KHAI PHÁ DỮ LIỆU

Trường Đại học Nha Trang
Khoa Công nghệ thông tin
Bộ môn Hệ thống thông tin
Giáo viên: TS.Nguyễn Khắc Cường

CHỦ ĐỀ 2

CHUẨN BỊ DỮ LIỆU (Phần 2)

Chuẩn bị dữ liệu

- Datasets:
  - Là tập hợp các dữ liệu
  - Các thuật toán KPDL thực hiện công việc phân tích và xử lý dữ liệu trong các datasets này
- Missing data:
  - Trong đa số các trường hợp thu thập dữ liệu,
    - có rất ít trường hợp các dữ liệu được thu thập là đầy đủ
    - đa số là thiếu dữ liệu
  - Ảnh hưởng của missing data:
    - Có các thuật toán KPDL ít bị ảnh hưởng bởi missing data
    - Và có các thuật toán bị ảnh hưởng rất lớn bởi các missing data

Chuẩn bị dữ liệu

- Một số phương pháp cơ bản xử lý missing data:
  - Giữ nguyên các sample có missing data:
    - Một số phương pháp KPDL chấp nhận missing data và kết quả KPDL không bị ảnh hưởng nhiều bởi missing data
  - Loại bỏ toàn bộ các samples có missing data:
    - Chỉ nên làm điều này nếu số lượng sample có missing data là chiếm tỉ lệ không nhiều trong toàn bộ dataset
  - Tìm cách bổ sung các missing data
    - Tự nghĩ ra các giá trị hợp lý theo cách lý giải nào đó để điền vào các missing data → cách này dễ gây ra noise trong dataset
    - Thay tất cả các giá trị missing data của các sample bằng một hằng số chung nào đó
    - Thay tất cả các giá trị missing data của các sample trong một cột bằng một giá trị chung là giá trị trung bình của feature (cột) đó

Chuẩn bị dữ liệu

- Một số phương pháp cơ bản xử lý missing data:
  - Tìm cách bổ sung các missing data

    → Cách này có nhược điểm: có thể gây ra bias
  - Thay sample có missing data bằng một tập hợp các sample. VD:
    - Sample có missing data: X = {1, ?, 3}
    - Thay thế bằng các sample sau:
      - X1 = {1,0,3}
      - X2 = {1,1,3}
      - X3 = {1,2,3|
      - X4 = {1,3,3}
      - X5 = {1,4,3}

Chuẩn bị dữ liệu

- Một số phương pháp cơ bản xử lý missing data:
  - Tùy vào đặc điểm riêng của từng loại data rồi áp dụng các phương pháp sau để suy luân các giá trị cho các missing data:
    - Regression
    - Bayesian formalism
    - Clustering
    - Decision-tree induction
- Time-dependent data
  - Các dữ liệu trong thực tế có thể
    - Phụ thuộc chặt chẽ vào yếu tố thời gian
    - Có phụ thuộc nhưng không qua chặt chẽ vào yếu tố thời gian
    - Không phụ thuộc vào yếu tố thời gian

Chuẩn bị dữ liệu

- Time-dependent data
  - Chuẩn bị dữ liệu:
    - Riêng đối với dữ liệu phụ thuộc thời gian, việc chuẩn bị dữ liệu là rất quan trọng → vì ảnh hưởng lớn đến kết quả KPDL
  - Một số kỹ thuật chuẩn bị dữ liệu cơ bản:
    - Với dữ liệu được lấy mẫu với khoảng cách thời gian đều đặn bằng nhau
      - VD: nhiệt độ đo từng giờ, hàng hóa bán hàng ngày, …
      - Lập luận: giá trị sau có sự liên quan nào đó với dữ liệu trước
      - Ký hiệu: $X = \{t(1), t(2), t(3), \ldots, t(n)\}$
      - Yêu cầu: chuân bị dữ liệu băng cách nào đó để
        - Dựa vào n giá trị đã thu được
        - Dự đoán được giá trị thứ n+1 với độ chính xác cao

Chuẩn bị dữ liệu

- Time-dependent data
  - Một số kỹ thuật chuẩn bị dữ liệu cơ bản:
    - Với dữ liệu được lấy mẫu với khoảng cách thời gian đều đặn bằng nhau
      - Kỹ thuật cơ bản: dùng một window để cắt các dãy giá trị liên tiếp thành các samples
      - VD:
        - Thu thập được chuỗi dữ liệu theo thời gian như sau

$$X = \{t(0), t(1), t(2), t(3), t(4), t(5), t(6), t(7), t(8), t(9), t(10)\}$$

        - Giả sử chọn window có size = 5 → chuỗi dữ liệu trên được biến đổi thành bảng dữ liệu như sau

Chuẩn bị dữ liệu

- Time-dependent data
  - Một số kỹ thuật chuẩn bị dữ liệu cơ bản:
    - Với dữ liệu được lấy mẫu với khoảng cách thời gian đều đặn bằng nhau
      - Giả sử chọn window có size = 5 → chuỗi dữ liệu trên được biến đổi thành bảng dữ liệu có 6 sample như sau:

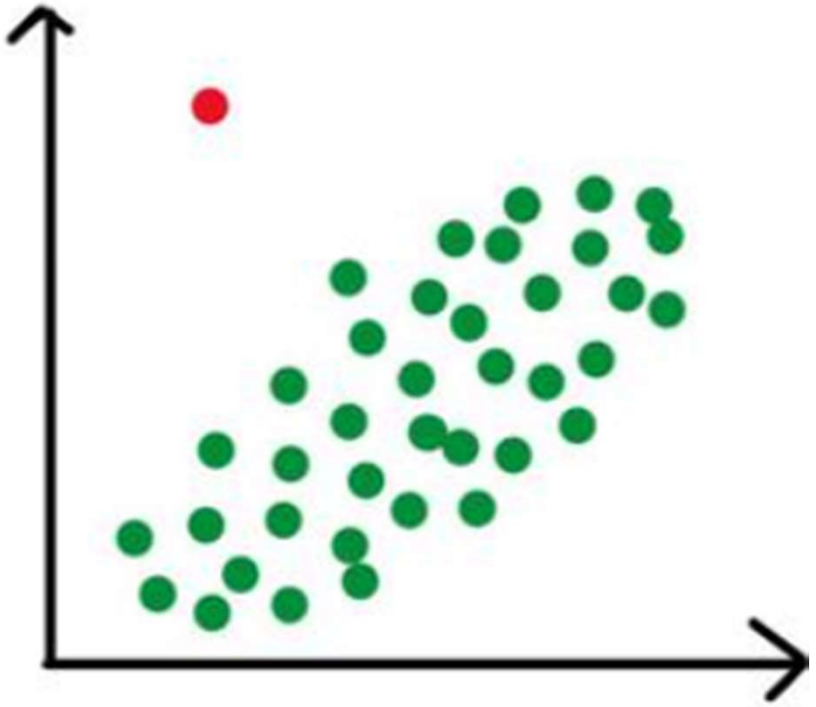
Sample	Window					Next Value
	M1	M2	M3	M4	M5	
1	$t(0)$	$t(1)$	$t(2)$	$t(3)$	$t(4)$	$t(5)$
2	$t(1)$	$t(2)$	$t(3)$	$t(4)$	$t(5)$	$t(6)$
3	$t(2)$	$t(3)$	$t(4)$	$t(5)$	$t(6)$	$t(7)$
4	$t(3)$	$t(4)$	$t(5)$	$t(6)$	$t(7)$	$t(8)$
5	$t(4)$	$t(5)$	$t(6)$	$t(7)$	$t(8)$	$t(9)$
6	$t(5)$	$t(6)$	$t(7)$	$t(8)$	$t(9)$	$t(10)$

Chuẩn bị dữ liệu

- Time-dependent data
  - Một số kỹ thuật chuẩn bị dữ liệu cơ bản:
    - Với dữ liệu được lấy mẫu với khoảng cách thời gian đều đặn bằng nhau
      - Vấn đề: Kích thước của window bao nhiêu là tốt?
      - Giải quyết: tìm ra bằng thực nghiệm đối với từng loại data khác nhau
      - Kỹ thuật khác:
        - Dùng difference: $t(n+1)-t(n)$ thay vì dùng trực tiếp $t(n+1)$
        - Dùng ratio: $t(n+1)/t(n)$ thay vì dùng trực tiếp $t(n+1)$
        - . . .

Chuẩn bị dữ liệu

- Outlier analysis
  - Outlier:
    - trong một dataset, có một hoặc vài samples có giá trị rất khác biệt so với số đông các giá trị còn lại trong dataset

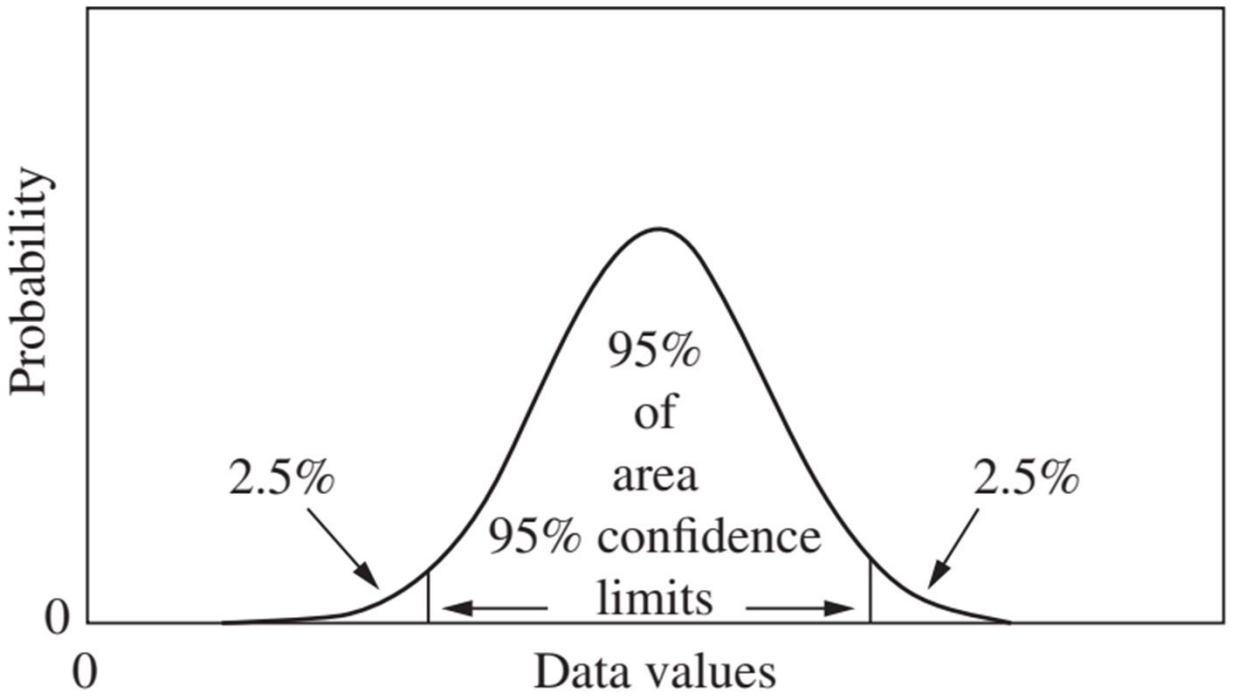


Chuẩn bị dữ liệu

- Outlier analysis
  - Outlier:
    - Có rất nhiều lý do làm xuất hiện các sample là các outlier trong dataset
    - Các thuật toán KPDL
      - cố gắng giảm sự ảnh hưởng của các outlier đến kết quả KPDK
      - hoặc cố gắng loại bỏ outlier trước khi thực hiện KPDL
  - Một số phương pháp cơ bản dùng phát hiện outlier
    - graphical / visualization techniques
    - statistical-based techniques
    - distance-based techniques
    - model-based techniques

Chuẩn bị dữ liệu

- Outlier analysis
  - VD1: statistical-based techniques cho rằng dùng mean value và standard deviation có thể loại bỏ được outlier



Chuẩn bị dữ liệu

- Outlier analysis
  - VD2:
    - Giả sử có dataset chứa dữ liệu 1-D như sau

$$Age = \{3, 56, 23, 39, 156, 52, 41, 22, 9, 28, 139, 31, 55, 20, -67, 37, 11, 55, 45, 37\}$$

    - Tính được:
      - Mean = 39.9
      - Standard deviation = 45.65
    - Cho rằng các dữ liệu trên tuân theo normal distribution → đề xuất chọn threshold value cho phân bố trên là:

$$\text{Threshold} = \text{Mean} \pm 2 \times \text{Standard deviation}$$

→ Threshold_max = 39.9 + 2 x 45.56 = 131.02

Threshold_min = 39.9 - 2 x 45.56 = -51.22

    - Kết luận: mọi dữ liệu ngoài khoảng [-51.22, 131.02] là outlier

Chuẩn bị dữ liệu

- Outlier analysis
  - VD3: Với dataset có dữ liệu nhiều chiều, một kỹ thuật cơ bản có thể phát hiện được outlier như sau
    - Tính covariance matrix

$$V_n = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}_n)(x_i - \bar{x}_n)^T$$

Trong đó

- n : số lượng sample
- x_i : từng sample
- $\bar{x}_n$: mean vector

Chuẩn bị dữ liệu

- Outlier analysis
  - VD3: Với dataset có dữ liệu nhiều chiều, một kỹ thuật cơ bản có thể phát hiện được outlier như sau
    - Tính Mahalanobis distance cho từng sample đối với mean vector

$$M_i = \left(\sum_{i=1}^{n} (x_i - \bar{x}_n)^T \mathrm{V}_n^{-1} (x_i - \bar{x}_n) \right)^{1/2}$$

    - Kết luận:
      - Các sample nào có giá trị Mahalanobis lớn nhất là các outlier

Chuẩn bị dữ liệu

- Biến đổi dữ liệu
  - Tác dụng:
    - Các phép biến đổi dữ liệu giúp nâng cao hiệu quả, độ chính xác của các giải thuật KPDL
    - Dựa vào kiểu dữ liệu và đặc điểm của từng loại dữ liệu để chọn phương pháp biến đổi phù hợp
  - Normalization
    - Thường dùng cho các phương pháp KPDL distance-based
    - Mục tiêu chung: biến đổi toàn bộ giá trị vào khoảng [0,1] hoặc [-1,1]
    - Một số phương pháp normalization cơ bản:
      - Decimal scaling: $v'(i) = \frac{v(i)}{10^k}$ scaling các giá trị vào [-1,1]
        - v(i) : giá trị của sample thứ i tại feature v
        - v'(i) : giá trị sau khi đã scaling
        - Chọn k là nhỏ nhất sao cho sao cho $\max(|v'(i)|) < 1$

Chuẩn bị dữ liệu

- Biến đổi dữ liệu
  - Một số phương pháp normalization cơ bản:
    - Decimal scaling:
      - VD:
        - Xét tất cả các giá trị của các sample tại tất cả các feature, có:
          - Giá trị lớn nhất là 455
          - Giá trị nhỏ nhất là -834
        - Như vậy, chọn k=3 sẽ đáp ứng được điều kiện $\max\left(|v'(i)|\right) < 1$ vì |-834| : 1000 = 0.834 < 1
    - Min-Max normalization:
      - Công thức:

$$v'(i) = \frac{(v(i) - \min(v(i)))}{(\max(v(i)) - \min(v(i)))}$$

      - Scaling các giá trị vào [0,1]

Chuẩn bị dữ liệu

- Biến đổi dữ liệu
  - Một số phương pháp normalization cơ bản:
    - Standard deviation normalization:
      - Công thức:

$$v*(i) = \frac{(v(i) - \text{mean}(v))}{\text{sd}(v)}$$

      - Scaling các giá trị vào [-1,1]
- Làm trơn dữ liệu:
  - Một số giải thuật KPDL rất nhạy với mức độ sai khác giữa các giá trị của dữ liệu
  - Việc giảm bớt sự sai khác này có thể giúp nâng cao hiệu quả và độ chính xác của các giải thuật KPDL

Chuẩn bị dữ liệu

- Làm trơn dữ liệu:
  - Có nhiều kỹ thuật làm trơn dữ liệu
  - Làm tròn số là một trong số các kỹ thuật đơn giản nhất của các kỹ thuật làm trơn dữ liệu
  - VD:
    - Giả sử có các giá trị tại feature F như sau:

$$\{0.93,\ 1.01,\ 1.001,\ 3.02,\ 2.99,\ 5.03,\ 5.01,\ 4.98\}$$

    - Làm trơn:

$$F_{\text{smoothed}} = \{1.0,\ 1.0,\ 1.0,\ 3.0,\ 3.0,\ 5.0,\ 5.0,\ 5.0\}$$

    - Nhận xét:
      - Rõ ràng là làm tròn số đã làm giảm mức đô sai khác
      - Và không làm thay đổi quá nhiều chất lượng của dữ liệu

Chuẩn bị dữ liệu

Q / A