

TRƯỜNG ĐẠI HỌC BÁCH KHOA TP. HCM
Khoa KTXD - Bộ môn Kỹ thuật & Quản lý Tài nguyên nước

PHÂN TÍCH DỮ LIỆU



Giảng viên: PGS. TS. NGUYỄN THÔNG
E-mail: nguyenthong@hcmut.edu.vn or nthong56@yahoo.fr
Web: <http://www4.hcmut.edu.vn/~nguyenthong/index>
PGS. TS. Nguyễn Thông | Tél. (08) 38 691 592- 098 99 66 719

PHÂN TÍCH DỮ LIỆU

Chương 5: Phân nhóm dữ liệu (Cluster)

NỘI DUNG MÔN HỌC

Chương 1. Thống kê mô tả (ôn).
Chương 1bis. Xác suất & phân phối thống kê (ôn)..
Chương 2. Khoảng tin cậy.
Chương 2bis: Kỹ thuật chọn mẫu
Chương 3. Kiểm định thống kê.
Chương 4. Phân khúc dữ liệu (Segmentation).
Chương 5. Phân nhóm dữ liệu (Cluster).
Chương 6. Phân tích thành phần chính (PCA).
Chương 7. Phân tích chuỗi thời gian.
Chương 8. Hồi quy tuyến tính.
Chương 9. Xử lý số liệu thực nghiệm.
Chương 10. Giới thiệu phần mềm SPSS or R

PHÂN TÍCH DỮ LIỆU

Chương 5: Phân nhóm dữ liệu (Cluster)

GIỚI THIỆU VÀ ĐẶT VẤN ĐỀ

- Phân nhóm (cluster, groupe) là một kỹ thuật được ứng dụng trong nhiều lĩnh vực. Ví dụ :
 - Sinh học : phân nhóm các loại cây, các động vật theo một số đặc tính nào đó (định lượng hoặc định tính).
 - Y học : phân loại các loại bệnh nhân theo một số chỉ tiêu y học.

PGS. TS. Nguyễn Thông

PHÂN TÍCH DỮ LIỆU

Chương 5: Phân nhóm dữ liệu (Cluster)

→ Marketing và nghiên cứu thị trường

- Phân loại khách hàng** – nhận biết các thành phần có cùng cách tiêu thụ hàng hóa giống nhau.
- Phân loại sản phẩm** – nghiên cứu nhiều nhãn hiệu khác nhau của 1 sản phẩm theo các thuộc tính của chúng bởi người tiêu dùng. Các nhãn hiệu xuất hiện trong cùng một nhóm sẽ có các thuộc tính « gần gũi » nhau nhất, trên cơ sở các thuộc tính khảo sát.

PGS. TS. Nguyễn Thông

PHÂN TÍCH DỮ LIỆU

Chương 5: Phân nhóm dữ liệu (Cluster)

- Nghiên cứu một tập thể cá nhân theo các yếu tố : hoạt động, thu nhập, quan điểm.
 - Các cá nhân xuất hiện, trong cùng một nhóm, thể hiện đó là các cá nhân có **tổng hợp các yếu tố là gần nhau nhất**. Nó cho phép chúng ta hiểu được các cung cách sống khác nhau.
 - Mục đích của phương pháp phân nhóm là nhận ra các nhóm **đồng nhất** trong tập dữ liệu các cá nhân đang xét cùng với các biến liên quan (định tính, định lượng).

PGS. TS. Nguyễn Thông

PHÂN TÍCH DỮ LIỆU

Chương 5: Phân nhóm dữ liệu (Cluster)

LÝ THUYẾT PHÂN NHÓM

Cũng như trong bất kỳ một phương pháp phân tích thống kê nào, một số công tác chuẩn bị phải được thực hiện trước. Các vấn đề cơ bản đặt ra và phải quyết định như sau :

- Các biến (chỉ tiêu) nào sẽ được sử dụng trong việc thành lập nhóm ?
- Định nghĩa về « khoảng cách » giữa các biến được sử dụng.

PGS. TS. Nguyễn Thông

PHÂN TÍCH DỮ LIỆU

Chương 5: Phân nhóm dữ liệu (Cluster)

- Tiêu chuẩn gì sẽ được sử dụng để nhóm các « phần tử» hoặc các nhóm lại với nhau.
- Trong đó việc chọn các biến để thành lập nhóm có một vai trò quan trọng nhất.
 - Việc xác định các chỉ tiêu để tiến hành thành lập nhóm có một vai trò quyết định đến kết quả phân nhóm.

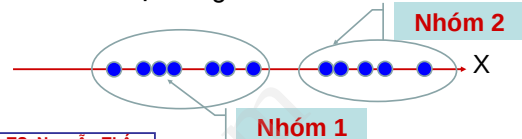
PGS. TS. Nguyễn Thống

PHÂN TÍCH DỮ LIỆU

Chương 5: Phân nhóm dữ liệu (Cluster)

PHÂN NHÓM DỰA TRÊN 1 BIẾN (BÀI TOÁN 1 CHIỀU)

- Một loại thí nghiệm VLXD cho kết quả chỉ tiêu khảo sát X. Giả sử ta muốn phân loại X thành 2 nhóm dựa vào giá trị. Một cách định tính ta có thể có được lời giải sau :

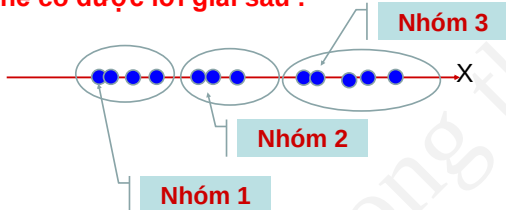


PGS. TS. Nguyễn Thống

PHÂN TÍCH DỮ LIỆU

Chương 5: Phân nhóm dữ liệu (Cluster)

Trường hợp muốn phân loại X thành 3 nhóm dựa vào giá trị. Một cách định tính ta có thể có được lời giải sau :



PGS. TS. Nguyễn Thống

PHÂN TÍCH DỮ LIỆU

Chương 5: Phân nhóm dữ liệu (Cluster)

PHÂN NHÓM DỰA TRÊN 2 BIẾN (BÀI TOÁN 2 CHIỀU)

Ví dụ: Một Tổng Công ty hoạt động có 14 Công ty trực thuộc. Để thực hiện chủ trương của Ban Giám Đốc về việc thành lập các Công ty có quy mô « lớn » hơn so với hiện tại để có thể cạnh tranh với các doanh nghiệp cùng ngành → Tổng Công ty sẽ tiến hành sáp nhập một số Công ty trực thuộc.

PGS. TS. Nguyễn Thống

PHÂN TÍCH DỮ LIỆU

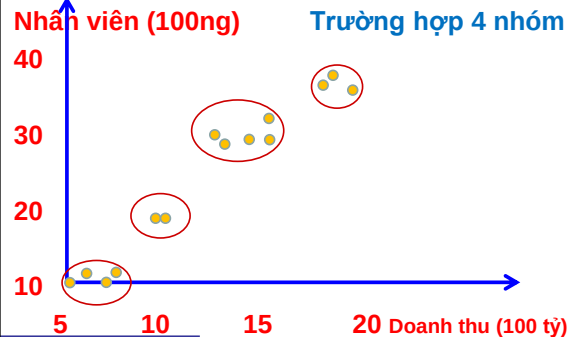
Chương 5: Phân nhóm dữ liệu (Cluster)

Tiêu chí sáp nhập đưa ra được dựa vào doanh thu X_1 và số lượng nhân sự X_2 . Để có thể tiến hành sắp xếp lại các Công ty trực thuộc, Tổng Công ty cần phân nhóm các Công ty trực thuộc có doanh thu và nhân sự « gần giống » nhau → Về mặt đồ thị ta có thể biểu diễn 14 Công ty dưới dạng sau :

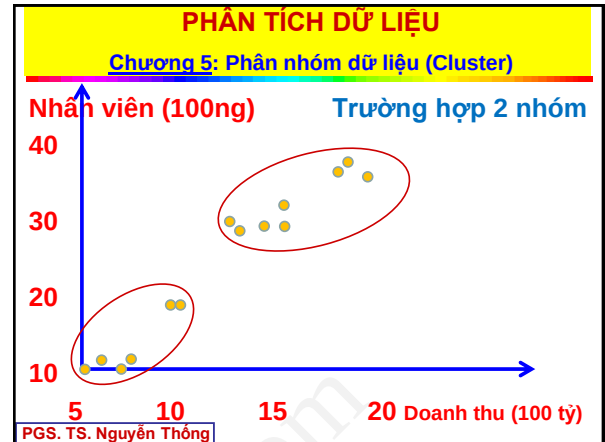
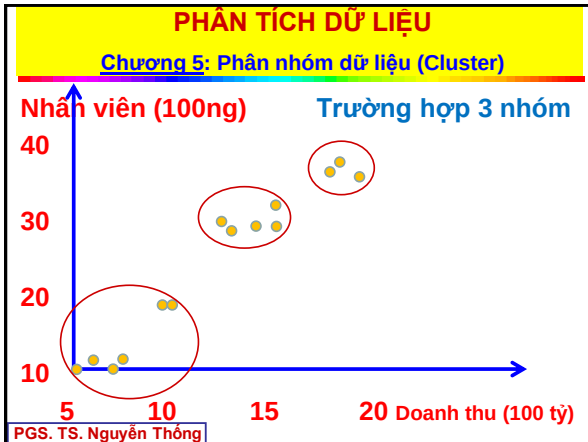
PGS. TS. Nguyễn Thống

PHÂN TÍCH DỮ LIỆU

Chương 5: Phân nhóm dữ liệu (Cluster)



PGS. TS. Nguyễn Thống



PHÂN TÍCH DỮ LIỆU
Chương 5: Phân nhóm dữ liệu (Cluster)

Ghi chú: Ta có thể quan niệm đây là trường hợp có 14 nhóm, trong đó mỗi phần tử (Công ty) hình thành 1 nhóm.

Bài toán tương tự sẽ được tổng quát hoá trong trường hợp vấn đề nhiều biến (n biến) → Trong không gian n chiều.

→

PGS. TS. Nguyễn Thông

PHÂN TÍCH DỮ LIỆU
Chương 5: Phân nhóm dữ liệu (Cluster)

→ Mỗi cá nhân nghiên cứu sẽ được « định vị » trong không gian n chiều nhờ vào n tọa độ là giá trị các biến tương ứng (sẽ nói rõ hơn trong chủ đề: Phân tích nhân tố các thành phần chính).

PGS. TS. Nguyễn Thông

PHÂN TÍCH DỮ LIỆU
Chương 5: Phân nhóm dữ liệu (Cluster)

TIÊU CHUẨN « KẾT HỢP » THÀNH NHÓM

Các ví dụ trên chỉ cho chúng ta một cách nhìn có tính cách trực giác trong việc thành lập nhóm. Về mặt lý thuyết, để tiến hành kết hợp 2 hay nhiều « cá thể » vào trong một nhóm ta sẽ dựa vào các khái niệm cơ bản trình bày sau.

PGS. TS. Nguyễn Thông

PHÂN TÍCH DỮ LIỆU
Chương 5: Phân nhóm dữ liệu (Cluster)

→ Hai cá thể (hình thành một tiêu nhóm) sẽ được nhóm lại khi nó có các thuộc tính « gần gũi » nhất, so với các cá thể (nhóm) còn lại.

→ Để đánh giá sự gần gũi của 2 cá thể, ta dựa vào các tiêu chuẩn phổ biến sau đây :

PGS. TS. Nguyễn Thông

PHÂN TÍCH DỮ LIỆU

Chương 5: Phân nhóm dữ liệu (Cluster)

Khoảng cách Euclid : Xét trong không gian n chiều, hai cá thể A và B sẽ được xác định « vị trí » trong không gian này nhờ vào các giá trị (tọa độ) tương ứng : $A(x_1^A, x_2^A, \dots, x_n^A)$ và $B(x_1^B, x_2^B, \dots, x_n^B)$.

Khoảng cách Euclid giữa A & B định nghĩa bởi:

$$D(A, B) = \sqrt{\sum_i (x_i^A - x_i^B)^2}$$

PGS. TS. Nguyễn Thống

PHÂN TÍCH DỮ LIỆU

Chương 5: Phân nhóm dữ liệu (Cluster)

- **Khoảng cách Chebychev :** Được định nghĩa như giá trị tuyệt đối lớn nhất của các sai biệt của A và B :

$$D(A, B) = \max_i |x_i^A - x_i^B|$$

PGS. TS. Nguyễn Thống

PHÂN TÍCH DỮ LIỆU

Chương 5: Phân nhóm dữ liệu (Cluster)

- Trong trường hợp số liệu ở dạng tần số « frequency », « khoảng cách » được đo thông qua sự đánh giá « độc lập » giữa 2 biến, dựa vào định nghĩa χ^2 và được định nghĩa như sau :

$$D(A, B) = \sqrt{\sum_i \frac{[x_i^A - E(x_i^A)]^2}{E(x_i^A)} + \sum_i \frac{[x_i^B - E(x_i^B)]^2}{E(x_i^B)}}$$

PGS. TS. Nguyễn Thống

PHÂN TÍCH DỮ LIỆU

Chương 5: Phân nhóm dữ liệu (Cluster)

SỰ TƯƠNG TỰ

Để có thể « nhóm » các cá thể, ngoài khái niệm dựa vào sự « gần » nhau theo thông số khoảng cách Euclid (hoặc định nghĩa khoảng cách khác) như trên.



PGS. TS. Nguyễn Thống

PHÂN TÍCH DỮ LIỆU

Chương 5: Phân nhóm dữ liệu (Cluster)

SỰ TƯƠNG TỰ

→ Đôi khi người ta còn dựa vào tiêu chuẩn sự tương tự bởi tiêu chuẩn sau đây :

→ Tiêu chuẩn cosin:

$$S(A, B) = \frac{\sum_i x_i^A x_i^B}{\sqrt{\sum_i (x_i^A)^2} \cdot \sqrt{\sum_i (x_i^B)^2}}$$

PGS. TS. Nguyễn Thống

PHÂN TÍCH DỮ LIỆU

Chương 5: Phân nhóm dữ liệu (Cluster)

- Tiêu chuẩn tương quan theo Pearson :

$$S(A, B) = \frac{\sum_{i=1}^N X_i Y_i}{N-1}$$

- Với X_i , Y_i là biến x_i & y_i được trung tâm hóa & chuẩn hóa

PGS. TS. Nguyễn Thống

PHÂN TÍCH DỮ LIỆU

Chương 5: Phân nhóm dữ liệu (Cluster)

Chú ý : Tính khoảng cách trong trường hợp cho 2 nhóm, mà mỗi nhóm có chứa hơn 1 phần tử. Lúc đó ta sẽ có 3 định nghĩa khác nhau về khoảng cách giữa 2 nhóm :

- Khoảng cách được định nghĩa là giá trị khoảng cách trung bình còn gọi là khoảng cách có tính trọng số (khoảng cách giữa 2 tâm trọng trường của 2 nhóm)

PGS. TS. Nguyễn Thống

PHÂN TÍCH DỮ LIỆU

Chương 5: Phân nhóm dữ liệu (Cluster)

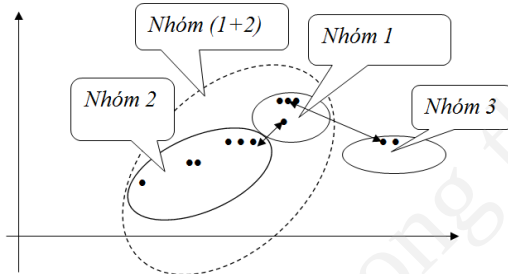
- Khoảng cách được định nghĩa là khoảng cách min của 2 phần tử thuộc 2 nhóm.
 - Khoảng cách được định nghĩa là khoảng cách max của 2 phần tử thuộc 2 nhóm.
- Tùy theo sự lựa chọn 1 trong 3 cách tính khoảng cách nêu trên, có thể sự sát nhập nhóm dựa vào khoảng cách sẽ cho kết quả không giống nhau.

PGS. TS. Nguyễn Thống

PHÂN TÍCH DỮ LIỆU

Chương 5: Phân nhóm dữ liệu (Cluster)

Ví dụ: Phân nhóm theo tiêu chí min



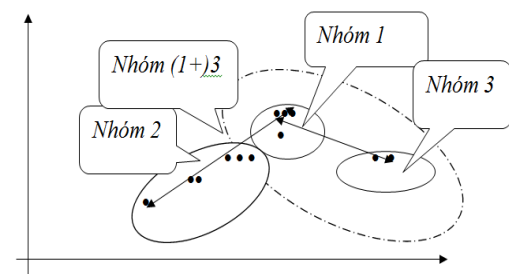
Hình thành nhóm theo tiêu chuẩn khoảng cách min

PGS. TS. Nguyễn Thống

PHÂN TÍCH DỮ LIỆU

Chương 5: Phân nhóm dữ liệu (Cluster)

Ví dụ: Phân nhóm theo tiêu chí max



Trường hợp gộp nhóm theo tiêu chuẩn max khoảng cách

PGS. TS. Nguyễn Thống

PHÂN TÍCH DỮ LIỆU

Chương 5: Phân nhóm dữ liệu (Cluster)

PHƯƠNG PHÁP « KẾT HỢP » THÀNH NHÓM

- **Phương pháp 1: Từ dưới lên (Ascendant).** Đây là phương pháp phổ biến :
 - Tập hợp gồm N cá thể cần nhóm lại với nhau để thành M nhóm (<N).
 - Ta tiến hành thực hiện phép nhóm đầu tiên cho 2 cá thể « gần » nhau nhất (ví dụ theo tiêu chuẩn khoảng cách Euclid nhỏ nhất) và như vậy ta còn có N-1 nhóm.

PGS. TS. Nguyễn Thống

PHÂN TÍCH DỮ LIỆU

Chương 5: Phân nhóm dữ liệu (Cluster)

- Và nhóm mới thành lập này sẽ thay thế 2 cá thể vừa được nhóm lại, được hiểu là một « cá thể ».
 - Công việc tiếp tục và ta sẽ lần lượt có số số nhóm giảm dần sau mỗi lần nhóm lại.
- Note:** Dĩ nhiên lần kết hợp thứ (N-1) sẽ cho ta còn 1 nhóm duy nhất gồm N phần tử.

PGS. TS. Nguyễn Thống

PHÂN TÍCH DỮ LIỆU

Chương 5: Phân nhóm dữ liệu (Cluster)

• Phương pháp 2: Từ trên xuống (Descendant)

→ Đây là một quy trình ngược lại. Ban đầu ta xem N cá thể chỉ trong duy nhất 1 nhóm, sau đó ta tiến hành tách ra để thành lập 2 nhóm, ví dụ dựa theo tiêu chuẩn khoảng cách Euclid « lớn nhất ».

→ Quy trình tiếp tục trong logic như vậy cho đến khi ta có được N nhóm với mỗi nhóm là một phần tử của số liệu ban đầu.

PGS. TS. Nguyễn Thông

PHÂN TÍCH DỮ LIỆU

Chương 5: Phân nhóm dữ liệu (Cluster)

CHÚ Ý

→ Giả thiết số liệu 1D, khi gộp còn 2 nhóm → Khoảng cách 2 nhóm là « xa nhất » (tách biệt rõ nhất).

→

PGS. TS. Nguyễn Thông

PHÂN TÍCH DỮ LIỆU

Chương 5: Phân nhóm dữ liệu (Cluster)

Bài tập 0: Cho số liệu sau:

Cá nhân Biến	a	b	c	d	e
x1	8	2	6	5	6
x2	10	8	3	2	9

Dùng tiêu chí để nhóm là khoảng cách Euclid.

- Tính ma trận khoảng cách
- Thực hiện tạo nhóm

PGS. TS. Nguyễn Thông

PHÂN TÍCH DỮ LIỆU

Chương 5: Phân nhóm dữ liệu (Cluster)

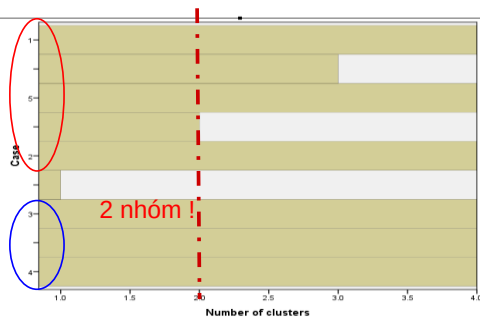
Ma trận khoảng cách Euclid:

	Proximity Matrix				
	Euclidean Distance				
Case	a	b	c	d	e
a	0.000	6.325	7.280	8.544	2.236
b	6.325	0.000	6.403	6.708	4.123
c	7.280	6.403	0.000	1.414	6.000
d	8.544	6.708	1.414	0.000	7.071
e	2.236	4.123	6.000	7.071	0.000

PGS. TS. Nguyễn Thông

PHÂN TÍCH DỮ LIỆU

Chương 5: Phân nhóm dữ liệu (Cluster)



PGS. TS. Nguyễn Thông

PHÂN TÍCH DỮ LIỆU

Chương 5: Phân nhóm dữ liệu (Cluster)

→ (c+d) → ma trận distance mới

→ (a+e) → ma trận distance mới

→ (ae+b) → ma trận distance mới

Cá nhân		
Cá nhân	(aeb)	(cd)
(aeb)	0	6
(cd)		0

PGS. TS. Nguyễn Thông

PHÂN TÍCH DỮ LIỆU

Chương 5: Phân nhóm dữ liệu (Cluster)

Bài tập 0bis: Một thí nghiệm nén R(kg/cm²) cho 6 mẫu bê tông cho kết quả như sau. Hãy phân dữ liệu trên thành 2 nhóm theo phương pháp min khoảng cách Euclid của trọng tâm nhóm.

Mẫu	1	2	3	4	5	6
R	80	85	100	76	82	95

PGS. TS. Nguyễn Thông

PHÂN TÍCH DỮ LIỆU

Chương 5: Phân nhóm dữ liệu (Cluster)

Bài tập 1: Một thí nghiệm nén R(kg/cm²) cho 10 mẫu bê tông cho kết quả như sau. Hãy phân dữ liệu trên thành 3 & 2 nhóm theo phương pháp min khoảng cách Euclid của trọng tâm nhóm.

Mẫu	1	2	3	4	5	6	7	8	9	10
R	80	85	100	76	82	95	90	98	86	102

Số liệu: SPSS → ../PTDuLieu/Cluster_1D.sav

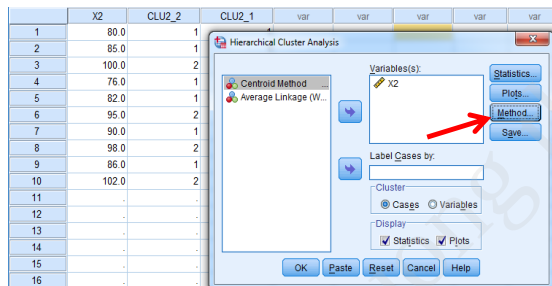
Giải Excel → ../PTDuLieu/Cluster_1D.xls

PGS. TS. Nguyễn Thông

PHÂN TÍCH DỮ LIỆU

Chương 5: Phân nhóm dữ liệu (Cluster)

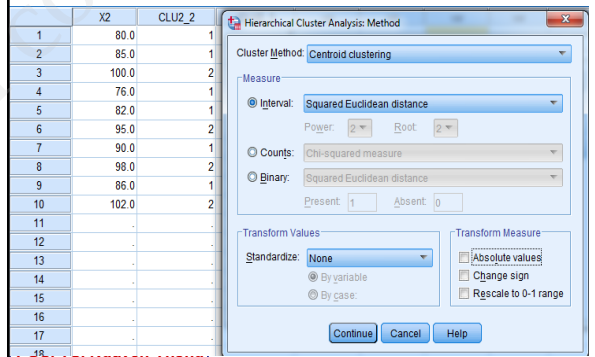
Hướng dẫn: SPSS\Analyze\Classify.



PGS. TS. Nguyễn Thông

PHÂN TÍCH DỮ LIỆU

Chương 5: Phân nhóm dữ liệu (Cluster)



PHÂN TÍCH DỮ LIỆU

Chương 5: Phân nhóm dữ liệu (Cluster)

Case	Squared Euclidean Distance									
	1	2	3	4	5	6	7	8	9	10
1	.000	25.000	400.000	16.000	4.000	225.000	100.000	324.000	36.000	484.000
2	25.000	.000	225.000	81.000	9.000	100.000	25.000	169.000	1.000	289.000
3	400.000	225.000	.000	576.000	324.000	25.000	100.000	4.000	196.000	4.000
4	16.000	81.000	576.000	.000	36.000	361.000	196.000	484.000	100.000	676.000
5	4.000	9.000	324.000	36.000	.000	169.000	64.000	256.000	16.000	400.000
6	225.000	100.000	25.000	361.000	169.000	.000	25.000	9.000	81.000	49.000
7	100.000	25.000	100.000	196.000	64.000	25.000	.000	64.000	16.000	144.000
8	324.000	169.000	4.000	484.000	256.000	9.000	64.000	.000	144.000	16.000
9	36.000	1.000	196.000	100.000	16.000	81.000	16.000	144.000	.000	256.000
10	484.000	289.000	4.000	676.000	400.000	49.000	144.000	16.000	256.000	.000

PGS. TS. Nguyễn Thông

PHÂN TÍCH DỮ LIỆU

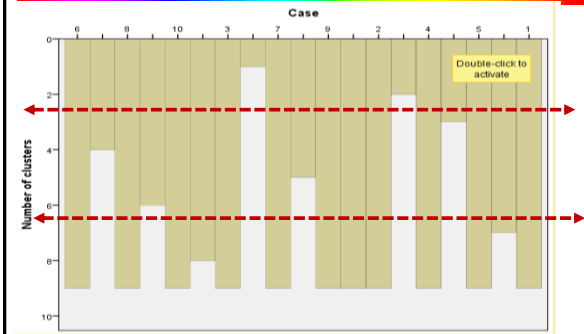
Chương 5: Phân nhóm dữ liệu (Cluster)

Cluster Membership	
Case	2 Clusters
1	1
2	1
3	2
4	1
5	1
6	2
7	1
8	2
9	1
10	2

PGS. TS. Nguyễn Thông

PHÂN TÍCH DỮ LIỆU

Chương 5: Phân nhóm dữ liệu (Cluster)



PGS. TS. Nguyễn Thông

PHÂN TÍCH DỮ LIỆU

Chương 5: Phân nhóm dữ liệu (Cluster)

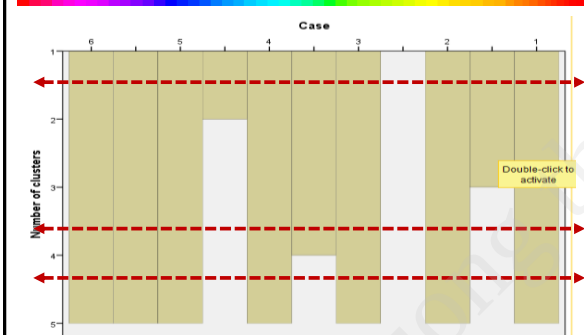
Bài tập 2: Thí nghiệm 6 mẫu vật liệu về lực kéo (R) & cắt (T) cho kết quả như sau. Người ta muốn chia 6 mẫu này thành 2 nhóm theo phương pháp tính khoảng cách Euclid (giả thiết không kể đến đơn vị).

Mẫu	1	2	3	4	5	6
R	2.9	3.0	3.2	3.3	3.7	3.8
T	1.8	1.9	2.4	2.5	2.6	2.7

PGS. TS. Nguyễn Thông

PHÂN TÍCH DỮ LIỆU

Chương 5: Phân nhóm dữ liệu (Cluster)



PGS. TS. Nguyễn Thông

PHÂN TÍCH DỮ LIỆU

Chương 5: Phân nhóm dữ liệu (Cluster)

Hướng dẫn:

Cluster Membership	
Case	2 Clusters
1	1
2	1
3	2
4	2
5	2
6	2

PGS. TS. Nguyễn Thông

PHÂN TÍCH DỮ LIỆU

Chương 5: Phân nhóm dữ liệu (Cluster)

Bài tập 3: Thí nghiệm 5 mẫu vật liệu về lực kéo (R) & cắt (T) cho kết quả như sau. Người ta muốn chia 5 mẫu này thành 2 nhóm theo phương pháp tính khoảng cách Euclid theo trọng tâm (giả thiết không kể đến đơn vị).

Mẫu	1	2	3	4	5
R	2.9	3.0	3.2	3.8	4.0
T	1.8	2.6	1.2	2.5	1.8

PGS. TS. Nguyễn Thông

PHÂN TÍCH DỮ LIỆU

Chương 5: Phân nhóm dữ liệu (Cluster)

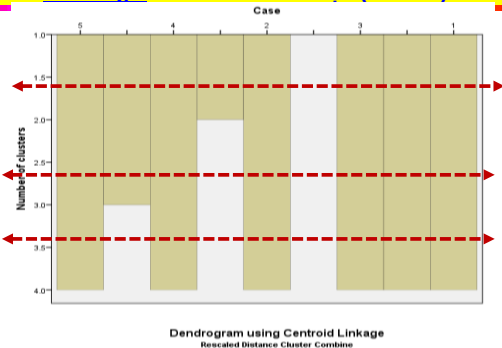
Hướng dẫn:

Cluster Membership	
Case	2 Clusters
1	1
2	2
3	1
4	2
5	2

PGS. TS. Nguyễn Thông

PHÂN TÍCH DỮ LIỆU

Chương 5: Phân nhóm dữ liệu (Cluster)



PGS. TS. Nguyễn Thông

THỐNG KÊ ỨNG DỤNG TRONG QUẢN LÝ & KỸ THUẬT

Chương 1: Phân phối thống kê & xác suất

HẾT CHƯƠNG

PGS. TS. Nguyễn Thông