ĐẠI HỌC BÁCH KHOA HÀ NỘI
VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

BÀI 3: TRỰC QUAN HÓA DỮ LIỆU

Nội dung

1. Biểu đồ tĩnh
2. Trực quan hóa theo điểm ảnh
3. Trực quan hóa trên không gian véc-tơ
4. Cây siêu cầu
5. SOM

1. Biểu đồ tĩnh
1.1 Thuộc tính

- Đối tượng DL đại diện cho các thực thể trong DL (vd khách hàng, sản phẩm, giao dịch)
- Đối tượng DL còn được gọi là mẫu, ví dụ hoặc điểm DL
- Thuộc tính là một trường DL, thể hiện một tính chất hoặc đặc trưng của DL
- Thuộc tính còn được gọi là chiều, đặc trưng hoặc biến

Thuộc tính (tiếp)

- Các giá trị của một thuộc tính cho trước được gọi là các quan sát
- Tập hợp các thuộc tính mô tả một đối tượng cho trước được gọi là một véc-tơ thuộc tính (hoặc véc-tơ đặc trưng)
- Kiểu thuộc tính được xác định bởi tập hợp các giá trị của thuộc tính

Thuộc tính định danh

- Có giá trị là các biểu tượng hoặc tên
- VD: ‘màu tóc’ gồm ‘xanh’, ‘đỏ’, ‘đen’, ‘trắng’, ‘bạch kim’
- Mô tả các thể loại, mã, trạng thái
- Giá trị phổ biến dựa trên hàm *mode*

Thuộc tính nhị phân

- Thuộc tính thể loại chỉ có hai thể loại hoặc hai trạng thái
  - 0 ~ vắng mặt, 1 ~ tồn tại
  - hoặc 0 ~ sai, 1 ~ đúng
- Thuộc tính đối xứng (vd: ‘giới tính’ gồm ‘nam’ và ‘nữ’)
- Thuộc tính bất đối xứng (vd: ‘kết quả’ gồm ‘dương tính’ và ‘âm tính’

Thuộc tính thứ tự

- Các giá trị tuân theo thứ tự nhất định
- VD: ‘kích cỡ’ gồm ‘nhỏ’, ‘bình thường’, ‘lớn’ và ‘ngoại cỡ’
- Giá trị phổ biến dựa trên hàm *mode* và *median*

Thuộc tính khoảng cách

- Thuộc tính số đo đạc theo tỉ lệ của giá trị đơn vị
- Có thể so sánh, tính khoảng cách giữa các giá trị
- VD: Nhiệt độ theo thang đo Celcius

Thuộc tính tỉ lệ

- Thuộc tính số có giá trị 0
- Có thể nhân các giá trị với nhau
- VD: Các giá trị đếm và đo đạc:
  - Số lượng
  - Trọng lượng
  - Chiều cao
  - Số tiền
  - ...

Thuộc tính rời rạc vs liên tục

- Thuộc tính rời rạc có tập giá trị hữu hạn hoặc tập giá trị vô hạn đếm được. VD:
  - Tập hữu hạn: màu sắc, tuổi
  - Tập vô hạn đếm được: ID của khách hàng
- Thuộc tính là liên tục nếu không phải là rời rạc

1.2 Các phép thống kê DL cơ bản

- Mô tả DL:
  - Giá trị trung tâm
  - Phạm vi phân bố
  - Trực quan hóa dựa trên các biểu đồ
- Nhận diện phần tử ngoại lai

mean (trung bình)

- Các giá trị có vai trò như nhau

$$x = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

- Các giá trị có trọng số khác nhau

$$x = \frac{w_1x_1 + w_2x_2 + \ldots + w_nx_n}{n}$$

- Phép đo phổ biến nhất, tuy nhiên nhạy cảm với phần tử ngoại vi

median (trung vị)

- Giá trị trung vị chia DL thành hai phần lớn hơn và nhỏ hơn; hai phần này có số phần tử bằng nhau
- Các tính xấp xỉ trung vị
  - Nhóm DL vào các khoảng giá trị
  - Tính tần xuất giá trị trong mỗi khoảng
  - Tìm khoảng có chứa tần xuất trung vị

median (trung vị) (tiếp)

- Xấp xỉ trung vị theo công thức:

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

trong đó:

- L_1 là biên dưới của khoảng trung vị
- N là số giá trị
- $(\Sigma freq)_l$ là tổng số tần xuất của các khoảng bé hơn khoảng trung vị
- $freq_{median}$ là tần xuất của khoảng trung vị
- *width* là độ rộng của khoảng trung vị

mode

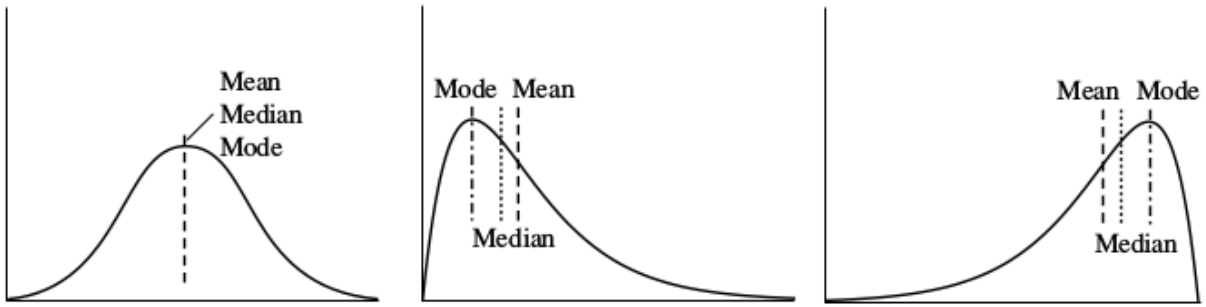
- Giá trị phổ biến nhất trong tập DL
- Multimodal: Tập có nhiều giá trị phổ biến
- Tập chỉ chứa các giá trị duy nhất không có mode
- Với tập unimodal:

$$mean - mode \approx 3 \times (mean - median)$$

midrange

- Trung bình của giá trị lớn nhất và giá trị nhỏ nhất trong tập

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



(a) Symmetric data (b) Positively skewed data (c) Negatively skewed data

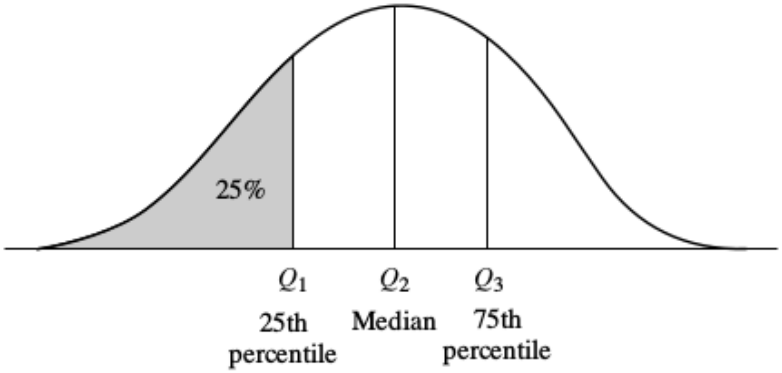
range

- Khoảng cách giữa giá trị lớn nhất và nhỏ nhất trong tập

$$range = max - min$$

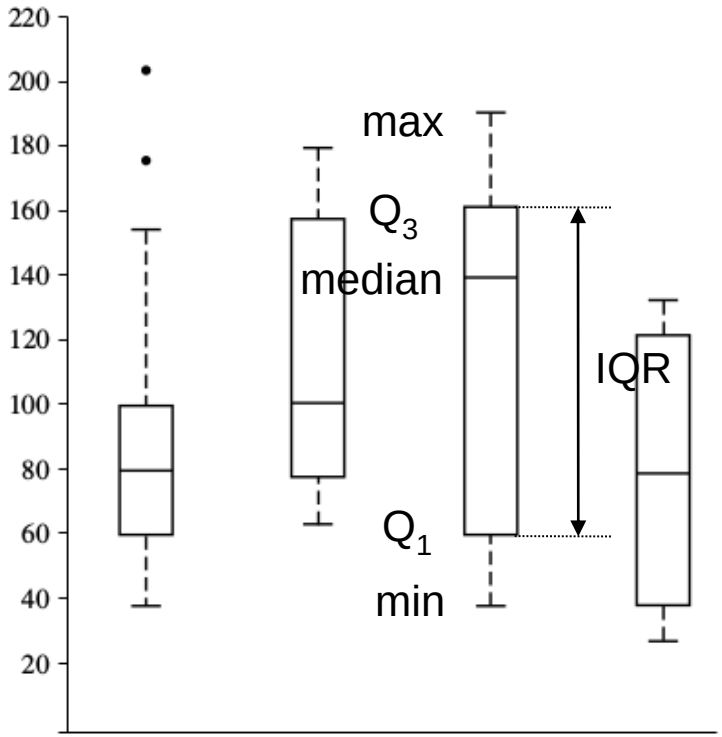
quantile

- Quantile là các điểm chia DL thành các phần (gần) bằng nhau (có số phần tử bằng nhau)
  - 2-quantile: một điểm chia DL thành hai phần bằng nhau ~ trung vị
  - 4-quantile (quartile)
  - 100-quantile (percentile)
- Interquartile range IQR = $Q_3 - Q_1$



boxplot (biểu đồ hộp)

- Biểu đồ hộp bao gồm:
  - Q_1, Q_3: Điểm đầu và cuối của hộp
  - IQR: Độ dài của hộp
  - Trung vị
  - Giá trị min và max



variance, standard deviation

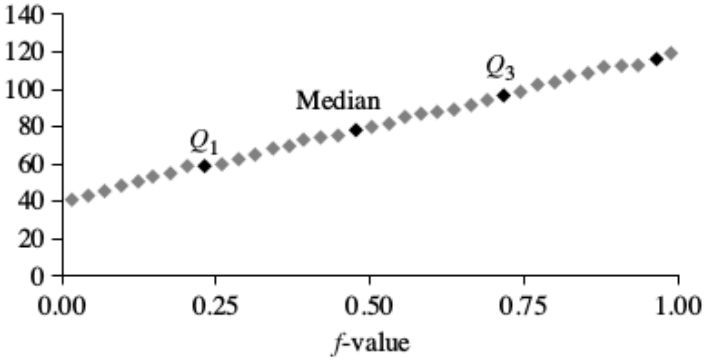
- Variance (phương sai)

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

- σ: standard deviation (độ lệch chuẩn) thể hiện mức độ phân tán của DL so với giá trị trung bình (mean)

Biểu đồ quantile

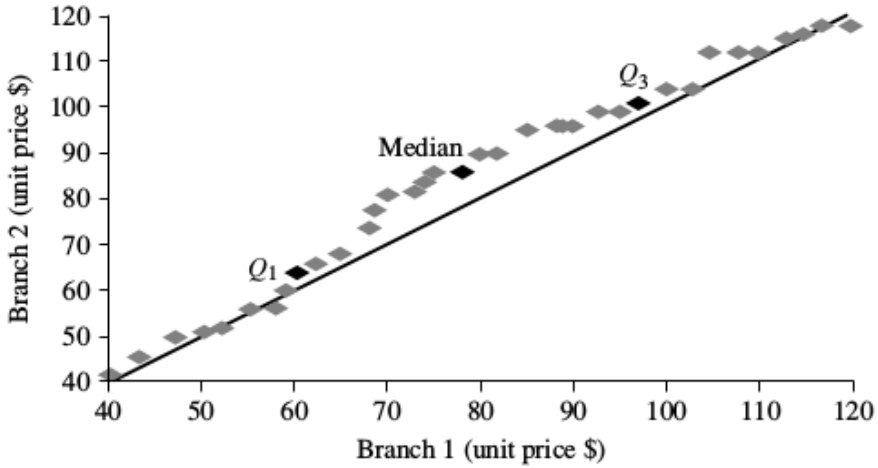
- Sắp xếp các giá trị theo thứ tự tăng dần $x_1 < x_2 < \ldots x_n$
- Tần xuất f_i tương ứng với x_i là tỉ lệ phần trăm DL có giá trị dưới x_i



$$f_i = \frac{i - 0.5}{N}$$

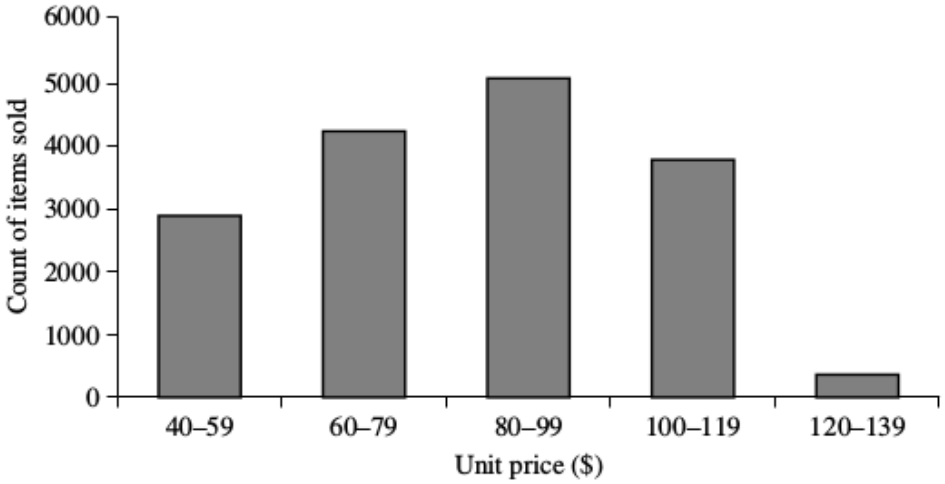
Biểu đồ quantile - quantile

- Thể hiện quan hệ giữa các giá trị quantile của hai phân phối đơn biến



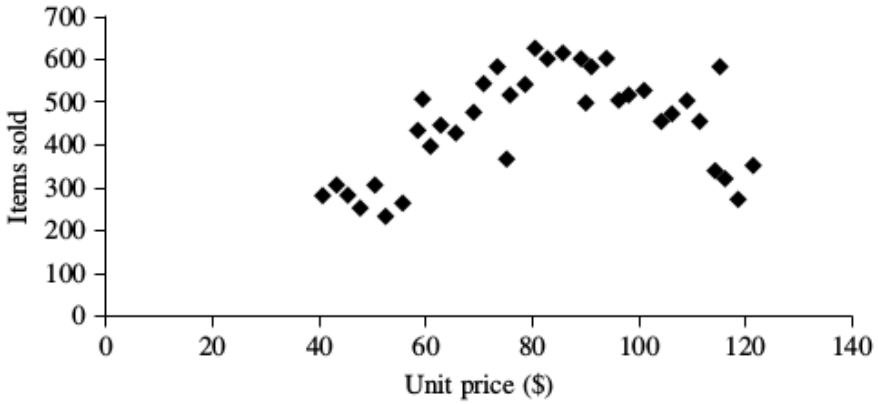
Histogram

- Các giá trị được gom vào các khoảng bằng nhau gọi là các bin (bucket)

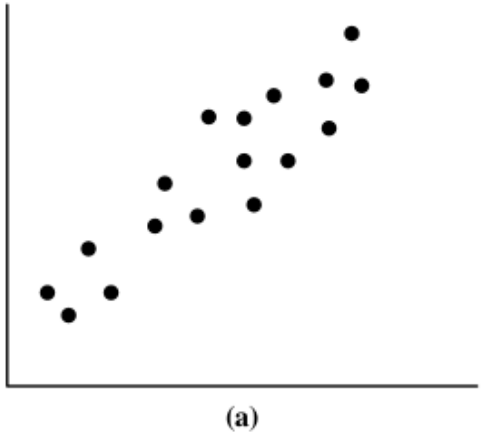


Biểu đồ scatter

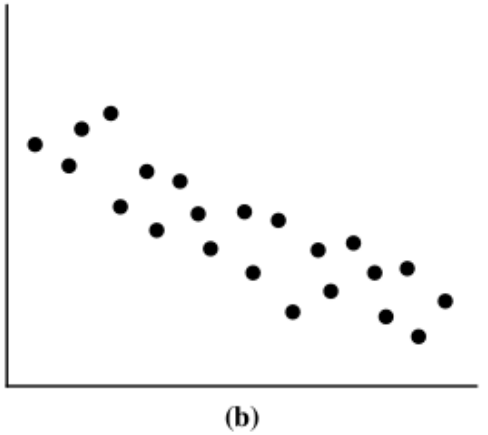
- Xác định tính tương hỗ giữa hai thuộc tính số



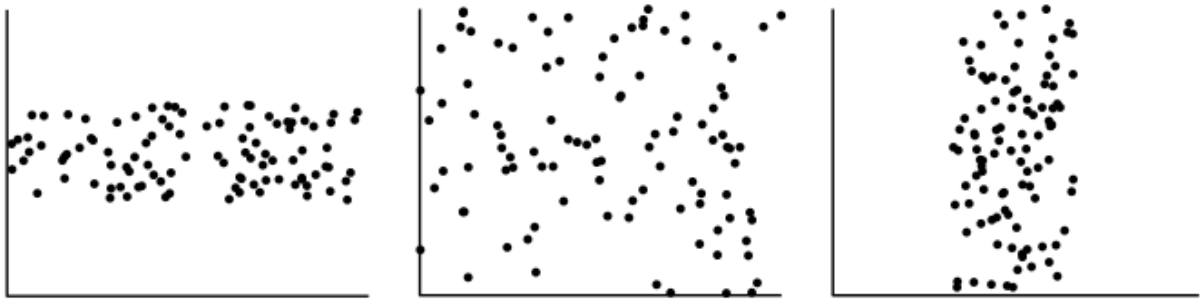
Biểu đồ scatter (tiếp)



Tương hỗ dương



Tương hỗ âm



Tương hỗ null (không tương hỗ)

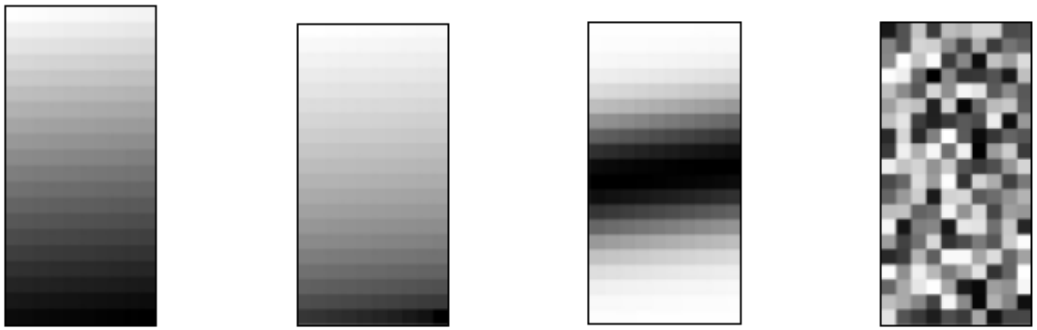
2. Trực quan hóa theo điểm ảnh

- Giá trị của một chiều DL được biểu diễn bằng một điểm ảnh với màu tương ứng với giá trị
- VD: Giá trị nhỏ tương ứng với màu sáng, giá trị lớn tương ứng với màu tối
- *m* chiều tương ứng với *m* cửa sổ. Một điểm DL có *m* chiều được biểu diễn bởi *m* điểm ảnh ở các vị trí tương ứng tại mỗi cửa sổ

Trực quan hóa theo điểm ảnh (tiếp)

- Các bản ghi thường được sắp xếp theo một chiều DL được quan tâm
- Tương quan nếu có giữa các chiều DL được thể hiện thông qua phân bố màu (giá trị DL) trên các cửa sổ

Trực quan hóa theo điểm ảnh (tiếp)



(a) *income* **(b)** *credit_limit* **(c)** *transaction_volume* **(d)** *age*

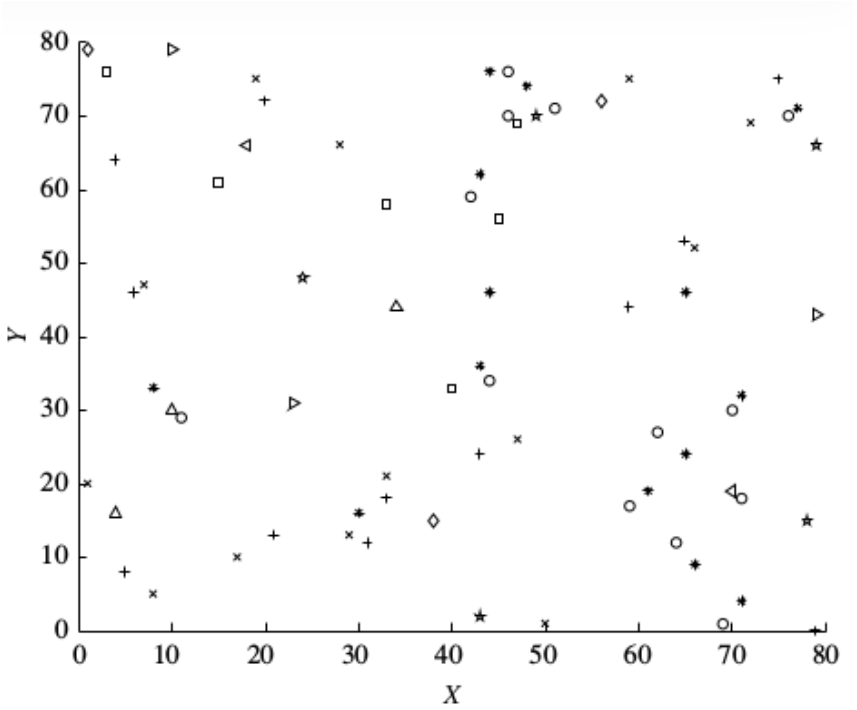
3. Trực quan hóa trên không gian véc-tơ

- Trực quan hóa dựa trên điểm ảnh không thể hiện được mật độ của các điểm DL
- Trực quan hóa trên không gian véc-tơ dựa trên kĩ thuật chiếu để biểu diễn DL đa chiều trên không gian 2 chiều

Biểu đồ scatter

- Cách biểu diễn:
  - Hai trục X và Y dùng để biểu diễn hai chiều số theo tọa độ Cartesian
  - Chiều thứ ba được biểu diễn bởi các hình khác nhau
  - Chiều thứ tư có thể được biểu diễn bởi màu sắc

Biểu đồ scatter (tiếp)

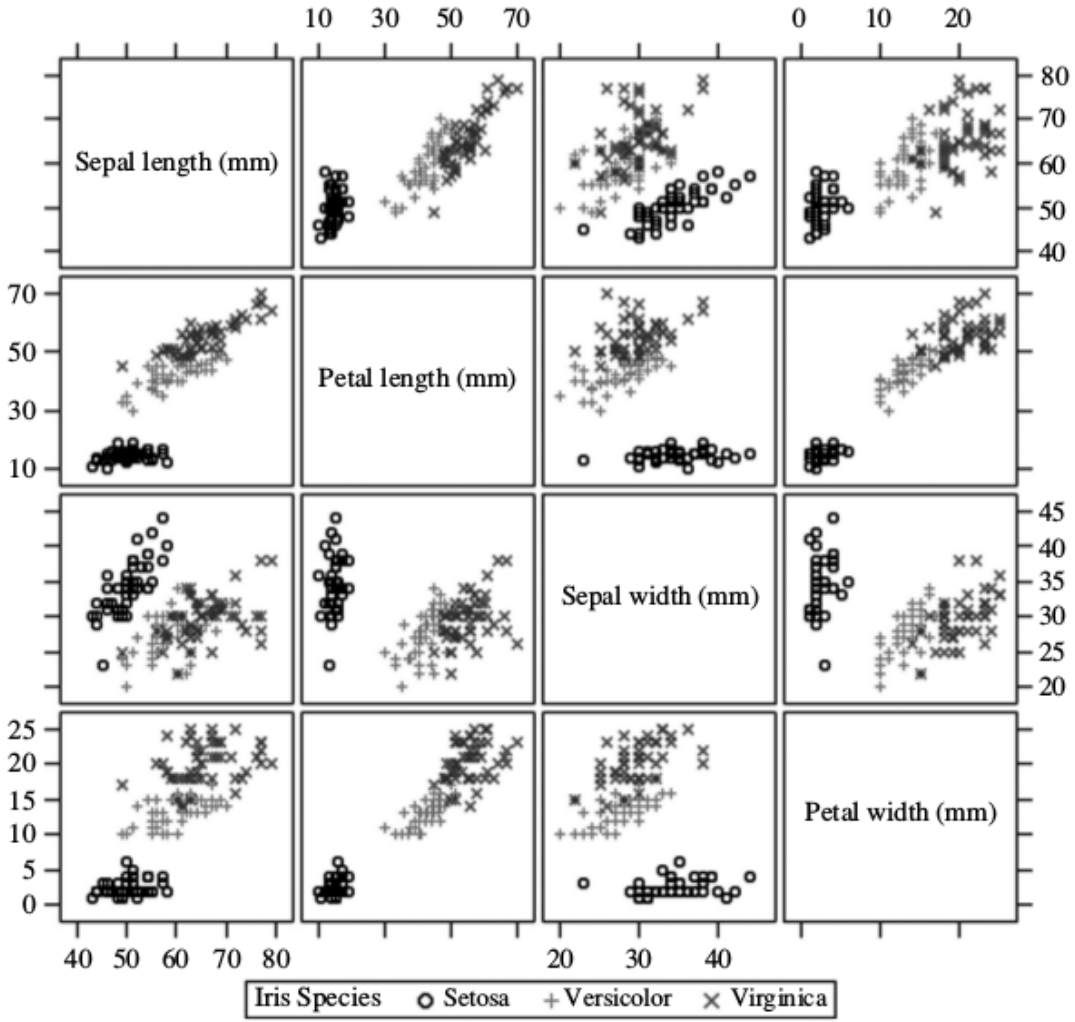


www.cs.sfu.ca/jpei/publications/rareevent-geoinformatica06.pd

Ma trận biểu đồ scatter

- Biểu đồ scatter chỉ có thể biểu diễn được tối đa 4 chiều
- Với DL có nhiều hơn 4 chiều, sử dụng phương pháp mở rộng của biểu đồ scatter
  - DL có *m* chiều
  - Sử dụng một ma trận m x m biểu đồ scatter 2D để biểu diễn mỗi chiều DL với các chiều DL còn lại
  - VD: Tập DL *Iris* có 5 chiều được trực quan hóa bởi ma trận 4 x 4 gồm 24 biểu đồ scatter 3D

VD: Tập DL *Iris*

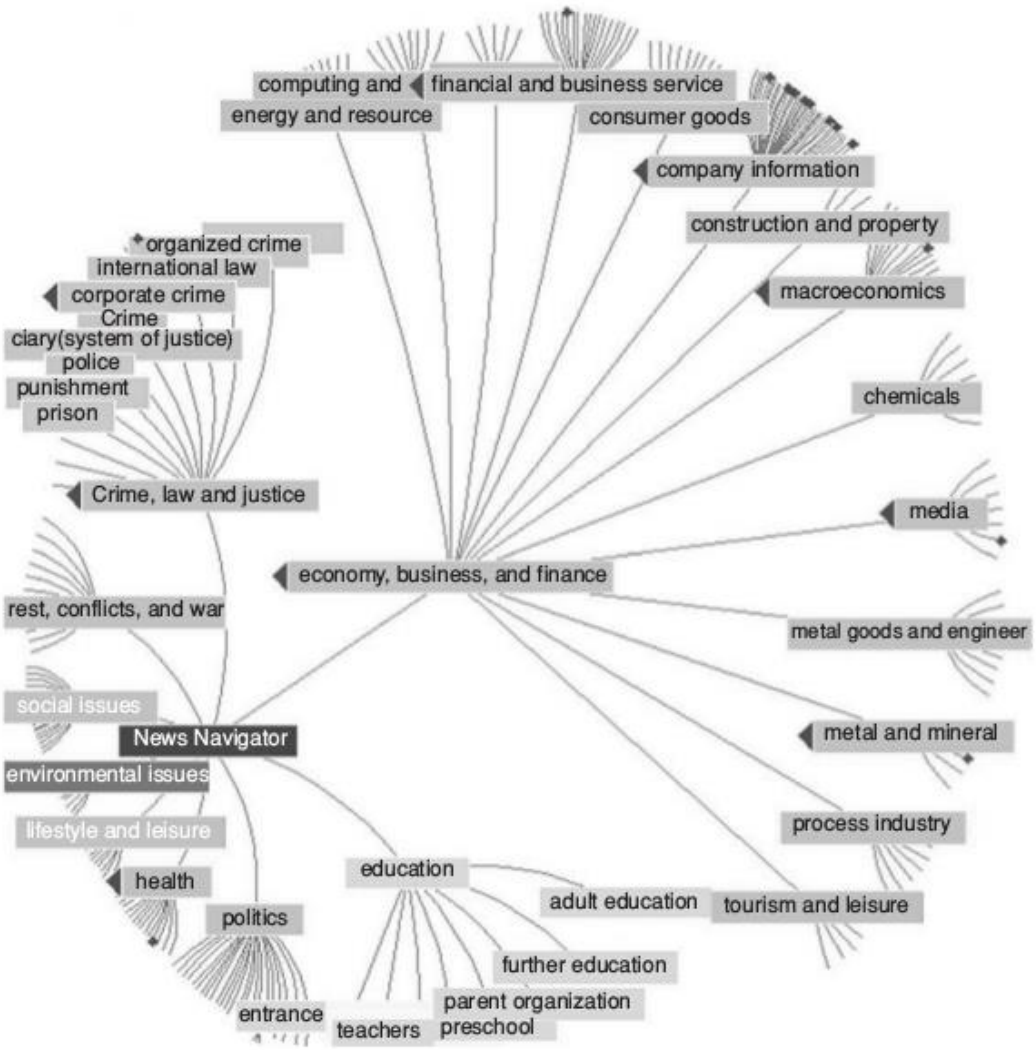


http://support.sas.com/documentation/cdl/en/grstatproc/61948/HTML/default/images/gsgscmat.gif

4. Cây siêu cầu

- Trực quán hóa lượng DL lớn
- DL có cấu trúc cây
- Một mặt tập trung vào một phần của DL, mặt khác vẫn biểu diễn ngữ cảnh chung của DL
- Tính chất mắt cá (fisheye):
  - Kích thước các nút không được tập trung nhanh chóng giảm đi khi
  - Kích thước các nút được tập trung nhanh chóng tăng lên

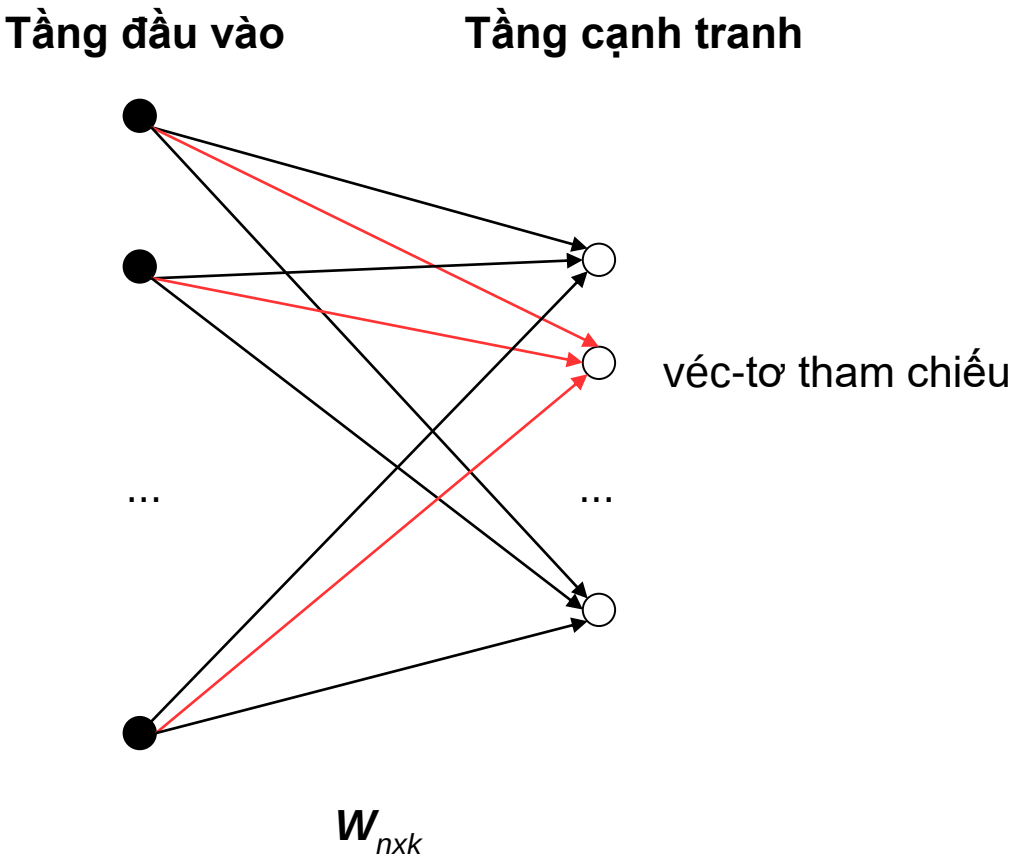
VD


computing and
financial and business service
energy and resource
consumer goods
company information
construction and property
organized crime
international law
corporate crime
Crime
ciary(system of justice)
police
punishment
prison
macroeconomics
chemicals
Crime, law and justice
media
economy, business, and finance
rest, conflicts, and war
metal goods and engineer
social issues
News Navigator
metal and mineral
environmental issues
process industry
lifestyle and leisure
education
health
adult education
tourism and leisure
politics
further education
parent organization
entrance
teachers
preschool

5. SOM (Self Organizing Map)

- SOM học ra biểu diễn 2 chiều của DL đa chiều
- SOM là một mạng nơ-ron tiến 2 tầng
  - Tầng đầu vào nhận tín hiệu từ DL đầu vào, có số chiều bằng số chiều của DL
  - Tầng cạnh tranh được tổ chức theo một hình trạng nhất định (hình chữ nhật, hình lục giác…) thể hiện mối quan hệ không gian giữa các nơ-ron
  - Mỗi nơ-ron ở tầng cạnh tranh có các trọng số liên kết từ tầng đầu vào gọi là véc-tơ tham chiếu

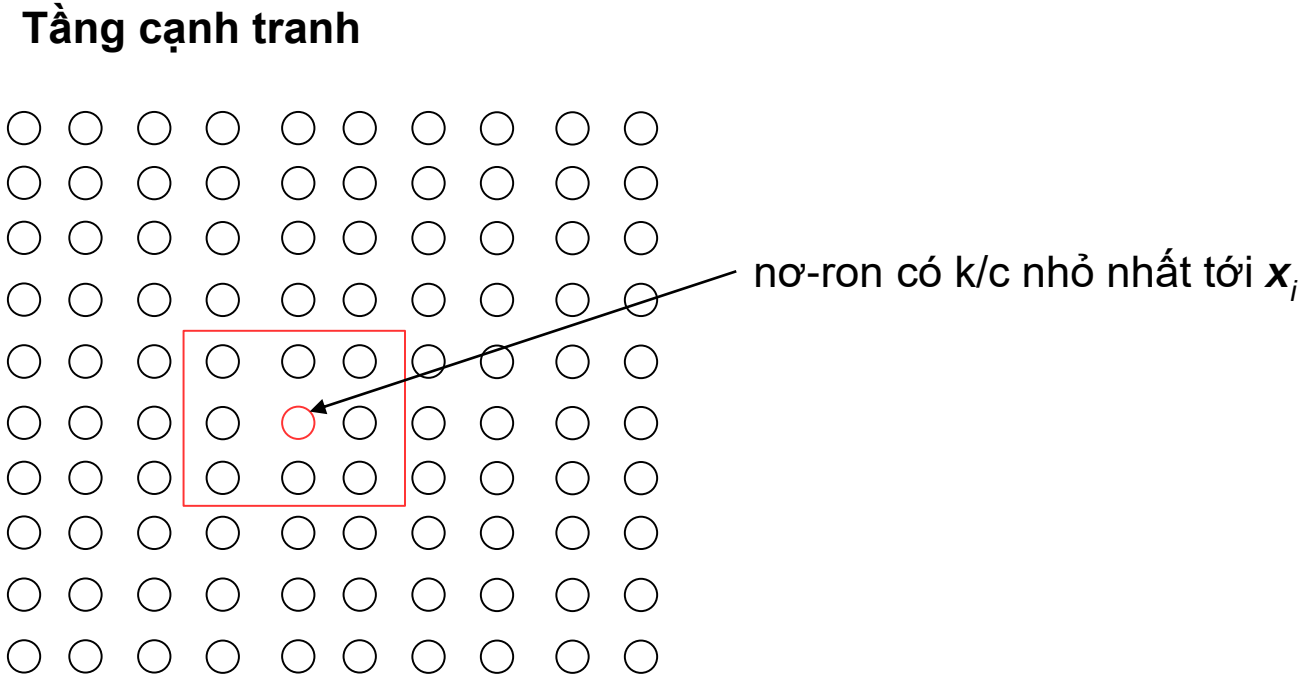
Kiến trúc SOM



Học cạnh tranh

- Với mỗi tín hiệu đầu vào $\boldsymbol{x}_i$, cần lựa chọn ra nơ-ron m_k có khoảng cách tới $\boldsymbol{x}_i$ là bé nhất
- Khoảng cách giữa $\boldsymbol{x}_i$ và m_k: Khoảng cách euclide giữa x_i và véc-tơ tham chiếu của m_k
- Hàm mục tiêu: Cực tiểu hóa tổng khoảng cách giữa các tín hiệu đầu vào và nơ-ron gần nhất tương ứng
- Cập nhật trọng số: Chỉ cập nhật trọng số của m_k và các nơ-ron lân cận

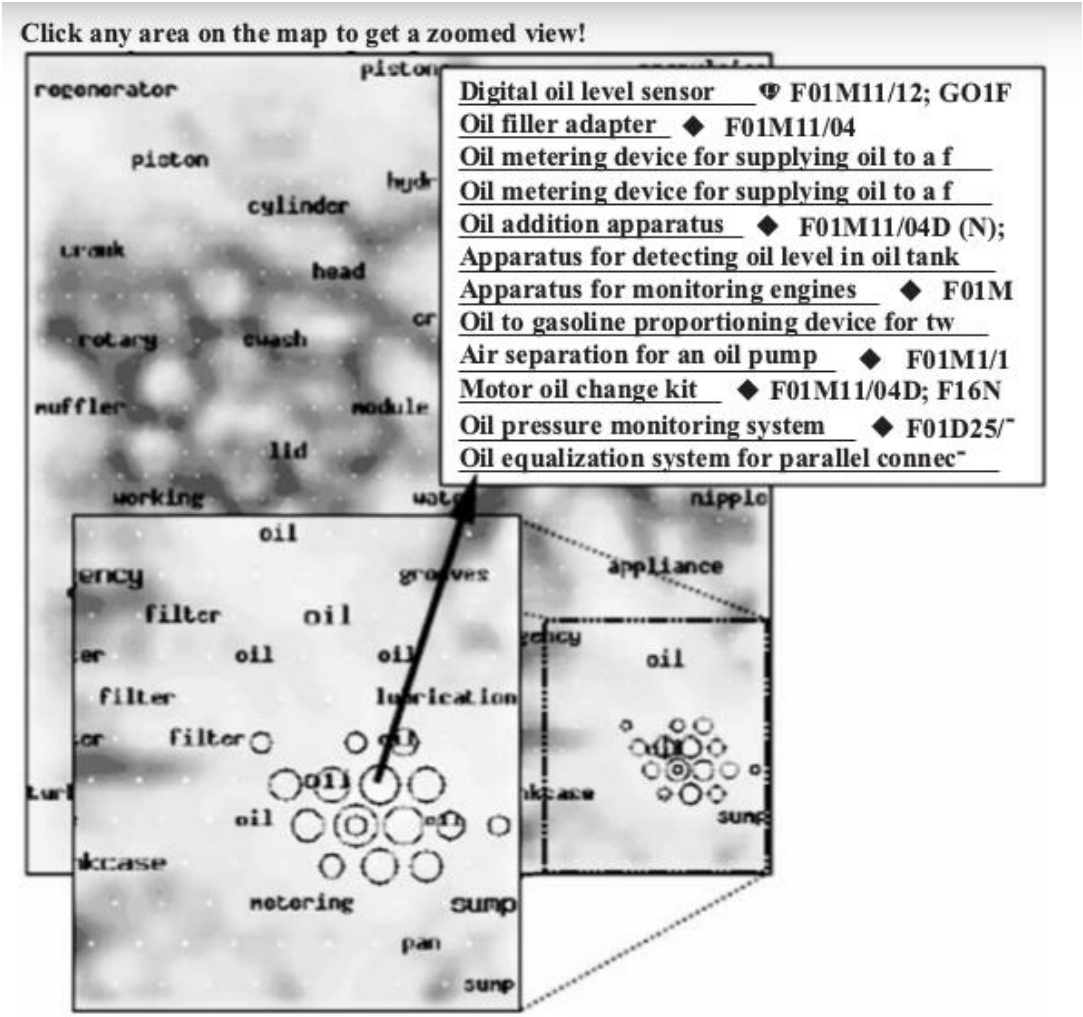
VD: Nơ-ron lân cận



WEBSOM

- Biểu diễn tập các văn bản trên bản đồ 2 chiều
- Các văn bản được biểu diễn dưới dạng túi từ
- Sau khi học, mỗi nhóm văn bản có thể được biểu diễn bởi các từ khóa đặc trưng
- Các vùng có mật độ lớn tập trung nhiều văn bản

WEBSOM


Click any area on the map to get a zoomed view!
Digital oil level sensor ◆ F01M11/12; GO1F
Oil filler adapter ◆ F01M11/04
Oil metering device for supplying oil to a f
Oil metering device for supplying oil to a f
Oil addition apparatus ◆ F01M11/04D (N);
Apparatus for detecting oil level in oil tank
Apparatus for monitoring engines ◆ F01M
Oil to gasoline proportioning device for tw
Air separation for an oil pump ◆ F01M1/1
Motor oil change kit ◆ F01M11/04D; F16N
Oil pressure monitoring system ◆ F01D25/
Oil equalization system for parallel connec
regenerator
piston
cylinder
crank
head
rotary
muffler
module
lid
working
nipple
appliance
oil
filter
lubrication
metering
sump
pan

ĐẠI HỌC BÁCH KHOA

25 YEARS ANNIVERSARY SOICT

VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY

Thank you for your attentions!

soict.hust.edu.vn/ fb.com/groups/soict