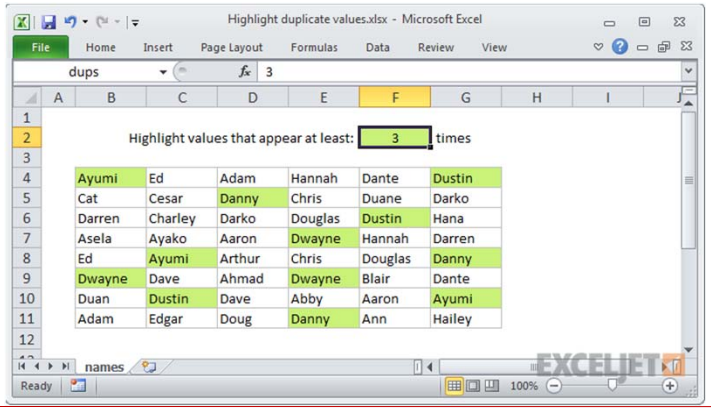
- Nguyên nhân dữ liệu bị khuyết giá trị
  - Thông tin không thu thập được
    (ví dụ người bệnh không cung cấp được thông tin về tuổi và khối lượng)
  - Một số thuộc tính không phù hợp cho tất cả đối tượng
    (ví dụ: thu nhập hàng tháng của một đứa trẻ)
- Phối hợp với tập dữ liệu bị khuyết giá trị
  - Loại bỏ dữ liệu đó ra khỏi tập dữ liệu
  - Ước lượng giá trị thiếu dựa vào những giá trị đã có
    (trung bình cộng, max, min…)
  - Thay thế chúng theo xác suất
  - …

18

18

Dữ liệu lặp lại - Duplicate Data

- Tập dữ liệu có các đối tượng dữ liệu giống nhau
  - Nguyên nhân có thể do trộn dữ liệu từ nhiều nguồn khác nhau



Tiền xử lý dữ liệu

- Là công đoạn làm cho dữ liệu phù hợp hơn với việc khai phá
  - Một số độ đo thống kê
  - Làm sạch dữ liệu
  - Thu gọn dữ liệu (giảm chiều, giảm kích thước)
  - Chuyển đổi dữ liệu
  - Rời rạc hóa dữ liệu

Một số độ đo thống kê dữ liệu

- Mean (Giá trị trung bình): được tính đơn giản bằng tổng của tất cả các giá trị của dữ liệu trong mẫu chia cho kích thước mẫu.
- Median (Trung vị): Sắp xếp các giá trị dữ liệu của mẫu tăng dần. Nếu mẫu có số phần tử lẻ, thì median là giá trị chính giữa của dãy đã sắp xếp, ngược lại thì median là trung bình cộng của 2 mẫu chính giữa.
- Mode (Số yếu vị): giá trị của phần tử có số lần xuất hiện nhiều nhất trong tập mẫu
- Phương sai, độ lệch chuẩn
- Min, max, range

Một số độ đo thống kê dữ liệu

- Mean, Median và Mode:
  - Ví dụ: Cho X = {10, 9, 8, 11, 8, 12, 8 }; Y = {21, 21, 24, 25}, tìm Mean, median, mode của X, Y
  - Trong 3 tham số Mean, Median và Mode thì Median có khả năng đo lường xu hướng tập trung của dữ liệu mạnh nhất
  - Mode rất hữu ích đối với dữ liệu có kiểu dữ liệu phân loại (nominal). Ví dụ nếu dữ liệu mô tả giới tính là nominal và 1 là nam, 0 là nữ thì Mean hay Median là 0.5 không có ý nghĩa gì. Trong khi đó Mode cho biết tần suất nam hay nữ xuất hiện nhiều nhất.
  - Trung bình bị ảnh hưởng bởi các giá trị ngoại lệ (outliers) trong dữ liệu
  - Trung vị được sử dụng để đo giá trị trung tâm của dữ liệu mà ít bị ảnh hưởng bởi các giá trị ngoại lệ

Một số độ đo thống kê dữ liệu

- Phương sai tổng thể, phương sai, độ lệch chuẩn:
  - Phương sai tổng thể: $s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$

  Phương sai tổng thể - bình phương của sai lệch (độ lệch)

  - Phương sai mẫu (phương sai mẫu hiệu chỉnh} : $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$
  - Độ lêch chuẩn: $s = \sqrt{s^2}$
  - Độ lệch chuẩn hiệu chỉnh

  N kích thước mẫu, $\bar{x}$ là giá trị trung bình (mean)

Một số độ đo thống kê dữ liệu

- Phương sai tổng thể, phương sai, độ lệch chuẩn:

```
from statistics import variance
import numpy as np
#phương sai tổng thể
print(np.var([1,2,3,4]))
#Phuong sai hiệu chỉnh
print(variance([1,2,3,4]))
//1.25
//1.6666666666666667
```

Dùng pstdev() và stdev() để tính độ lệch chuẩn

Làm sạch dữ liệu

- Làm sạch dữ liệu liên quan đến việc sửa chữa các vấn đề về chất lượng dữ liệu, bao gồm:
  - Xử lý dữ liệu bị khuyết
  - Xử lý dữ liệu nhiễu
  - Xử lý dữ liệu ngoại lai
  - Chỉnh sửa dữ liệu không nhất quán
  - Loại dữ liệu trùng lặp

Xử lý dữ liệu bị khuyết

- Xác định các dữ liệu bị khuyết
- Xác định nguyên nhân xảy ra sự thiếu dữ liệu
- Xử lý theo 2 cách sau:
  - Loại những dòng thiếu dữ liệu ra khỏi tập dữ liệu
  - Thay thế giá trị dữ liệu thiếu bằng giá trị khác
    - Giá trị trung bình của các giá trị trên thuộc tính có giá trị thiếu
    - Giá trị trung bình của các giá trị trên thuộc tính có giá trị thiếu và có cùng lớp
    - Sử dụng các giá trị thống kê khác như trung vị, giá trị xuất hiện nhiều nhất (mode)…
    - …

Xử lý dữ liệu bị khuyết

- Ví dụ: Thay thế giá trị bị khuyết ở tập dữ liệu bằng 10impor00 trong Python

```
t pandas as pd
import numpy as np
df = pd.DataFrame({'a':[1,2,3,4,np.nan],
                   'b':[6,7,8,np.nan,np.nan],
                   'x':[11,12,13,np.nan,np.nan],
                   'y':[16,np.nan,np.nan,19,np.nan]})
print(df)
df.fillna(1000, inplace=True)
print(df)
```


Xử lý dữ liệu bị khuyết

- Thay thế giá trị bị khuyết ở thuộc tính x bởi 100, thuộc tính y bởi 0

```
df.fillna({'x':100, 'y':0}, inplace=True)
```

- Thay thế giá trị bị khuyết ở thuộc tính x bởi giá trị trung vị của thuộc tính này

```
median = df['x'].median()
df['x'].fillna(median, inplace=True)
```

- Xóa các dòng bị khuyết

```
df.dropna(inplace=True)
```

Xử lý dữ liệu nhiễu

- Phương pháp binning
  - Trước tiên sắp dữ liệu theo thứ tự và phân hoạch thành các thùng (bin) có kích thước bằng nhau
  - Sau đó, giảm nhiễu bằng cách thay các giá trị trong mỗi thùng bằng các giá trị trung bình hoặc trung vị hoặc giá trị cận trên, cận dưới của mỗi thùng
- Phân cụm
- Hồi quy

Xử lý dữ liệu nhiễu

- Ví dụ về phương pháp binning
  - Cho tập dữ liệu bị nhiễu: 9, 21, 24, 21, 4, 26, 28, 34, 29, 8, 15, 25
  - Sắp xếp tăng dần và phân hoạch thành 3 thùng:
    4, 8, 9,15; 21, 21, 24, 25; 26, 28, 29, 34
  - Giảm nhiễu bằng giá trị trung bình trong mỗi thùng
    9, 9, 9, 9; 22, 22, 22, 22; 29, 29, 29, 29
  - Giảm nhiễu bằng giá trị biên trong mỗi thùng
    4, 4, 4, 15; 21, 21, 25, 25; 26, 26, 26, 34

Xử lý dữ liệu ngoại lai

- Thế nào là dữ liệu ngoại lai:
  - Là các giá trị cực so với các giá trị khác được quan sát trong cùng một điều kiện. Outlier có thể là một giá trị đơn lẻ, nhưng cũng có thể là giá trị từ hai hay nhiều biến số
- Phương pháp xử lý dữ liệu ngoại lai:
  - Dựa vào giả định của phân phối chuẩn (normal distribution)
  - Dựa vào giá trị trung vị
  - Phương pháp tứ phân vị - *IQR -interquartile range*
  - …

Xử lý dữ liệu ngoại lai

- Phương pháp dựa vào giả định của phân phối chuẩn
  - Nếu thuộc tính X tuân theo luật phân phối chuẩn với trung bình *m* và độ lệch chuẩn *s* thì 99% các giá trị của X phải nằm trong đoạn [*m - 3×s, m + 3×s*] .
  - Do đó, bất cứ số *xi* nào có giá trị thấp hơn (m - 3×s) hay cao hơn (m + 3×s) thì có thể nghi ngờ là outlier.

Xử lý dữ liệu ngoại lai

- Phương pháp dựa vào giá trị trung vị
  - Tính trung vị của các giá trị, giả sử đó là *M*
  - Tính độ khác biệt tuyệt đối giữa từng số trong thuộc tính X và *M*, và gọi kết quả là d_i :

$$d_i = |x_i - M|$$

  - Tính trung vị của các d_i và gọi là *Md*
  - Lấy d_i chia cho *Md*, và gọi chỉ số này là t_i
  - Nếu t_i cao hơn 4.5, có thể xem x_i tương ứng là outlier

33

33

Xử lý dữ liệu ngoại lai

- Phương pháp dựa vào IQR – Tứ phân vị
  - Sắp xếp dữ liệu theo thứ tự tăng
  - Tính giá trị *Q1* – giá trị bách phân 25:
  - Tính giá trị *Q3* – giá trị bách phân 75:
  - Tính Interquartile Range – *IQR = Q3 - Q1*
  - Outliers: giá trị nằm dưới *Q1 - IQR×1.5* và nằm trên Q3 + IQR × 1,5 được xem là bất thường

34

34

Xử lý dữ liệu ngoại lai

- Tính Q1, Q2, Q3 (kiểm tra lại)
  - Sắp xếp dữ liệu theo thứ tự tăng
  - Nếu kích thước mẫu là chẵn (dạng 2n) thì *Q1* là *median* của n số nhỏ nhất và Q3 là median của n số lớn nhất
  - Nếu kích thước mẫu là lẻ (dạng 2n + 1) thì Q1 là *median* của n số nhỏ nhất và *Q3* là median của *n* số lớn nhất
  - Q2 – median của mẫu dữ liệu

Xử lý dữ liệu ngoại lai

- Tính IQR trong Python sử dụng numpy và pandas

```
dataset= dataset= [10,12,12,13,12,11,14,13,15,10,10, 12, 17]
dataset = sorted(dataset)
print('Tap du lieu sau khi sap: ', dataset)
q1, q3= np.percentile(dataset,[25,75])
print('q1: ', q1);print('q3: ', q3)
iqr = q3 - q1; print('iqr: ', iqr)
can_duoi= q1 -(1.5 * iqr); can_tren= q3 +(1.5 * iqr)
print('q1 -(1.5 * iqr)  : ', can_duoi); print(' q3 +(1.5 * iqr) : ', can_tren)
outliers=[]
for x in dataset:
    if (x < can_duoi or x > can_tren):
        outliers.append(x)
print('Phan tu ngoai lai:', outliers)
```

Xử lý dữ liệu không nhất quán

- Dữ liệu không nhất quán
  - Dữ liệu được ghi nhận khác nhau cho cùng một đối tượng/thực thể
  - Ví dụ
    - 2014/12/24 và 24/12/2014
  - Dữ liệu được ghi nhận không phản ánh đúng ngữ nghĩa cho các đối tượng/thực thể
  - Nguyên nhân
    - Không nhất quán trong các qui ước đặt tên hay mã hóa dữ liệu
    - Định dạng không nhất quán của các vùng nhập liệu
    - …

Biến đổi dữ liệu

- Là quá trình biến đổi hay kết hợp dữ liệu vào những dạng thích hợp cho quá trình khai phá dữ liệu
  - Chuẩn hóa (min-max, z-score, tỷ lệ thập phân)
    - Các giá trị thuộc tính được chuyển đổi vào một miền trị nhất định được định nghĩa trước
  - Tích hợp dữ liệu
    - Là quá trình trộn dữ liệu từ các nguồn khác nhau vào một
  - …

Chuẩn hóa dữ liệu

- Chuẩn hóa min-max

Thuộc tính A về từ min_A, max_B về miền mới new_min_A , new_min_B

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

Trong đó, v – giá trị cũ, v' là giá trị chuẩn hóa bằng phương pháp max min

- Ví dụ:

  - Thu nhập có giá trị từ 98000 đến 12000, chuẩn hóa max-min về [0.00,1.00]. Giá trị cần chuyển là 73600, khi đó giá trị mới là

- Nhận xét về chuẩn hóa min-max:

  - Là phép biến đổi tuyến tính - bảo tồn quan hệ giữa các giá trị ban đầu, tuy nhiên bị ảnh hưởng bởi các phần tử ngoại lai

39

39

Chuẩn hóa dữ liệu

- Chuẩn hóa z-score

$$v' = \frac{v - mean_A}{stand\_dev_A}$$

- Ví dụ: Giả sử có tập dữ liệu có mean là 54000 và stand_dev là 16000. Giá trị chuẩn hóa z-score của 73600 là 1,225
- Nhận xét:
  - Dựa trên giá trị trung bình và độ lệch chuẩn.
  - Z-score hữu ích khi chưa biết trị min và max thực sự của A, hoặc là khi các phân tử outliers ảnh hưởng mạnh đến kết quả chuẩn hóa min-max

40

40

Chuẩn hóa dữ liệu

- Chuẩn hóa theo tỷ lệ thập phân:
  - Giá trị cũ: v
  - Giá trị mới: $v' = \frac{v}{10^j}$

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

Thực chất là dịch chấm thập phân của các giá trị trong thuộc tính cũ , số bước di chuyển của chấm thập phân phụ thuộc vào giá trị có trị tuyệt đối lớn nhất trong tập giá trị cũ

- Ví dụ: A = { 568, -896, 500, 88}. Giá trị tuyệt đối lớn nhất trong A là: 896 → j = 3. Do đó, tập sau chuẩn hóa theo tỷ lệ thập phân A' = {0.568, -0.896, 0.50,0.088}

41

41

Thu gọn dữ liệu

- Thu gọn dữ liệu – giảm kích thước, số chiều dữ liệu gốc nhưng vẫn giữ được các "toàn vẹn dữ liệu"
- Kỹ thuật
  - Lấy mẫu
  - Lựa chọn 1 số thuộc tính
  - Giảm chiều dữ liêu
  - Rời rạc hóa
  - …

42

42

Lấy mẫu dữ liệu

- Lấy mẫu là một kỹ thuật cần thiết để xây dựng tập con dữ liệu, mục đích:
  - Giảm thời gian xử lý
  - Giảm không gian lưu trữ
  - Cân bằng giữa các lớp
- Kỹ thuật lấy mẫu:
  - Lấy mẫu ngẫu nhiên (Simple Random Sampling)
    - Nhược điểm: những lớp có tỷ lệ thấp có thể không được chọn
  - Lấy mẫu phân tầng
    - Chia tập dữ liệu thành các phân hoạch, lấy mẫu trong mỗi phân hoạch
    - Số lượng mẫu được chọn tỷ lệ kích thước của mỗi phân hoạch
    - Khắc phụ được sự không đồng đều của các lớp trong mẫu
  - …

43

43

Xây dựng tập con thuộc tính

- Loại các thuộc tính thừa
  - Thông tin được lấy từ một hoặc nhiều thuộc tính khác
  - Ví dụ: giá bán với thuế VAT
- Loại các thuộc tính không liên quan
  - Chứa thông tin không phục vụ cho mục đích khai phá dữ liệu
  - Ví dụ: mã số sinh viên không liên quan đến dự báo điểm học tập

44

44

Xây dựng tập con thuộc tính

- Lựa chọn một số thuộc tính
  - Phương pháp backward eliminaton: từ tập thuộc tính ban đầu, loại bỏ từng thuộc tính và đánh giá
  - Phương pháp forward selection: kết nạp từng thuộc tính và đánh giá

45

45

Vấn đề dư thừa dữ liệu

- Hiện tượng giá trị của một thuộc tính có thể được tính ra từ một/nhiều thuộc tính khác, vấn đề trùng lặp dữ liệu,
- Phát hiện dư thừa *thuộc tính số*: *phân tích tương quan* (correlation analysis)
  - Dựa trên tập dữ liệu hiện có, kiểm tra khả năng dẫn ra một thuộc tính B từ thuộc tính A
- Phát hiện dư thừa *thuộc tính phân loại*: sử dụng kiểm *định chi bình phương* (tự đọc)

46

46

Vấn đề dữ thừa dữ liệu

- Phân tích tương quan 2 thuộc tính A và B với kiểu dữ liệu số:
  - $r(A,B) = \frac{\sum_{i=1}^{N}(a_i - mean_A)(b_i - mean_B)}{N\sigma_A\sigma_B} \in [-1,1]$
  - Nếu $r(A,B) > 0$, thì A và B tương quan thuận với nhau, r(A,B) càng lớn thì mức độ tương quan càng cao, A hoặc B có thể được loại bỏ vì dư thừa.
  - Nếu $r(A,B) = 0$ thì A và B là độc lập
  - Nếu $r(A,B) < 0$, thì A và B tương quan nghịch với nhau

47

47

Bài tập 2 – Dư thừa thuộc tính

Income với Age? (phân tích tương quan)

Object	Income	Age	Class
1	3.000.000	23	Y
2	9.600.000	56	N
3	4.700.000	43	Y
4	7.000.000	30	N
5	6.200.000	65	N
6	2.200.000	26	Y
7	6.600.000	38	N
8	2.000.000	31	Y
9	6..300.000	37	Y
10	7.000.000	42	N
11	8.000.000	47	N
12	10.000.000	51	Y

48

48

Giảm chiều dữ liệu

- Giảm chiều: Chuyển không gian biểu diễn dữ liệu
- Ví dụ
  - Phân tích thành phần chính (PCA)
  - Phép biến đổi wavelet

Rời rạc hóa dữ liệu

- Rời rạc hóa dữ liệu
  - Giảm số lượng giá trị của một thuộc tính liên tục bằng các chia miền trị thuộc tính thành các thùng (bin)
  - Các nhãn được gán cho các bin này và được dùng thay giá trị thực của thuộc tính
  - Các trị thuộc tính có thể được phân hoạch theo một phân cấp hay ở nhiều mức phân cấp khác nhau

Rời rạc hóa dữ liệu

- Kỹ thuật rời rạc hóa
  - Phân hoạch theo chiều rộng
  - Phân hoạch theo chiều sâu
  - Phân hoạch dựa vào Entropy
  - …

Rời rạc hóa dữ liệu

- Phân hoạch theo chiều rộng (khoảng cách):
  - Chia phạm vi thành N thùng có kích thước bằng nhau
  - Giả sử A và B là giá trị thấp nhất và cao nhất của thuộc tính, thì độ rộng của các thùng sẽ là: W = (B - A) / N.
  - Ưu điểm: Đơn giản
  - Nhược điểm:
    - Tập dữ liệu không cân bằng sẽ không được xử lý tốt
    - Dữ liệu ngoại lai

Rời rạc hóa dữ liệu

- Phân hoạch theo chiều sâu (tần suất) :
  - Chia phạm vi thành N thùng, mỗi thùng chứa số lượng mẫu như nhau
  - Nhược điểm:
    - Cùng một giá trị liên tục có thể được gán vào các thùng khác nhau
    - Quản lý các thuộc tính phân loại có thể khó khăn

Rời rạc hóa dữ liệu

- Phân hoạch theo Entropy:
  - Ý tưởng chính là phân chia giá trị thuộc tính theo cách tạo ra các thùng càng "tinh khiết" càng tốt
  - Thước đo tạp chất của một thùng sao cho
    - Với thùng có phân phối chuẩn sẽ có tạp chất cao nhất
    - Một thùng có tất cả các giá trị thuộc cùng một lớp có tạp chất là zero (tinh khiết nhất)
    - Sự phân bố lớp trong thùng càng nhiều thì tạp chất càng nhỏ
  - Entropy có thể là một độ đo tạp chất

Tóm tắt

- Dữ liệu
  - Thuộc tính, kiểu thuộc tính
  - Các loại tập dữ liệu
- Chất lượng dữ liệu
  - Dữ liệu bị nhiễu, ngoại lai
  - Dữ liệu bị thiếu
- Tiền xử lý dữ liệu
  - Làm sạch dữ liệu
  - Biến đổi dữ liệu
  - Thu gọn dữ liệu