TRƯỜNG ĐẠI HỌC BÁCH KHOA HÀ NỘI
VIỆN TOÁN ỨNG DỤNG VÀ TIN HỌC

PHÂN CỤM DỮ LIỆU

Môn: Phân tích số liệu

Giảng viên hướng dẫn: **ThS. Lê Xuân Lý**

Mã lớp: 137946 **Học kỳ 20221**

Nhóm sinh viên thực hiện: **Nhóm 5**

Hà Nội, 02/2023

Nhận xét của giảng viên

Nhận xét của giảng viên	Điểm

Hà Nội, ngày ..... tháng ..... năm 2023

Giảng viên hướng dẫn

ThS. Lê Xuân Lý

Đánh giá thành viên nhóm

Tên thành viên	MSSV	Phân công	Đánh giá
Nguyễn Hữu An	20195836	1., 2.1	Tích cực
Đào Bảo Đại	20200126	5. + chạy dữ liệu	Rất tích cực
Hà Quang Hanh	20195870	6. (làm chung)	Tích cực
Nguyễn Lan Hương	20195884	3.1	Tích cực
Thân Thị Mỹ Huyền	20206288	4. + chạy dữ liệu	Rất tích cực
Vũ Ánh Nguyệt	20195905	2.2, 2.3	Tích cực
Trần Văn Quang	20195911	6. (làm chung)	Tích cực
Nguyễn Hương Quỳnh	20206300	3.2 + chạy dữ liệu	Rất tích cực
Nguyễn Cẩm Tú	20206267	3.3	Tích cực

Lời mở đầu

Cùng với sự phát triển như vũ bão và không ngừng nghỉ của khoa học và công nghệ, các hệ thống dữ liệu dùng để phục vụ cho các lĩnh vực kinh tế, tự nhiên và xã hội cũng không ngừng tăng lên. Sự phong phú dữ liệu có thể giúp những người quản lý dễ dàng hơn trong việc ra quyết định. Tuy vậy, khi lượng dữ liệu thu thập được ngày càng nhiều, các quyết định cũng đòi hỏi khắt khe hơn, người quản lý không những cần dữ liệu mà còn cần có thêm các tri thức hỗ trợ nhằm khai thác hiệu quả tập dữ liệu đang có. Chính vì thế, phân tích dữ liệu và số liệu ra đời, như một giải pháp hiệu quả để giải quyết những bài toán dữ liệu lớn và phức tạp.

Khai phá dữ liệu là một tập con trong phân tích dữ liệu, bao gồm các quá trình như phân loại và sắp xếp lại các tập dữ liệu lớn để thiết lập các mối liên hệ nhằm giải quyết các vấn đề nhờ phân tích dữ liệu. Được ra đời từ những năm 90 của thế kỉ trước, khai phá dữ liệu đã nhanh chóng trở thành một hướng nghiên cứu phổ biến và quan trọng trong lĩnh vực khoa học máy tính và khoa học tri thức. Nhiều kết quả nghiên cứu của khai phá dữ liệu đã ứng dụng thành công trong các lĩnh vực khoa học, kinh tế, xã hội. Khai phá dữ liệu bao gồm nhiều hướng nghiên cứu quan trọng, và một trong số đó là phân tích cụm (Clustering). Phân tích cụm là quá trình tìm kiếm và phát hiện ra các cụm hay các mẫu tự nhiên trong cơ sở dữ liệu lớn. Các kỹ thuật chính được áp dụng trong phân tích cụm phần lớn được kế thừa từ lĩnh vực thống kê số liệu trong các vấn đề tài chính, địa lý, sinh học, nhận dạng ảnh, ... Trong tài liệu này, chúng ta sẽ tìm hiểu kĩ hơn một số khái niệm trong phân tích cụm, cũng như một số phương pháp thường được dùng trong phân tích cụm.

Nội dung chính của tài liệu bao gồm 6 phần:

1. Giới thiệu.
2. Khoảng cách và độ đo tương tự.
3. Phương pháp phân cụm theo thứ bậc.
4. Phương pháp phân cụm không theo thứ bậc.
5. Phương pháp phân cụm dựa trên mô hình thống kê.
6. Thuật toán chia tỷ lệ đa chiều.

Xin trân trọng cảm ơn thầy Lê Xuân Lý đã hỗ trợ và có những ý kiến đóng góp để giúp nhóm 5 hoàn thành báo cáo này.

Mục lục

Chương 1

Giới thiệu

Phân tích cụm là sự phân chia một cơ sở dữ liệu lớn thành các nhóm mà trong đó, các đối tượng dữ liệu ở cùng một nhóm thì tương tự nhau. Ý tưởng khi thực hiện phân tích cụm là tìm kiếm sự tương tự giữa các đối tượng dữ liệu nhờ phân tích các đặc điểm và tính chất ẩn chứa bên trong tập dữ liệu. Phân tích cụm thường được dùng trong các thuật toán khai phá dữ liệu để phân loại và mô tả tính chất tập dữ liệu. Mục tiêu của phân tích cụm là xác định các nhóm tự nhiên tồn tại bên trong tập dữ liệu sao cho các nhóm này thỏa mãn 2 điều kiện sau:

i, Độ tương tự giữa các đối tượng dữ liệu trong cùng một nhóm là lớn nhất có thể, ta gọi đây là tiêu chuẩn liên kết trong phân tích cụm.

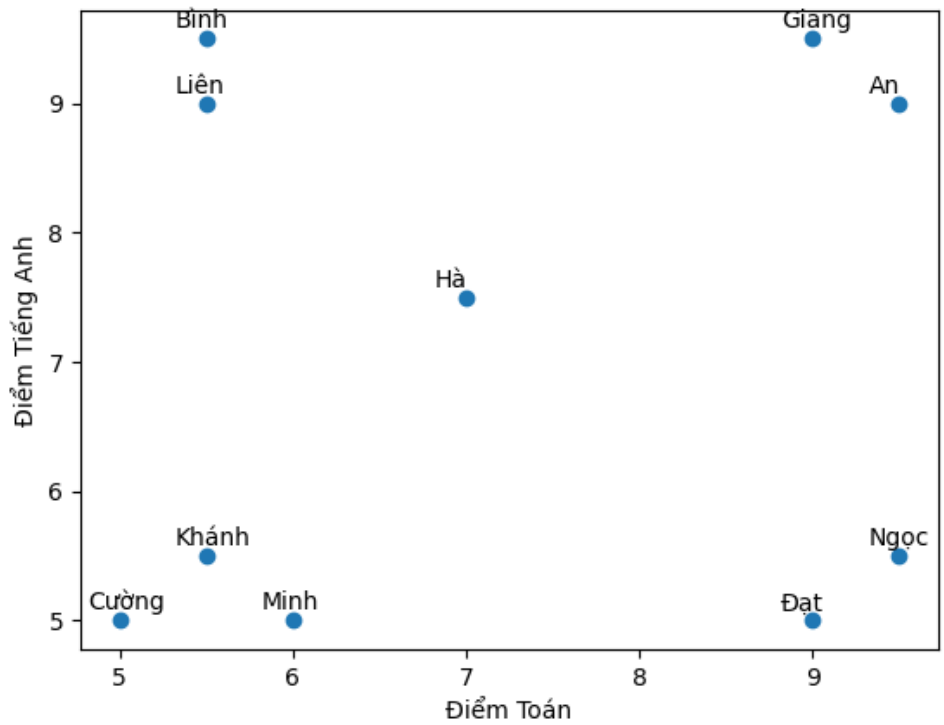
ii, Độ tương tự giữa các đối tượng ở các nhóm khau là thấp nhất có thể, đây là tiêu chuẩn tách rời trong phân tích cụm.

Để minh họa về phân tích cụm, cùng xem xét một ví dụ đơn giản trong thực tế dưới đây.

Ví dụ 1.0.1 Có 10 học sinh và bảng điểm thi giữa kì 2 môn Toán và Tiếng Anh của các học sinh đó. Giáo viên chủ nhiệm của muốn phân loại học lực để xếp lớp ôn thi cuối kỳ cho các học sinh. Vậy nên phân loại học sinh như thế nào?

Tên	Điểm Toán	Điểm Tiếng Anh
An	9.5	9.0
Bình	5.5	9.5
Cường	5.0	5.0
Đạt	9.0	5.0
Giang	9.0	9.5
Hà	7.0	7.5
Khánh	5.5	5.5
Liên	5.5	9.0
Minh	6.0	5.0
Ngọc	9.5	5.5

Trực quan hóa dữ liệu bằng cách sử dụng biểu đồ 2 trục, một trục là điểm Toán, trục còn lại là điểm Tiếng Anh, ta thu được:



Từ biểu đồ trên, có thể thấy phân chia 10 bạn học sinh thành bốn nhóm. Nhóm thứ nhất gồm những bạn giỏi ở cả 2 môn (An và Giang), nhóm thứ hai gồm những bạn giỏi Toán (Bình và Liên), nhóm thứ ba gồm những bạn giỏi Tiếng Anh (Đạt và Ngọc) và nhóm cuối cùng là những bạn học kém ở

cả 2 môn (Cường, Khánh, Minh). Những bạn học sinh ở cùng một nhóm thì ở rất gần nhau trên biểu đồ. Riêng trường hợp của bạn Hà, học bình thường ở cả hai môn, giáo viên chủ nhiệm có thể xếp riêng những bạn như Hà thành một lớp, hoặc cũng có thể cho Hà vào một trong bất kì 4 nhóm kể trên.

Đây chỉ là một ví dụ rất đơn giản về phân tích cụm và ta có thể nhìn thấy các cụm bằng mắt thường. Tuy nhiên trên thực tế, việc chỉ ra các cụm thường phức tạp và không dễ hình dung. Đối với tập các dữ liệu lớn nhiều chiều, ta dùng các thuật toán phân tích cụm để xử lí. Một thuật toán phân tích cụm được gọi là tốt khi nó đạt được các yêu cầu cơ bản như sau:

1. Có khả năng mở rộng quy mô, tức là khi số lượng đối tượng dữ liệu trong tập dữ liệu tăng lên, hiệu quả của thuật toán phân tích cụm không giảm đi đáng kể.

2. Tương thích với các kiểu thuộc tính khác nhau: Một số thuật toán phân cụm chỉ hiệu quả đối với dữ liệu dạng số và có bản chất định lượng. Tuy nhiên trên thực tế, dữ liệu có thể chia ra theo nhiều kiểu như định tính, định lượng, ...

3. Đa dạng về hình dáng cụm: Các thuật toán phân cụm dựa vào khoảng cách hình học thường cho ra những hình dáng cụm hạn chế như hình cầu hoặc gần cầu. Do vậy khi phân cụm, cần phân tích nhiều đặc điểm khác ngoài khoảng cách giữa các đối tượng dữ liệu.

4. Tối thiểu lượng tri thức cần để xác định các tham số đầu vào: Khi các tham số đầu vào càng đơn giản để xác định, thuật toán càng linh hoạt để triển khai.

5. Khả năng thích nghi với dữ liệu nhiễu: Trong quá trình thu thập dữ liệu, rất khó tránh khỏi sai sót hoặc thiếu. Nhiều thuật toán phân tích cụm thường có một bước bắt buộc trước khi chạy, đó là tiền xử lí dữ liệu, loại bỏ nhiễu.

6. Ít nhạy cảm với thứ tự dữ liệu đầu vào: Kết quả cuối cùng khi thay đổi thứ tự dữ liệu đầu vào là giống nhau hoặc hiệu quả như nhau.

7. Có thể dùng được với tập dữ liệu có số chiều lớn.

8. Dễ hiểu, dễ sử dụng và triển khai.

Có ba phương pháp chính được sử dụng trong phân tích cụm, đó là:

- Phân cụm theo thứ bậc: sẽ được trình bày ở chương 3. Phương pháp này phù hợp với tập dữ liệu vừa và nhỏ, kết quả không ổn định.

- Phân cụm không theo thứ bậc: sẽ được trình bày ở chương 4. Phương pháp này cải thiện một số nhược điểm của phương pháp phân cụm theo thứ bậc và cho kết quả ổn định hơn.
- Phân cụm dựa trên mô hình thống kê: sẽ được trình bày ở chương 5. Đây là phương pháp rất có ý nghĩa về mặt thống kê, khi nó cho phép chúng ta giải thích quan sát thu thập được.

Tuy vậy, dù là phân cụm theo bất kì phương pháp nào, thì các đối tượng dữ liệu ở các cụm khác nhau đều đánh đánh giá thông sự tương tự và bất tương tự. Không có định nghĩa duy nhất về sự tương tự hoặc bất tương tự giữa các đối tượng tự liệu. Các định nghĩa về tương tự phụ thuộc vào các đặc điểm, tính chất của tập dữ liệu cần khảo sát. Độ tương tự giữa 2 đối tượng dữ liệu thường được đánh giá gián tiếp thông qua một độ đo khoảng cách.

Định nghĩa 1.0.2 Khoảng cách giữa 2 đối tượng x, y thông qua độ đo D, kí hiệu là $d(x, y)$, thỏa mãn các tính chất sau:

- Tính không âm: $d(x, y) \geq 0, d(x, y) = 0 \Leftrightarrow x = y$.
- Tính đối xứng: $d(x, y) = d(y, x)$.
- Bất đẳng thức tam giác: $d(x, y) \neq d(x, z) + d(z, y)$.

Ở chương tiếp theo, ta sẽ tìm hiểu một số độ đo khoảng cách và độ đo tương tự giữa các đối tượng dữ liệu thường được dùng.

Chương 2

Khoảng cách và độ đo tương tự

2.1 Khoảng cách

Ma trận dữ liệu của n đối tượng dữ liệu với p thuộc tính:

$$\begin{bmatrix} a_{1,1} & a_{1,2} & \dots & a_{1,p} \\ a_{2,1} & a_{2,2} & \dots & a_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n,1} & a_{n,2} & \dots & a_{n,p} \end{bmatrix},$$

với $a_{i,j}$ là dữ liệu thuộc tính thứ j của phần tử thứ i.

Ví dụ 2.1.1 Viết ma trận dữ liệu điểm của 10 bạn học sinh ở ví dụ 1.0.1:

Ma trận dữ liệu ban đầu:

$$\begin{bmatrix} 9.5 & 9.0 \\ 5.5 & 9.5 \\ 5.0 & 5.0 \\ 9.0 & 5.0 \\ 9.0 & 9.5 \\ 7.0 & 7.5 \\ 5.5 & 5.5 \\ 5.5 & 9.0 \\ 6.0 & 5.0 \\ 9.5 & 5.5 \end{bmatrix}.$$

Chuẩn hóa ma trận dữ liệu, ta thu được:

$$\begin{bmatrix} 1.3132 & 1.0107 \\ -0.9220 & 1.2698 \\ -1.2014 & -1.0625 \\ 1.0338 & -1.0625 \\ 1.0338 & 1.2698 \\ -0.0838 & 0.2332 \\ -0.9220 & -0.8034 \\ -0.9220 & 1.0107 \\ -0.6426 & -1.0625 \\ 1.3132 & -0.8034 \end{bmatrix}.$$

Một số lí do mà ta nên chuẩn hóa dữ liệu trước khi thực hiện tính toán:

- Các biến có đơn vị đo khác nhau $\Rightarrow$ đưa về cùng một đơn vị.
- Các biến có biên độ giao động không giống nhau $\Rightarrow$ đưa về cùng biên độ giao động.

Ma trận khoảng cách của n đối tượng dữ liệu:

$$\begin{bmatrix} 0 & d(1,2) & \dots & d(1,n) \\ d(2,1) & 0 & \dots & d(2,n) \\ \vdots & \vdots & \ddots & \vdots \\ d(n,1) & d(n,2) & \dots & 0 \end{bmatrix},$$

với $d(i,j) = d(j,i)$ là khoảng cách giữa 2 đối tượng dữ liệu thứ i và thứ j.

2.1.1 Khoảng cách Euclidean

Định nghĩa 2.1.2 Khoảng cách Euclidean giữa 2 điểm dữ liệu x và y, được cho bởi công thức sau:

$$\begin{aligned} d(x,y) &= \sqrt{(x_1 - y_1)^2 + x_2 - y_2)^2 + \cdots + x_p - y_p)^2} \\ &= \sqrt{(x-y)^T(x-y)}, \end{aligned}$$

trong đó $x^T = (x_1, x_2, \dots, x_p)$ và $y^T = (y_1, y_2, \dots, y_p)$.

Khoảng cách Euclidean còn được gọi là khoảng cách hình học. Cách tính khoảng cách Euclidean khá gần gũi với cách tính khoảng cách vật lí trong cuộc sống thường ngày của chúng ta nên rất dễ sử dụng. Tuy vậy, khoảng

cách Euclidean coi mọi điểm trong không gian dữ liệu nhiều chiều có vai trò như nhau. Do vậy, người ta thường không dùng trực tiếp khoảng cách Euclidean trong thống kê do mỗi biến dữ liệu sẽ có sự biến thiên khác nhau. Thay vào đó, ta có một loại khoảng cách gọi là khoảng cách thống kê giữa 2 điểm dữ liệu.

Định nghĩa 2.1.3 Khoảng cách thống kê giữa 2 điểm dữ liệu x và y, được cho bởi công thức sau:

$$d(x,y) = \sqrt{(x-y)^T A(x-y)},$$

trong đó $A = S^{-1}, S$ là ma trận hiệp phương sai mẫu.

Định nghĩa 2.1.4 Ma trận hiệp phương sai mẫu S chứa các hiệp phương sai giữa các cặp biến trong tập quan sát, phần tử $S_{i,j}$ là hiệp hương sai giữa biến thứ i và biến thứ j. Công thức tính hiệp phương sai giữa 2 biến X và Y được cho bởi công thức:

$$cov(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}),$$

trong đó $\overline{X}, \overline{Y}$ lần lượt là giá trị kỳ vọng của X và Y trong tập quan sát. Các phần tử trên đường chéo chính của S các phương sai của các biến. Đối với những tập mẫu lớn, người ta thường hay dùng hiệp phương sai mẫu hiệu chỉnh. Hiệp phương sai mẫu hiệu chỉnh được đưa ra bởi công thức:

$$cov(X,Y) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}),$$

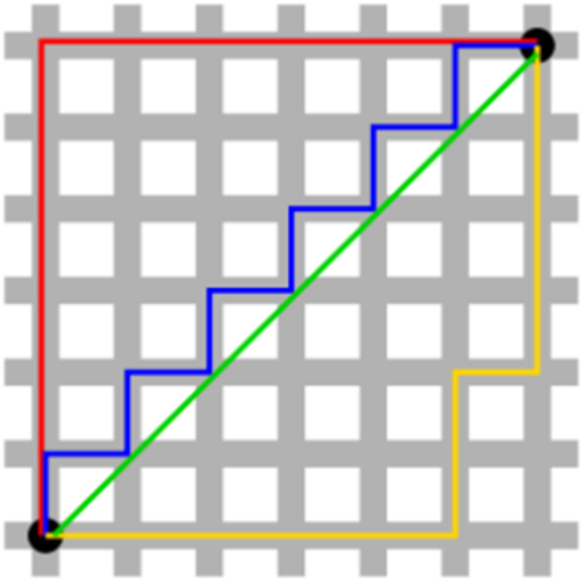
với n ở cả 2 công thức đều mang ý nghĩa là số lượng quan sát trong tập mẫu. Khi đó S sẽ trở thành ma trận hiệp phương sai mẫu hiệu chỉnh.

2.1.2 Khoảng cách Manhattan

Định nghĩa 2.1.5 Khoảng cách Manhattan giữa 2 điểm dữ liệu x và y, được cho bởi công thức sau:

$$d(x,y) = \sum_{i=1}^{p}|x_i - y_i|.$$

Để ý rằng, khoảng cách Manhattan bằng tổng độ dài hình chiếu của khoảng cách Euclidean lên các thành phần chính. Cùng xem hình dưới đây:



Đường màu xanh lá cây là khoảng cách Euclidean. Khoảng cách Manhattan bằng tổng độ dài của một trong ba đường còn lại. Cái tên Manhattan xuất phát từ thành phố Manhattan, nơi có nhiều nhà cao tầng được xây dựng liền kề nhau gần giống dạng lưới vuông. Khi đó người ta đã nghĩ ra việc tính khoảng cách khi đi lại giữa 2 toàn nhà trong mạng lưới. Có một biến thể khác của khoảng cách Manhattan, đó là Canberra Metric, được đưa ra bởi công thức sau:

$$d(x,y) = \sum_{i=1}^{p} \frac{|x_i - y_i|}{|x_i| + |y_i|}$$

Canberra Metric là phiên bản có trọng số của Manhattan. Canberra Metric thường được dùng cho tập dữ liệu gần gốc tọa độ do nó rất nhạy cảm với những giá trị gần bằng 0.

2.1.3 Khoảng cách Minskowski

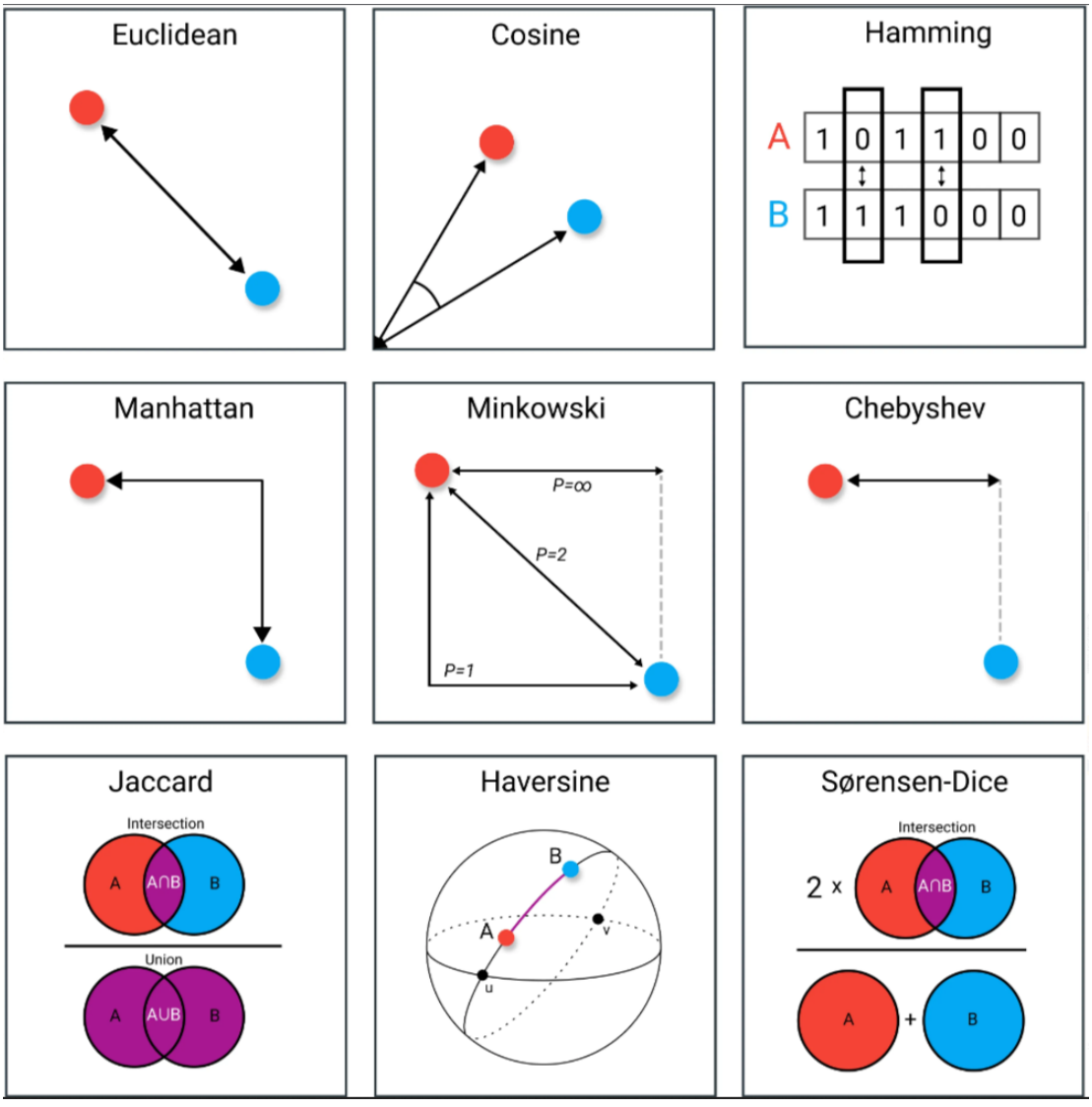
Cả 2 loại khoảng cách Euclidean và Manhattan đều là trường hợp của loại khoảng cách dưới đây.

Định nghĩa 2.1.6 Khoảng cách Minskowski giữa 2 điểm dữ liệu x và y, được cho bởi công thức sau:

$$d(x,y) = \left| \sum_{i=1}^{p} |x_i - y_i|^m \right|^{\frac{1}{m}},$$

với $m \geq 1$. Khoảng cách Minskowski khá linh hoạt do khi m được thay đổi có nghĩa là độ đo khoảng cách đã thay đổi. Mặc dù vậy, khoảng cách Minskowski thường kém hiệu quả về mặt tính toán. Khi $m = 2$, khoảng cách Minskowski trở thành khoảng cách Euclidean và là khoảng cách ngắn nhất giữa 2 điểm.

Ngoài các thước đo khoảng cách kể trên, còn có rất nhiều thước đo khoảng cách khác như Cosine, Hamming, Chebyshev, ... Chúng đều có những ứng dụng cho những bài toán cụ thể.



Ở phần tiếp theo, chúng ta sẽ đi tìm hiểu về hệ số tương tự và độ đo tương giữa các cặp điểm dữ liệu thông qua các phép đo khoảng cách.

2.2 Hệ số tương tự

Khi các mục không thể được biểu diễn bằng phép đo p chiều có ý nghĩa, các cặp mục thường được so sánh trên cơ sở có hoặc không có một vài đặc tính nhất định. Các cặp mục tương tự có nhiều đặc điểm chung hơn là các cặp mục không tương tự. Sự hiện diện hay vắng mặt của một đặc tính có thể được mô tả toán học bằng cách đưa vào một biến nhị phân, giả định giá trị biến là 1 nếu có đặc tính và giá trị biến là 0 nếu không có đặc tính.

Ví dụ 2.2.1 Có p = 5 biến nhị phân, "điểm" cho hai mục i và k có thể trông như sau:

	Varriables				
	1	2	3	4	5
Item i	0	1	1	0	0
Item k	1	1	0	0	1

Có 2 cặp tương đồng: 1 cặp (1, 1) và 1 cặp (0, 0); có 3 cặp không tương đồng. Ta có:

$$\left(x_{i_j} - x_{k_j}\right)^2 = \begin{cases} 0 & \text{nếu } x_{i_j} = x_{k_j} = 1 \text{ hoặc } x_{i_j} = x_{k_j} = 0 \\ 1 & \text{nếu } x_{i_j} \neq x_{k_j} \end{cases}$$

Khi đó khoảng cách Euclid bình phương $\sum_{j=1}^{p}(x_{i_j} - x_{k_j})^2$ sẽ cung cấp một đếm số không khớp, chính là số cặp mục không tương đồng:

$$\sum_{j=1}^{5}(x_{i_j} - x_{k_j})^2 = (0-1)^2 + (1-1)^2 + (1-0)^2 + (0-0)^2 + (0-1)^2 = 3.$$

Mặc dù khoảng cách dựa trên công thức này có thể được sử dụng để đo độ tương đồng, nhưng nó sẽ dẫn đến việc cân bằng trọng số các cặp 1-1 và 0-0. Trong một số trường hợp, 1-1 thể hiện dấu hiệu tương tự mạnh hơn so với 0-0. Để xử lý sự khác biệt giữa 1-1 và 0-0, một số phương án xác định hệ số tương tự được đưa ra. Ta sẽ sắp xếp tần số của các kết quả trùng khớp và khác nhau của các mục i và k dưới dạng một bảng dự phòng.

		Item k		
		1	0	Totals
Item i	1	a	b	a + b
	0	c	d	c + d
Totals		a + c	b + d	p = a + b+ c + d

Trong đó:

- a: tổng số thuộc tính mà cả i và k đều có (đều có giá trị là 1),
- b: tổng số thuộc tính mà i có, k không có (i có giá trị là 1, k có giá trị là 0),
- c: tổng số thuộc tính mà i không có, k có (i có giá trị là 0, k có giá trị là 1),
- d: tổng số thuộc tính mà cả i và k đều không có (đều có giá trị là 0).

Với cặp i, k ở ví dụ vừa rồi:

	Varriables				
	1	2	3	4	5
Item i	0	1	1	0	0
Item k	1	1	0	0	1

- a = 1
- b = 1
- c = 2
- d = 1

Ta có bảng liệt kê các hệ số tương tự phổ biến được xác định theo tần số trong bảng sau:

Bảng 1. Hệ số tương tự xác định theo tần số

Hệ số	Mô tả
1. $\frac{a+d}{p}$	Trọng số cặp (1, 1) và (0, 0) là như nhau
2. $\frac{2(a+d)}{2(a+d)+b+c}$	Gấp đôi trọng số cho cặp (1, 1) và (0, 0)
3. $\frac{a+d}{a+d+2(b+c)}$	Gấp đôi trọng số cho cặp (0, 1) và (1, 0)
4. $\frac{a}{p}$	Không có cặp (0, 0) trên tử số
5. $\frac{a}{a+b+c}$	Không có cặp (0, 0) trên tử hay mẫu số (Các cặp (0, 0) được coi là không liên quan)
6. $\frac{2a}{2a+b+c}$	Không có cặp (0, 0) trên tử hay mẫu số (Gấp đôi trọng số cho cặp (1, 1))
7. $\frac{a}{a+2(b+c)}$	Không có cặp (0, 0) trên tử hay mẫu số (Gấp đôi trọng số cho cặp (0, 1) và (1, 0))
8. $\frac{a}{b+c}$	Tỉ số giữa cặp tương đồng và không tương đồng (đã loại trừ cặp (0, 0))

STT	Giới tính	Chiều cao	Cân nặng	Màu tóc	Màu mắt
I1	Nữ	1.60m	50kg	Đen	Đen
I2	Nam	1.72m	61kg	Vàng	Đen
I3	Nữ	1.56m	47kg	Đen	Nâu
I4	Nam	1.67m	66kg	Đen	Đen
I5	Nam	1.75m	65kg	Đen	Đen

Ví dụ 2.2.2 Xác định 5 biến nhị phân X_1, X_2, X_3, X_4, X_5 như sau:

$$X_1 = \begin{cases} 1 & nam \\ 0 & nữ \end{cases} \qquad X_2 = \begin{cases} 1 & chiều\ cao \geq 1m7 \\ 0 & chiều\ cao < 1m7 \end{cases}$$

$$X_3 = \begin{cases} 1 & cân\ nặng \geq 60kg \\ 0 & cân\ nặng < 60kg \end{cases} \qquad X_4 = \begin{cases} 1 & tóc\ đen \\ 0 & còn\ lại \end{cases}$$

$$X_5 = \begin{cases} 1 & mắt\ đen \\ 0 & còn\ lại \end{cases}$$

Ta có bảng cho I1 và I2 với p = 5:

	X_1	X_2	X_3	X_4	X_5
I1	0	0	0	1	1
I2	1	1	1	0	1

Số lượng các kết quả trùng khớp và khác nhau được chỉ ra trong bảng dự phòng sau:

		I2		
		1	0	Total
I1	1	1	1	2
	0	3	0	3
	Total	4	1	5

Sử dụng hệ số tương tự 1 trong bảng 1:

$$\frac{a+d}{p} = \frac{1+0}{5} = 0.2.$$

Tiếp tục với hệ số tương tự 1, ta tính các hệ số còn lại cho các cặp. Cuối cùng ta thu được một ma trận đối xứng:

	I1	I2	I3	I4	I5
I1	1				
I2	0.2	1			
I3	0.8	0	1		
I4	0.6	0.6	0.4	1	
I5	0.4	0.8	0.2	0.8	1

Dựa vào ma trận có thể thấy:

- I2 và I3 có hệ số tương tự nhỏ nhất, vì vậy 2 người khác nhau nhất.
- Các cặp (I1, I3), (I2, I5), (I4, I5) là những cặp có hệ số tương tự cao nhất, vì vậy giống nhau nhất.

2.3 Độ đo tương tự và độ đo liên kết cho cặp điểm

Đến đây, ta đã thảo luận về các phương pháp tương tự cho các mục. Trong một số ứng dụng sẽ là các biến, thay vì các mục, các đối tượng phải được

nhóm lại. Các thước đo độ tương tự cho các biến thường có dạng hệ số tương quan mẫu. Trong một số ứng dụng phân cụm, hệ số tương quan âm được thay thế bằng giá trị tuyệt đối của chúng. Khi các biến là nhị phân, dữ liệu lại có thể được sắp xếp dưới dạng một bảng phụ. Tuy nhiên, lần này sẽ là các biến thay vì các mục, mô tả các danh mục. Đối với mỗi cặp biến, có n mục được phân loại trong bảng. Với mã hóa 0 và 1 thông thường, bảng sẽ trở thành như sau:

		Variable k		
		1	0	Totals
Variable i	1	a	b	a+b
	0	c	d	c+d
Totals		a+c	b+d	n=a+b+c+d

Công thức tương quan thông thường được áp dụng cho các biến nhị phân trong bảng dự phòng:

$$r = \frac{ad - bc}{[(a+b)(c+d)(a+c)(b+d)]^{\frac{1}{2}}}.$$

r có thể được coi là thước đo mức độ giống nhau giữa 2 biến. Hệ số tương quan trong này có liên quan đến thống kê khi bình phương ($r^2 = \frac{X^2}{n}$) để kiểm tra tính độc lập của 2 biến. Từ bảng dự phòng trên có thể phát triển các phép đo liên kết (hoặc độ tương tự) tương tự chính xác với các phép đo được liệt kê trong Bảng 1. Thay đổi duy nhất được yêu cầu là thay n (số mục) cho p (số biến).

Chương 3

Phương pháp phân cụm theo thứ bậc

Chúng ta hiếm khi có thể kiểm tra tất cả các khả năng nhóm, ngay cả với những máy tính lớn nhất và nhanh nhất. Vì vấn đề này, nhiều thuật toán phân cụm đã xuất hiện để tìm các cụm "hợp lý" mà không cần phải xem xét tất cả các cấu hình.

Các kỹ thuật phân cụm theo thứ bậc tiến hành bằng một loạt các hợp nhất liên tiếp hoặc một loạt các phân chia liên tiếp.

Phương pháp kết hợp cụm thứ bậc bắt đầu với các phần tử riêng lẻ. Do đó, ban đầu số cụm bằng số phần tử. Các phần tử giống nhau nhất được hợp nhất đầu tiên và các cụm này được hợp nhất theo sự tương đồng của chúng. Cuối cùng, khi độ tương đồng giảm đi, tất cả các phần tử được hợp nhất thành một cụm duy nhất.

Phương pháp phân chia cụm thứ bậc hoạt động theo hướng ngược lại. Một cụm ban đầu của các phần tử được chia thành hai cụm nhỏ sao cho các phần tử trong một cụm nhỏ ở "xa" các phần tử trong cụm kia. Các cụm nhỏ này sau đó được chia thành các cụm nhỏ khác nhau; quá trình tiếp tục cho đến khi số cụm bằng số phần tử - nghĩa là cho đến khi mỗi phần tử tạo thành một cụm.

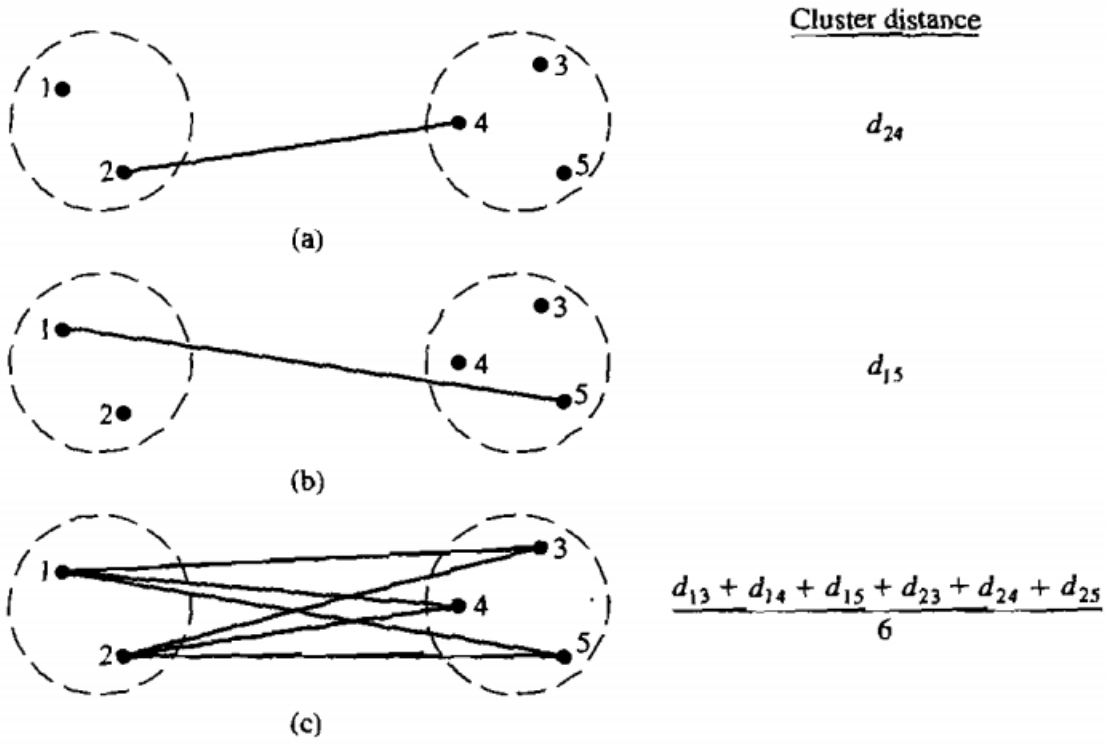
Kết quả của cả hai phương pháp kết hợp và phân chia có thể được hiển thị dưới dạng sơ đồ hai chiều Dendrogram. Dendrogram minh họa sự hợp nhất hoặc phân chia đã được thực hiện ở các cấp độ liên tiếp.

Trong phần này, chúng ta sẽ tập trung vào các phương pháp kết hợp cụm thứ bậc và đặc biệt là các phương pháp liên kết.

Các phương pháp liên kết thì phù hợp để phân cụm các biến. Điều này không đúng với tất cả các phương pháp kết hợp cụm thứ bậc. Ta sẽ lần lượt thảo luận về *liên kết đơn* (khoảng cách tối thiểu), *liên kết hoàn chỉnh* (khoảng cách tối đa) và *liên kết trung bình* (khoảng cách trung bình). Việc hợp nhất các cụm theo ba tiêu chí liên kết được minh họa bằng sơ đồ hình 3.1.

Từ hình vẽ, ta thấy rằng liên kết đơn xảy ra khi các cụm được hợp nhất theo khoảng cách giữa các phần tử gần nhất của chúng. Liên kết hoàn chỉnh xảy ra khi các cụm được hợp nhất theo khoảng cách giữa các phần tử xa nhất của chúng. Đối với liên kết trung bình, các cụm được hợp nhất theo khoảng cách trung bình giữa các cặp phần tử trong các cụm tương ứng.

Sau đây là các bước trong thuật toán kết hợp cụm thứ bậc để nhóm N đối tượng:
1. Bắt đầu với N cụm, mỗi cụm 1 phần tử và lập ma trận đối xứng NxN khoảng cách $D = \{d_{ik}\}$
2. Trên ma trận, tìm các cặp gần nhau nhất. Giả sử khoảng cách giữa 2 cụm gần nhất U và V là d_{UV}



Hình 3.1: Khoảng cách giữa các cụm cho (a) liên kết đơn, (b) liên kết hoàn chỉnh và (c) liên kết trung bình.

3. Hợp nhất cụm U và V. Gán nhãn cho cụm mới này là (UV). Cập nhật lại ma trận khoảng cách: (a) xóa các hàng và cột tương ứng với cụm U và V; (b) thêm 1 hàng và 1 cột gồm các khoảng cách giữa cụm (UV) và các cụm còn lại.
4. Lặp lại bước 2 và 3. Tổng lần lặp $N - 1$ lần. Tất cả các phần tử sẽ tạo thành 1 cụm duy nhất sau khi kết thúc thuật toán.

3.1 Phân cụm theo liên kết đơn

Đầu vào cho 1 thuật toán liên kết đơn có thể là khoảng cách hoặc sự tương đồng giữa các cặp phần tử. Ban đầu mỗi phần tử là 1 cụm riêng biệt. Thuật toán tạo các cụm lớn hơn bằng cách hợp nhất các cụm nhỏ hơn gần nhau (tương đồng) nhất.

Tìm khoảng cách nhỏ nhất trong $D = \{d_{ik}\}$ và hợp nhất các phần tử tương ứng, ví dụ, hợp nhất U và V để có cụm (UV). Khoảng cách giữa cụm (UV) và cụm W bất kỳ tính bằng:

$$d_{(UV)W} = \min\{d_{UW}, d_{vW}\}.$$

Kết quả có thể hiển thị bằng biểu đồ dendrogram hoặc sơ đồ cây. Các nhánh (cụm) hợp nhất tại các nút, chúng cho biết mức độ hợp nhất xảy ra.

Ví dụ: Ta xét ma trận khoảng cách của 5 phần tử:

$$\mathbf{D} = \{d_{ik}\} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 0 & & & & \\ 9 & 0 & & & \\ 3 & 7 & 0 & & \\ 6 & 5 & 9 & 0 & \\ 11 & 10 & \textcircled{2} & 8 & 0 \end{bmatrix} \end{array}.$$

Coi mỗi phần tử là 1 cụm, hợp nhất 2 cụm gần nhau nhất. Ta có:

$$\min_{i,k}(d_{ik}) = d_{53} = 2.$$

Kết hợp 3 với 5 thành cụm (35). Tính khoảng cách (35) đến các cụm còn lại

$$d_{(35)1} = \min\{d_{31}, d_{51}\} = 3;$$

$$d_{(35)2} = \min\{d_{32}, d_{52}\} = 7;$$

$$d_{(35)4} = \min\{d_{34}, d_{54}\} = 8.$$

Xóa các hàng và cột tương ứng với 3 và 5, thêm 1 hàng và 1 cột cho cụm (35), có ma trận sau cập nhật:

$$\begin{array}{c} \\ (35) \\ 1 \\ 2 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} (35) & 1 & 2 & 4 \end{array} \\ \begin{bmatrix} 0 & & & \\ \textcircled{3} & 0 & & \\ 7 & 9 & 0 & \\ 8 & 6 & 5 & 0 \end{bmatrix} \end{array}.$$

Giờ khoảng cách nhỏ nhất là $d_{(35)1} = 3$. Ta ghép 2 cụm (1) và (35) thành (135)

$$d_{(135)2} = \min\{d_{12}, d_{(35)2}\} = 7;$$

$$d_{(135)4} = \min\{d_{14}, d_{(35)4}\} = 6.$$

Xóa các hàng và cột tương ứng với 1 và (35), thêm 1 hàng và 1 cột cho cụm (135), có ma trận sau cập nhật:

$$\begin{array}{c} \\ (135) \\ 2 \\ 4 \end{array} \begin{array}{c} \begin{array}{ccc} (135) & 2 & 4 \end{array} \\ \begin{bmatrix} 0 & & \\ 7 & 0 & \\ 6 & \textcircled{5} & 0 \end{bmatrix} \end{array}.$$

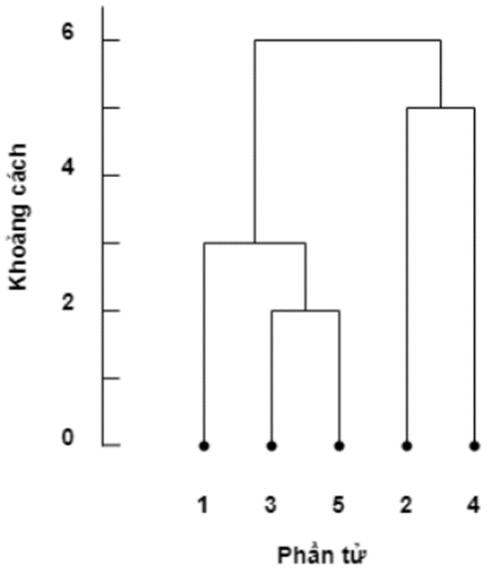
Giờ khoảng cách nhỏ nhất là $d_{42} = 5$. Ta ghép 2 cụm (2) và (4) thành (24)

$$d_{(135)(24)} = \min\{d_{(135)2}, d_{(135)4}\} = 6.$$

Ta có ma trận:

$$\begin{array}{c} \\ (135) \\ (24) \end{array} \begin{array}{c} \begin{array}{cc} (135) & (24) \end{array} \\ \begin{bmatrix} 0 & \\ ⑥ & 0 \end{bmatrix} \end{array}.$$

Cuối cùng, cụm (24) kết hợp với (135) thành 1 cụm duy nhất (12345) với khoảng cách gần nhất là 6. Biểu đồ 2 chiều của cách phân cụm vừa rồi:



Hình 3.2: Lược đồ dendrogram liên kết đơn cho khoảng cách giữa năm đối tượng

3.2 Phân cụm theo liên kết hoàn chỉnh

Phương pháp phân cụm theo liên kết hoàn chỉnh được thực hiện theo các bước tương tự phương pháp liên kết đơn. Điểm khác biệt so với phương pháp liên kết đơn là tiêu chí xác định khoảng cách giữa các cụm. khoảng cách giữa 2 cụm bất kì được xác định là khoảng cách giữa 2 phần tử xa nhau nhất của mỗi cụm.
Thay đổi công thức tính khoảng cách giữa cụm. Khoảng cách giữa cụm (UV) và bất kỳ cụm W nào khác được tính bằng công thức:

$$d_{(UV)W} = \max\{d_{UW}, d_{VW}\} \tag{3.2}$$

Ví dụ: Một giáo viên muôn chia học sinh của mình thành các nhóm khác nhau. Cô ấy có điểm của từng học sinh trong một môn học và dựa trên điểm

này, cô ấy muốn chia học sinh thành các nhóm. không có mục tiêu cố định ở đây là bao nhiêu nhóm.

id_sv	diem1
1	10
2	7
3	28
4	20
5	35

Xác định ma trận khoảng cách của các đối tượng

$$D = \{d_{ik}\} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{pmatrix} 0 & & & & \\ 3 & 0 & & & \\ 18 & 21 & 0 & & \\ 10 & 13 & 8 & 0 & \\ 25 & 28 & 7 & 15 & 0 \end{pmatrix} \end{array}.$$

Ta có $d_{12} = 3$ là khoảng cách ngắn nhất nên ta ghép 1 và 2 thành cụm (12).

Tính các khoảng cách từ cụm (12) đến các phần tử còn lại:

$$d_{(12)3} = \max\{d_{13}, d_{23}\} = \max\{18, 21\} = 21;$$
$$d_{(12)4} = \max\{d_{14}, d_{24}\} = \max\{10, 13\} = 13;$$
$$d_{(12)5} = \max\{d_{15}, d_{25}\} = \max\{25, 28\} = 28.$$

Xóa đi các hàng và các cột tương ứng với phần tử (1) và (2) đồng thời thêm một hàng và cột cho cụm (12) ta được ma trận khoảng cách mới:

$$\begin{array}{c} \\ (12) \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{cccc} (12) & 3 & 4 & 5 \end{array} \\ \begin{pmatrix} 0 & & & \\ 21 & 0 & & \\ 13 & 8 & 0 & \\ 28 & 7 & 15 & 0 \end{pmatrix} \end{array}.$$

Ta có $d_{35} = 7$ là khoảng cách lớn nhất nên ta hợp nhất 3 và 5 thành một cụm (35).
Tính các khoảng cách từ cụm (35) đến các phần tử còn lại:

$$(d_{(35)(12)} = max\left\{d_{3(12)}, d_{5(12)}\right\} = max\left\{21, 28\right\} = 28;$$
$$(d_{(35)4}) = max\left\{d_{34}, d_{54}\right\} = max\left\{8, 15\right\} = 15.$$

Xóa đi các hàng và các cột tương ứng với phần tử 3 và 5 đồng thời thêm vào một hàng và một cột cho cụm (35). Ta được ma trận khoảng cách mới:

$$\begin{array}{c} \\ (12) \\ (35) \\ 4 \end{array} \begin{array}{c} \begin{array}{ccc} (12) & (35) & 4 \end{array} \\ \begin{pmatrix} 0 & & \\ 28 & 0 & \\ 13 & 15 & 0 \end{pmatrix} \end{array}.$$

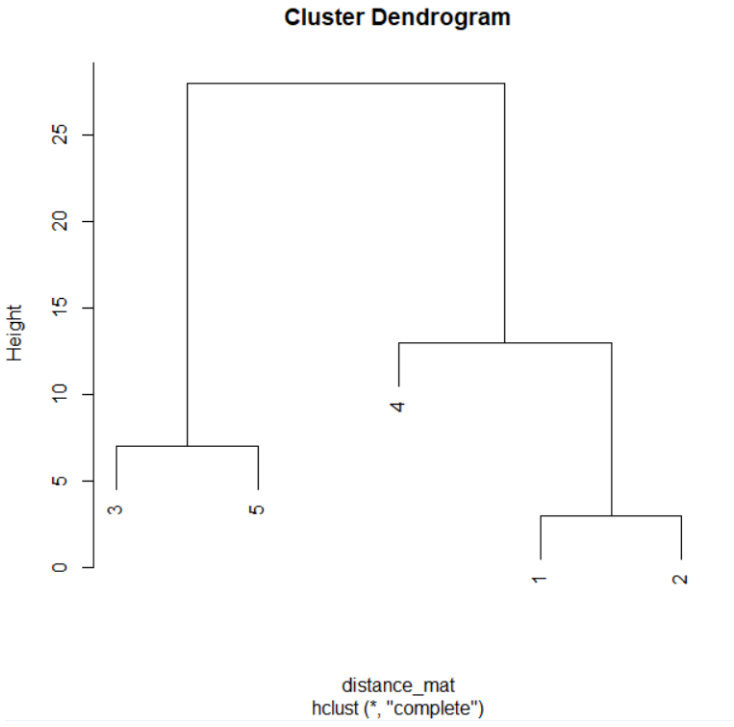
Ta có $d_{(12)(4)} = 13$ là khoảng cách lớn nhất nên ta hợp nhất (12) và (4) thành một cụm (124).
Tính các khoảng cách từ cụm (124) đến các phần tử còn lại:

$$(d_{(124)(35)}) = max\left\{d_{(12)(35)}, d_{4(35)}\right\} = max\left\{15, 28\right\} = 28.$$

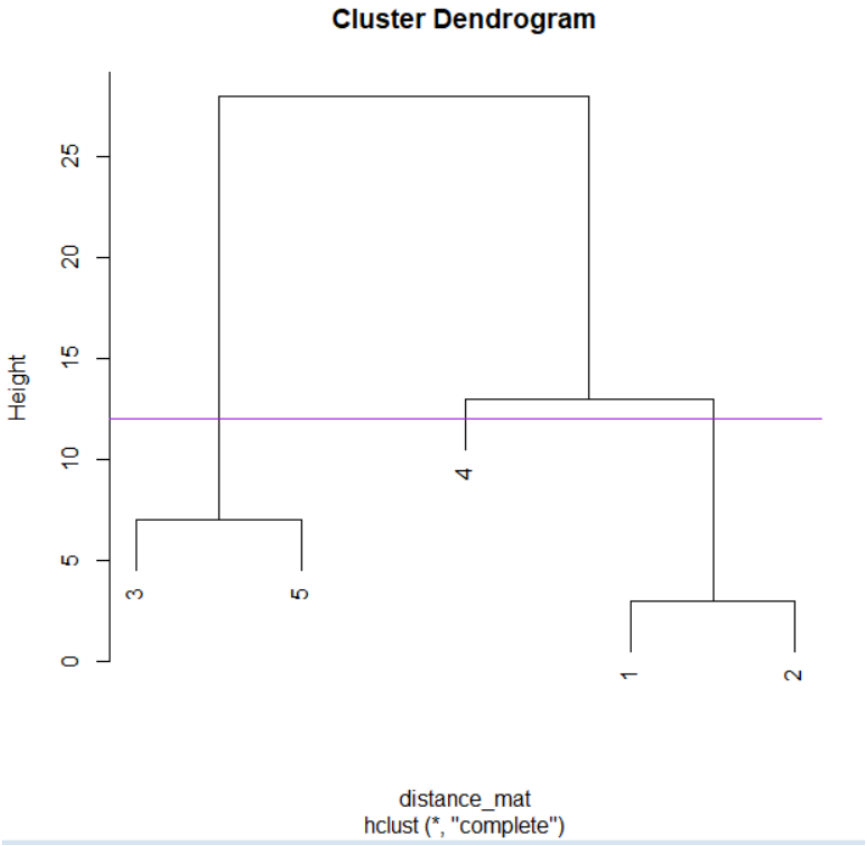
Xóa đi các hàng và các cột tương ứng với phần tử (12) và (4) đồng thời thêm vào một hàng và một cột cho cụm (124). Ta được ma trận khoảng cách mới:

$$\begin{array}{c} \\ (124) \\ (35) \end{array} \begin{array}{c} \begin{array}{cc} (124) & (35) \end{array} \\ \begin{pmatrix} 0 & \\ 28 & 0 \end{pmatrix} \end{array}.$$

Hợp nhất cụm (124) và cụm (35) thành 1 cụm (12345) với khoảng cách ngắn nhất là 28.
Ta thu được biểu đồ Dendrogram:



Cắt biểu đồ dendrogram bằng một đường nằm ngang tại một chiều cao nhất, ta sẽ xác định được ở khoảng cách đó sẽ có bao nhiêu cụm.



Phân cụm liên kết hoàn chỉnh tránh được nhược điểm của phương pháp liên kết đơn: có xu hướng tạo ra các cụm dài và lỏng lẻo, mà các cụm đó thường sẽ bao gồm các phần tử khác biệt trong cùng một cụm nên thường ít sử dụng trong các bài toán thực tế, các cụm có thể bị ràng buộc với nhau do các phần tử đơn lẻ ở gần nhau, mặc dù nhiều phần tử trong mỗi cụm có thể rất xa nhau.
Liên kết hoàn chỉnh có xu hướng tìm các cụm nhỏ gọn có đường kính xấp xỉ bằng nhau, tuy nhiên khá nhạy cảm với nhiễu.

3.3 Phân cụm theo liên kết trung bình

Phương pháp phân cụm theo liên kết trung bình coi khoảng cách giữa hai cụm là khoảng cách trung bình giữa tất cả các cặp phần tử, trong đó mỗi phần tử của cặp thuộc về mỗi cụm.
Thay đổi công thức tính khoảng cách giữa các cụm ở Bước 3. Trong thuật toán chung. Khi đó khoảng cách giữa (UV) và bất kỳ cụm W nào khác được

tính bằng công thức:

$$d_{(UV)W} = \frac{\sum_i \sum_k d_{ik}}{N_{(UV)} N_W} \tag{3.3}$$

trong đó, d_{ik} là khoảng cách giữa phần tử i trong cụm (UV) và phần tử k trong cụm W, $N_{(UV)}$ và N_W là số phần tử trong cụm (UV) và W tương ứng.
Ví dụ: Xác định ma trận khoảng cách của 5 phần tử

$$D = \{d_{ik}\} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{pmatrix} 0 & & & & \\ 9 & 0 & & & \\ 3 & 7 & 0 & & \\ 6 & 5 & 9 & 0 & \\ 11 & 10 & 2 & 8 & 0 \end{pmatrix} \end{array}.$$

Coi mỗi phần tử là 1 cụm, hợp nhất 2 cụm gần nhau nhất là cụm 3 và cụm 5. Ta có:

$$\min_{i,k}(d_{ik}) = d_{53} = 2.$$

Tính các khoảng cách từ cụm (35) đến các phần tử còn lại:

$$d_{(35)1} = AVG(d_{31}, d_{51}) = \frac{3+11}{2} = 7;$$
$$d_{(35)2} = AVG(d_{32}, d_{52}) = \frac{7+10}{2} = 8,5;$$
$$d_{(35)4} = AVG(d_{34}, d_{54}) = \frac{9+8}{2} = 8,5.$$

Xóa đi các hàng và các cột tương ứng với phần tử thứ 3 và thứ 5, đồng thời thêm một hàng và cột cho cụm (35) ta được ma trận khoảng cách mới:

$$\begin{array}{c} \\ (35) \\ 1 \\ 2 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} (35) & 1 & 2 & 4 \end{array} \\ \begin{pmatrix} 0 & & & \\ 7 & 0 & & \\ 8.5 & 9 & 0 & \\ 8.5 & 6 & 5 & 0 \end{pmatrix} \end{array}.$$

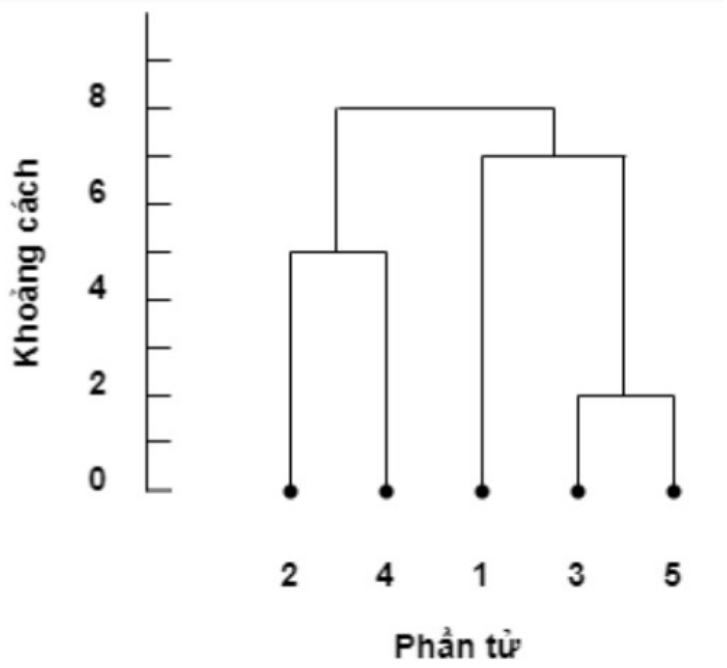
Lặp lại các bước 2, bước 3 ở những vòng lặp tiếp theo. Sau khi gộp cụm 2 và 4:

$$\begin{array}{c} \\ (35) \\ (24) \\ 1 \end{array} \begin{array}{c} \begin{array}{ccc} (35) & (24) & 1 \end{array} \\ \begin{pmatrix} 0 & & \\ 8.5 & 0 & \\ 7 & 8.6 & 0 \end{pmatrix} \end{array}.$$

Sau khi gộp cụm 1 và cụm (35):

$$\begin{array}{c} \\ (135) \\ (24) \end{array} \begin{array}{c} \begin{array}{cc} (135) & (24) \end{array} \\ \begin{pmatrix} 0 & \\ 8 & 0 \end{pmatrix} \end{array}.$$

Hợp nhất cụm (24) và cụm (135) thành một cụm duy nhất (12345)
Ta thu được biểu đồ Dendrogram:



Chương 4

Phương pháp phân cụm không theo thứ bậc

Mục tiêu:
Chương này giúp mọi người biết đến và hiểu phương pháp K-Mean, từ đó vận dụng nó vào trong phân tích số liệu thực tế.
Tóm tắt nội dung chương:

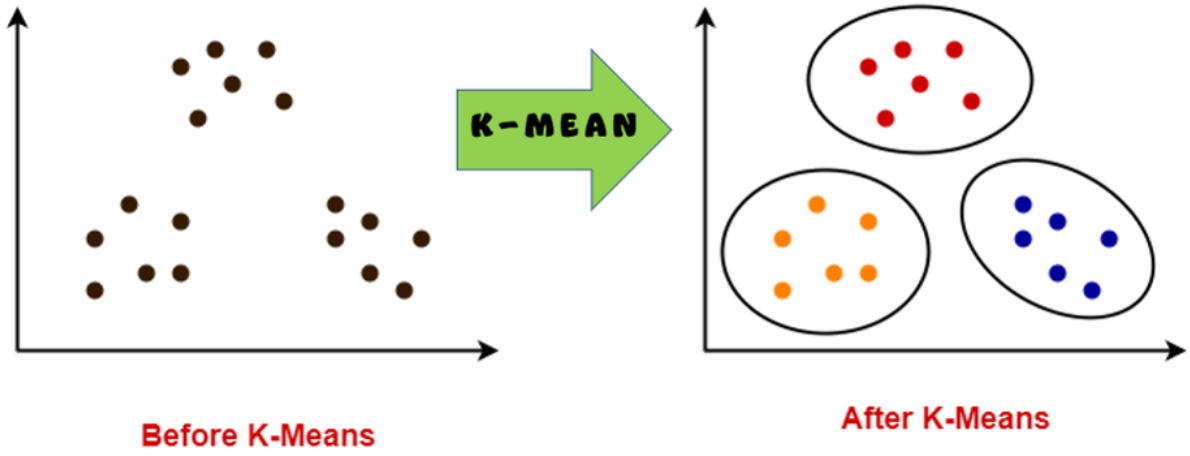
- **4.1** nói về nguồn gốc, ý tưởng phương pháp K-Mean và ứng dụng của nó trong 1 số lĩnh vực.
- **4.2** đưa ra thuật toán và thực hiện thuật toán thủ công trên dữ liệu nhỏ.
- **4.3** nói về cách tìm giá trị k hợp lý bằng phương pháp khuỷu tay (Elbow Method).
- **4.4** in ra kết quả chạy dữ liệu thực tế với công cụ R-Studio và giải thích kết quả tìm được.
- **4.5** nêu ưu/nhược điểm của phương pháp K-Mean.

4.1 Giới thiệu phương pháp K-Mean

Phương pháp K-Mean được đề xướng bởi nhà toán học nổi tiếng McQueen vào năm 1967.
Phân cụm K-Mean được đề xuất khi chúng ta không biết nhãn (label) của từng điểm dữ liệu. Mục đích là phân chia n giá trị quan sát vào K cụm khác nhau sao cho dữ liệu trong cùng một cụm có tính chất giống nhau.

Phân cụm K-Mean rất đơn giản, dễ hiểu, có nhiều hữu ích đặc biệt và cực kỳ phổ biến trong ngành khoa học dữ liệu.
Trong đó như :

- Phương pháp K-Mean được sử dụng rộng rãi trong nghiên cứu thị trường, nhận dạng mẫu, phân tích dữ liệu và xử lý ảnh.
- Phương pháp K-Mean cũng có thể giúp các nhà khoa học dữ liệu khám phá ra các nhóm khách hàng của họ. Và họ có thể mô tả đặc điểm nhóm khách hàng của mình dựa trên lịch sử mua hàng.
- Trong lĩnh vực sinh học, K-Mean có thể được sử dụng để xác định phân loại thực vật và động vật, phân loại các gen có chức năng tương tự và hiểu sâu hơn về các cấu trúc vốn có của quần thể.
- Trong lĩnh vực chứng khoán, K-Mean được dùng để phân các doanh nghiệp/ngành có cùng dao động với nhau. Người ta khuyến cáo không nên đầu tư nhiều chứng khoán trong cùng 1 cụm vì nó tương đương đánh 1 chứng khoán.



Hình 4.1: Ví dụ về phân 18 đối tượng như hình bên trái vào 3 cụm theo thuật toán K-Mean dựa trên 2 thuộc tính

4.2 Thuật toán K-Mean

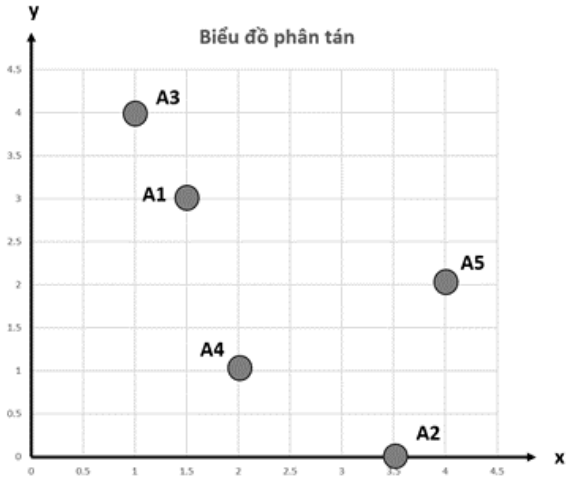
McQueen đưa ra phương pháp K-means để mô tả thuật toán của ông là phân phối mỗi đối tượng vào cụm có trung tâm gần nó nhất. Quá trình đó được tạo qua 5 bước:

Bước 1	Chọn số cụm (Chọn giá trị **K**)
Bước 2	Phân chia ngẫu nhiên các đối tượng vào **K** cụm ban đầu
Bước 3	Tính trung tâm cho từng cụm Công thức tính trung tâm cho cụm bất kỳ có n điểm dữ liệu: $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ Trong đó: X_i là các đối tượng trong cụm đang xét n là số lượng điểm dữ liệu trong cụm đang xét $\bar{X}$ là điểm trung tâm của cụm đang xét
Bước 4	Từ toàn bộ danh sách các điểm dữ liệu, phân phối từng điểm dữ liệu cho cụm có trung tâm gần nó nhất. Trường hợp khoảng cách điểm dữ liệu đến trung tâm 2 cụm bằng nhau thì phân điểm đó vào cụm có chỉ số nhỏ hơn. Tính lại trung tâm cho cụm nhận được đối tượng mới và cụm mất đối tượng.
Bước 5	Lặp lại bước 4 cho đến khi không có sự phân phối lại Người ta sẽ cho trước 1 số vòng lặp đủ lớn gọi là **m**. Trong trường hợp, sau mỗi lần lặp, luôn có sự phân phối lại, thì thuật toán sẽ dừng lại ngay khi số vòng lặp bằng **m**

Ở **bước 2 và bước 3**, thay vì nhóm các đối tượng thành K nhóm rồi mới tính tâm cụm như trên, chúng ta có thể chỉ định K điểm trung tâm ban đầu (lấy trong tập dữ liệu điểm đã có) rồi mới phân phối các đối tượng vào cụm có trung tâm gần nó nhất.

Ví dụ 4-1: Chia 5 đối tượng vào 2 nhóm bằng phương pháp K-Mean ?

Đối Tượng	Quan Sát	
	x	y
A1	1.5	3
A2	3.5	0
A3	1	4
A4	2	1
A5	4	2



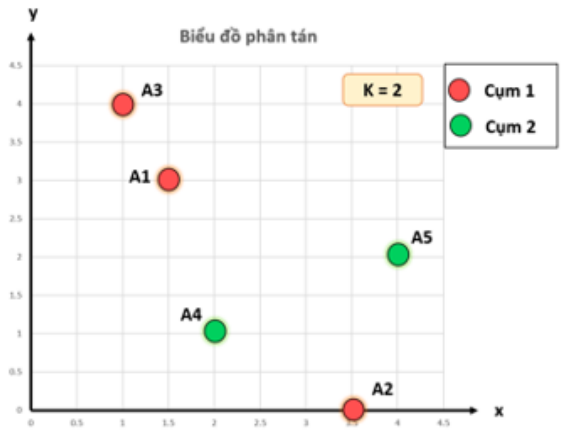
Bước 1: Chọn **K = 2**

Bước 2: Phân chia ngẫu nhiên các đối tượng vào **2 cụm**

Cụm 1 = (A1, A2, A3)

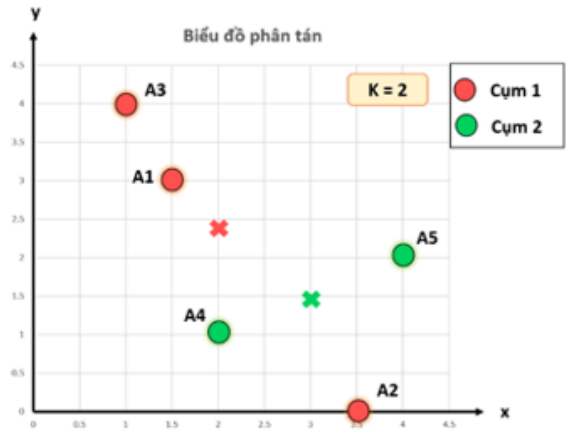
Cụm 2 = (A4, A5)

Đối Tượng	Quan Sát		Cụm
	x	y	
A1	1.5	3	1
A2	3.5	0	1
A3	1	4	1
A4	2	1	2
A5	4	2	2



Bước 3: Tính trung tâm cho 2 cụm

Cụm	$\overline{x}$	$\overline{y}$
Cụm 1 (A1, A2, A3)	$\frac{1.5+3.5+1}{3}=2$	$\frac{3+0+4}{3}=\frac{7}{3}$
Cụm 2 (A4, A5)	$\frac{2+4}{2}=3$	$\frac{1+2}{2}=1.5$



Xét 2 cụm ban đầu, cụm 1 và cụm 2 có trung tâm lần lượt là $(2; \frac{7}{3})$ và $(3; 2)$

Bước 4:

Ta tính khoảng cách Euclide cho từng đối tượng tới trung tâm của các cụm và phân phối lại mỗi đối tượng cho cụm có tâm gần nó nhất.

Nếu đối tượng được chuyển từ nhóm này sang nhóm khác thì trung tâm của các cụm liên quan phải được tính lại trước khi tiếp tục.

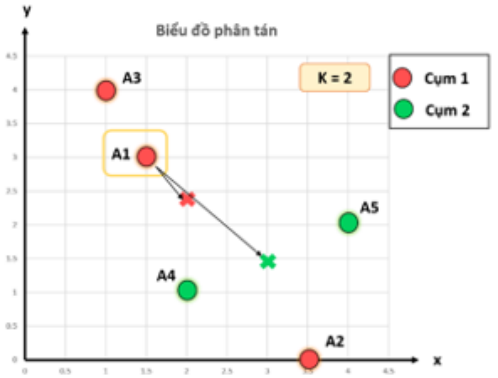
Ta tính bình phương khoảng cách từ đối tượng A1 đến tâm 2 cụm, ta có:

$$d^2\,(A1, tâm\ 1) = (1.5-2)^2 + \left(3-\frac{7}{3}\right)^2 = 0.6944$$

$$d^2\,(A1, tâm\ 2) = (1.5-3)^2 + (3-1.5)^2 = 4.5$$

⇨ *A1 (1.5; 3) gần tâm cụm 1 hơn tâm cụm 2*
Mà A1 đang ở cụm 1

⇨ *Không có sự phân phối lại*



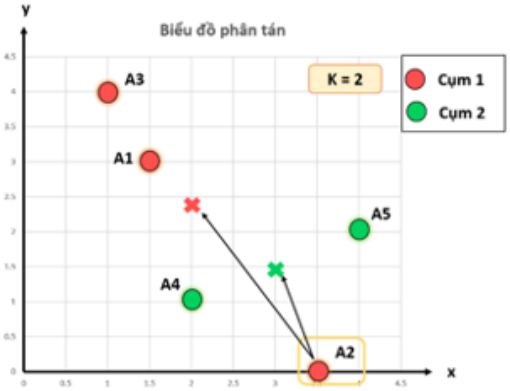
Tiếp tục xét đối tượng A2, ta có:

$$d^2\,(A2, tâm\ 1) = (3.5-2)^2 + \left(0-\frac{7}{3}\right)^2 = 7.69$$

$$d^2\,(A2, tâm\ 2) = (3.5-3)^2 + (0-1.5)^2 = 2.5$$

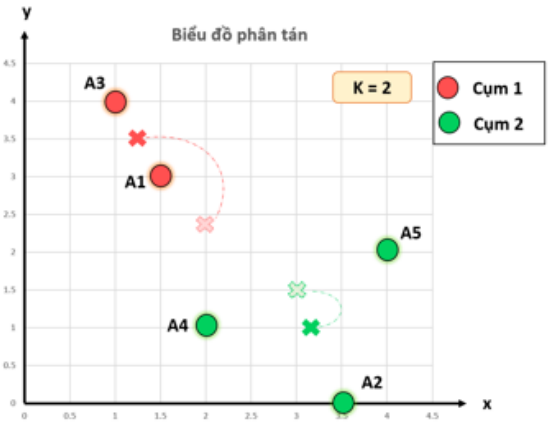
⇨ *A2 (3.5; 0) gần tâm cụm 2 hơn tâm cụm 1*
Mà A2 đang ở cụm 1

⇨ *A2 chuyển từ cụm 1 sang cụm 2*



Ta tính lại tâm các cụm:

Cụm	$\overline{x}$	$\overline{y}$
Cụm 1 (A1, A3)	$\frac{1.5+1}{2} = 1.25$	$\frac{3+4}{2} = 3.5$
Cụm 2 (A2, A4, A5)	$\frac{3.5+2+4}{3} = 3.17$	$\frac{0+1+2}{3} = 1$



Xét 2 cụm mới, cụm 1 và cụm 2 có trung tâm lần lượt là (1.25 ; 3.5) và (3.17 ; 1)

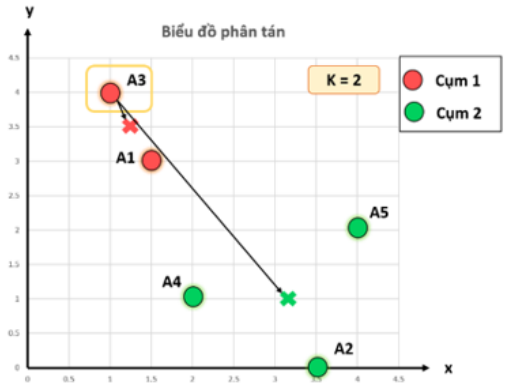
Bây giờ, ta tiếp tục xét các điểm tiếp theo với 2 cụm mới Cụm 1(A1, A3) và Cụm 2(A2, A4, A5)

Tiếp tục xét đối tượng A3, ta có:

$$d^2\,(A3, tâm\ 1) = (1 - 1.25)^2 + (4 - 3.5)^2 = 0.31$$

$$d^2\,(A3, tâm\ 2) = (1 - 3.17)^2 + (4 - 1)^2 = 13.69$$

⇨ *A3 (1; 4) gần tâm cụm 1 hơn tâm cụm 2*
Mà A3 đang ở cụm 1
⇨ *Không có sự phân phối lại*

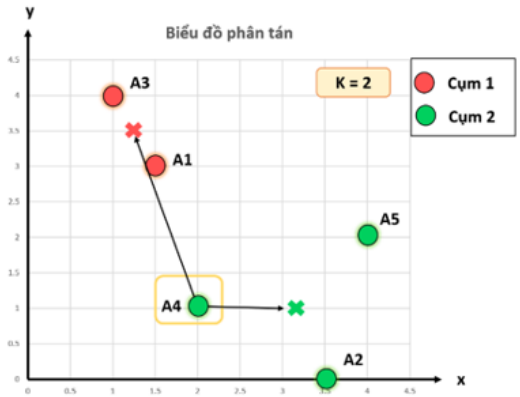


Tiếp tục xét đối tượng A4, ta có:

$$d^2\,(A4, tâm\ 1) = (2 - 1.25)^2 + (1 - 3.5)^2 = 6.81$$

$$d^2\,(A4, tâm\ 2) = (2 - 3.17)^2 + (1 - 1)^2 = 1.36$$

⇨ *A4 (2; 1) gần tâm cụm 2 hơn tâm cụm 1*
Mà A4 đang ở cụm 2
⇨ *Không có sự phân phối lại*

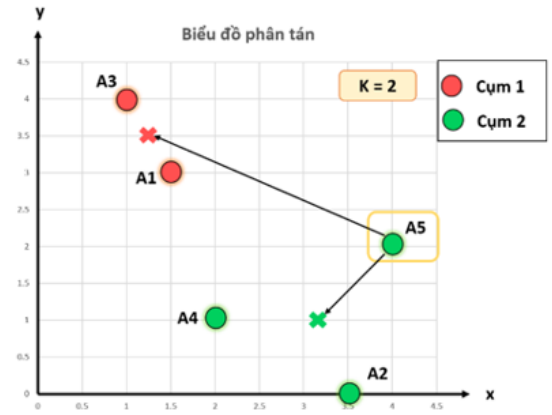


Tiếp tục xét đối tượng A5, ta có:

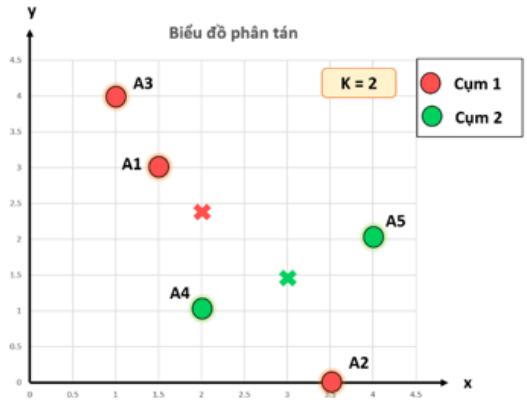
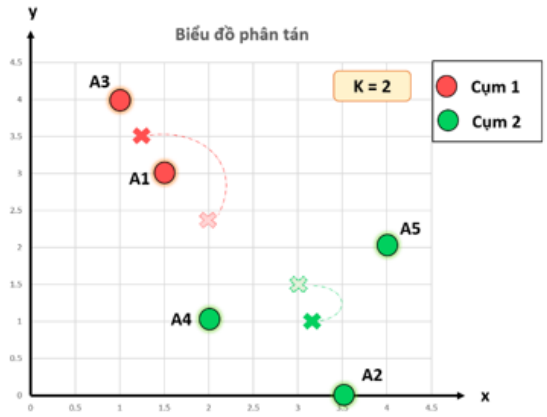
$$d^2\,(A5, tâm\ 1) = (4 - 1.25)^2 + (2 - 3.5)^2 = 9.81$$

$$d^2\,(A5, tâm\ 2) = (4 - 3.17)^2 + (2 - 1)^2 = 1.69$$

⇨ *A5 (4; 2) gần tâm cụm 2 hơn tâm cụm 1*
Mà A5 đang ở cụm 2
⇨ *Không có sự phân phối lại*

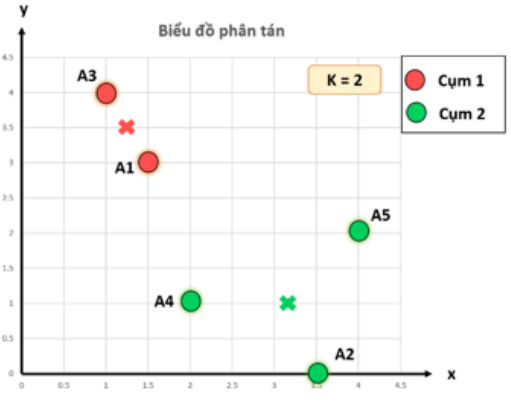
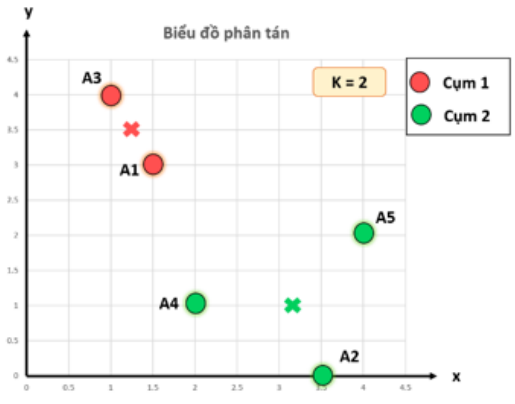


Bước 5:

Trước khi thực hiện bước 4	Sau khi duyệt hết 5 đối tượng ở bước 4
Cụm 1 = (A1, A2, A3) Cụm 2 = (A4, A5) 	Cụm 1 = (A1, A3) Cụm 2 = (A2, A4, A5) 

Khi so sánh các cụm trước và sau khi thực hiện bước 4, nhận thấy có sự phân phối lại (A2 từ cụm 1 sang cụm 2), nên ta lặp lại bước 4.

Trong lần lặp này, không có bất kỳ đối tượng nào được chuyển từ nhóm này sang nhóm khác.

Trước khi lặp lại bước 4	Sau khi lặp lại bước 4
Cụm 1 = (A1, A3) Cụm 2 = (A2, A4, A5) 	Cụm 1 = (A1, A3) Cụm 2 = (A2, A4, A5) 

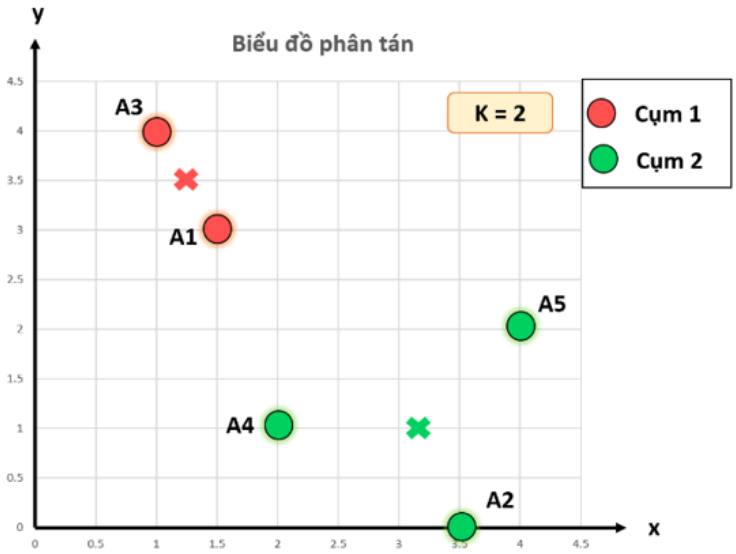
Khi so sánh các cụm trước và sau khi lặp bước 4, nhận thấy không có sự phân phối lại $\Rightarrow$ **Dừng thuật toán**

Kết quả thu được:

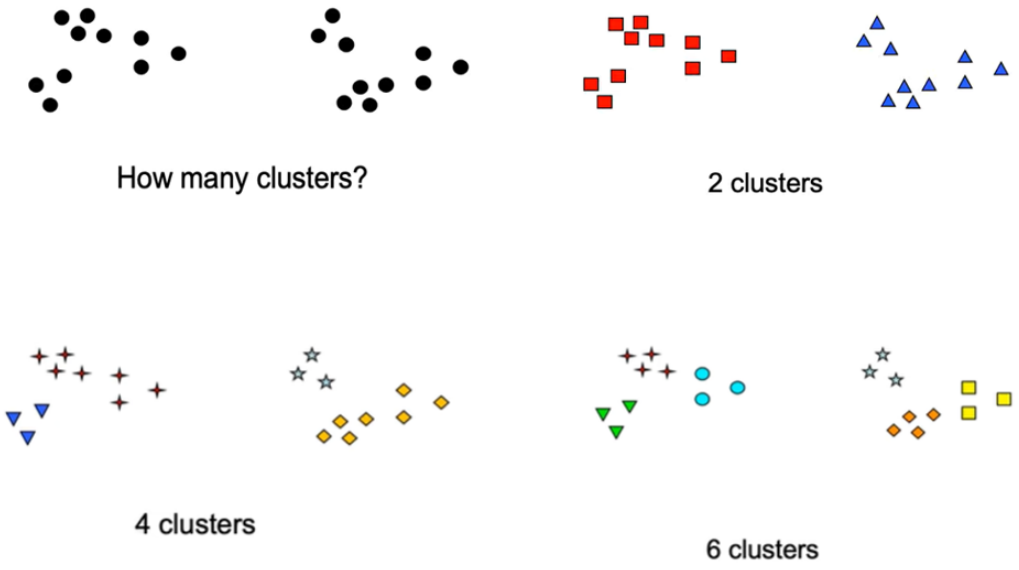
Đối Tượng	Quan Sát	
	x	y
A1	1.5	3
A3	1	4
Tâm cum 1	1.25	3.5

Đối Tượng	Quan Sát	
	x	y
A2	3.5	0
A4	2	1
A5	4	2
Tâm cum 2	1.25	3.5

Cum 1 = (A1, A3)
Cum 2 = (A2, A4, A5)



4.3 Tìm giá trị K tối ưu bằng phương pháp Elbow



Vấn đề:
Trong bài toán phân tích dữ liệu thực tế, việc lựa chọn K chính xác thường không rõ ràng, nó còn tùy thuộc vào hình dạng và tỷ lệ phân bố các điểm trong tập dữ liệu và độ phân giải phân cụm mong muốn của người dùng. Ngoài ra, việc tăng giá trị k sẽ làm giảm số lượng lỗi trong quá trình phân cụm, đến trường hợp xấu nhất là không có lỗi nếu mỗi điểm dữ liệu được coi là cụm của chính nó (nghĩa là khi k bằng số điểm dữ liệu).
Khi đó, theo trực giác, lựa chọn K tối ưu sẽ tạo ra sự cân bằng giữa việc nén dữ liệu tối đa bằng cách sử dụng một cụm duy nhất và độ chính xác tối đa bằng cách gán từng điểm dữ liệu cho cụm riêng của nó.

Giải pháp:
Trong phân tích cụm, phương pháp khuỷu tay (Elbow Method) là một kỹ thuật được sử dụng để xác định số lượng cụm trong một tập dữ liệu. Phương pháp này bao gồm vẽ đồ thị và phát hiện khuỷu tay (nơi cánh tay bắt đầu đóng lại), chọn giá trị k tương ứng tại điểm đó là số lượng cụm sẽ sử dụng.

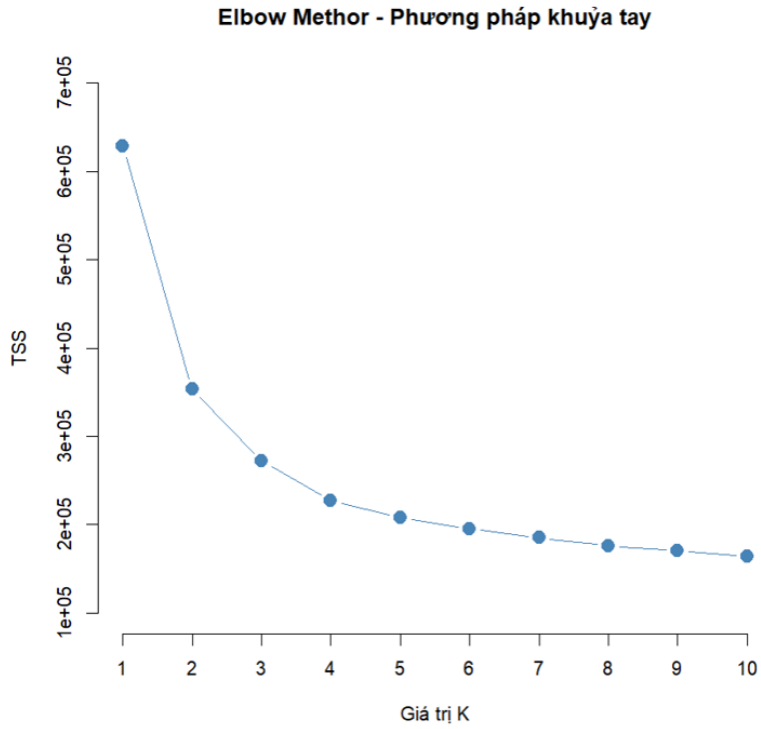
Thuật toán:

- Bước 1: Ta dùng thuật toán K-Mean phân tích bộ dữ liệu với nhiều giá trị k khác nhau (k nhận các giá trị dương liên tiếp bắt đầu từ 1).
- Với mỗi kết quả phân cụm tương ứng với mỗi giá trị k, ta tính tổng bình phương khoảng cách tới tâm bên trong các cụm **TSS** (**Total within-clusters sum of squares**).

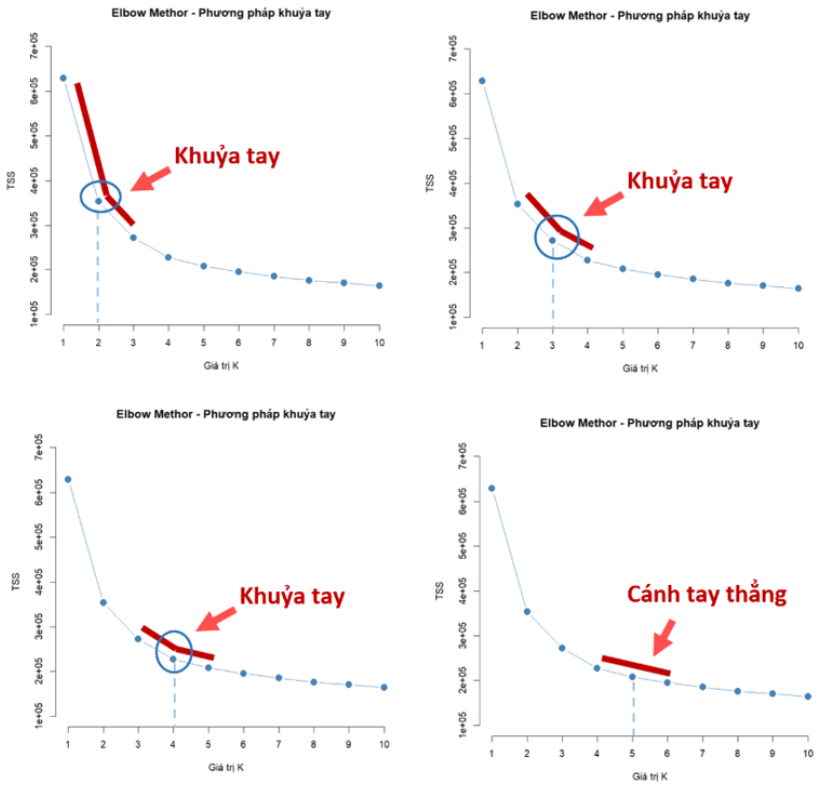
$$TSS(k) = \sum_{i=1}^{k} \sum_{X \in C_i} d(X, \overline{X_i})^2$$

Trong đó:
X là các đối tượng trong bộ dữ liệu
C_i là tên cụm thứ i ($i \in \{1, 2, \ldots, k\}$)
$\overline{X_i}$ là tâm cụm thứ i

- Bước 3: Biểu diễn giá trị TSS và giá trị k tương ứng qua đồ thị như hình dưới.

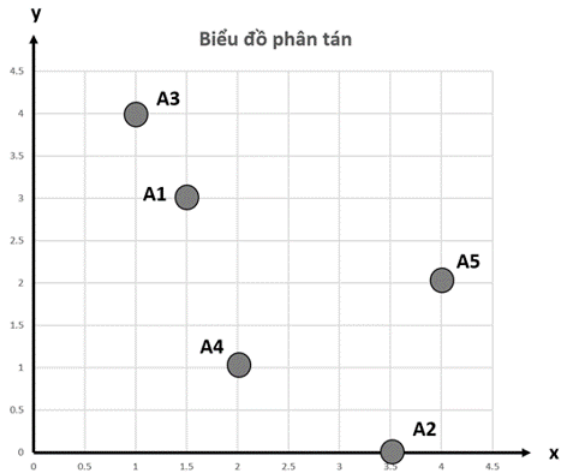


- Bước 4: Phát hiện điểm mà tại điểm đó cánh tay bắt đầu đóng lại, để lộ ra khuỷu tay. Như hình dưới, ta có thể chọn k=2, k=3 hoặc k=4 là số cụm thích hợp. Vì $TSS_{k=4}$ nhỏ hơn nên chọn giá trị k=4 là tốt hơn cả.

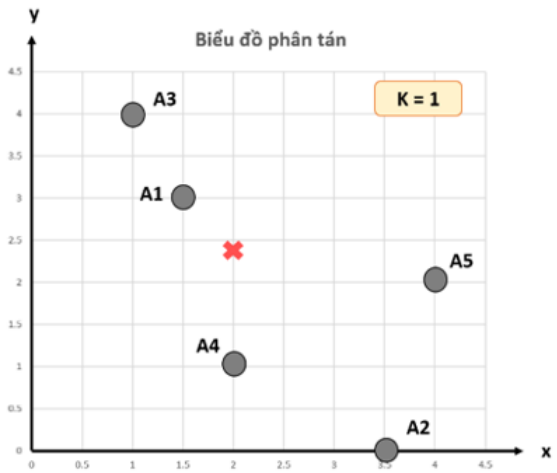


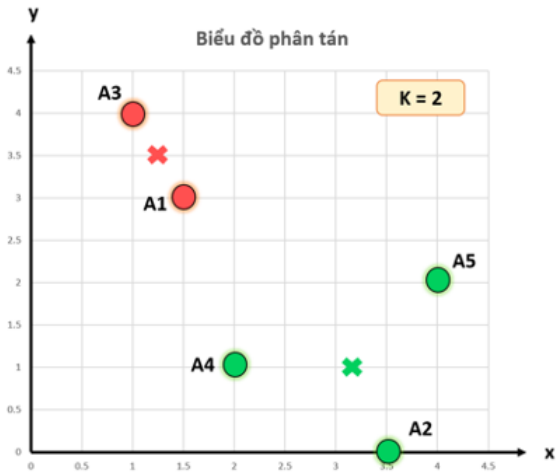
Ví dụ 4-2: Chia 5 đối tượng vào bao nhiêu nhóm thì hợp lý ?

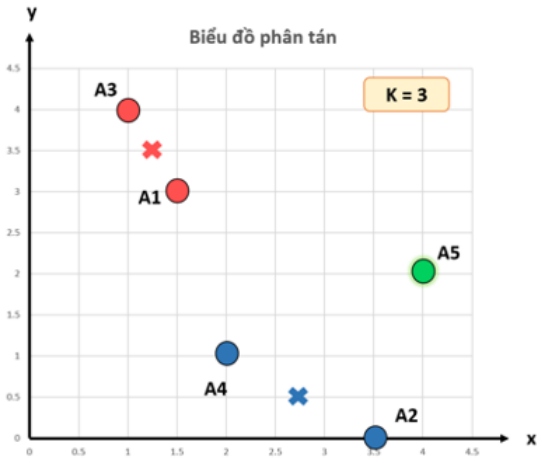
Đối Tượng	Quan Sát	
	x	y
A1	1.5	3
A2	3.5	0
A3	1	4
A4	2	1
A5	4	2

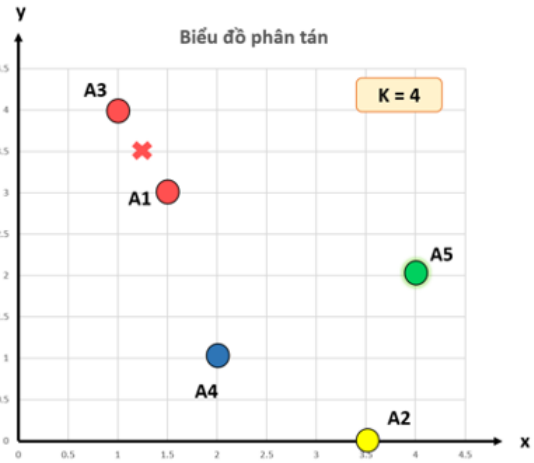


Bước 1: Ta dùng thuật toán K-Mean phân tích bộ dữ liệu với các giá trị k từ 1 đến 4, thu được kết quả như sau.

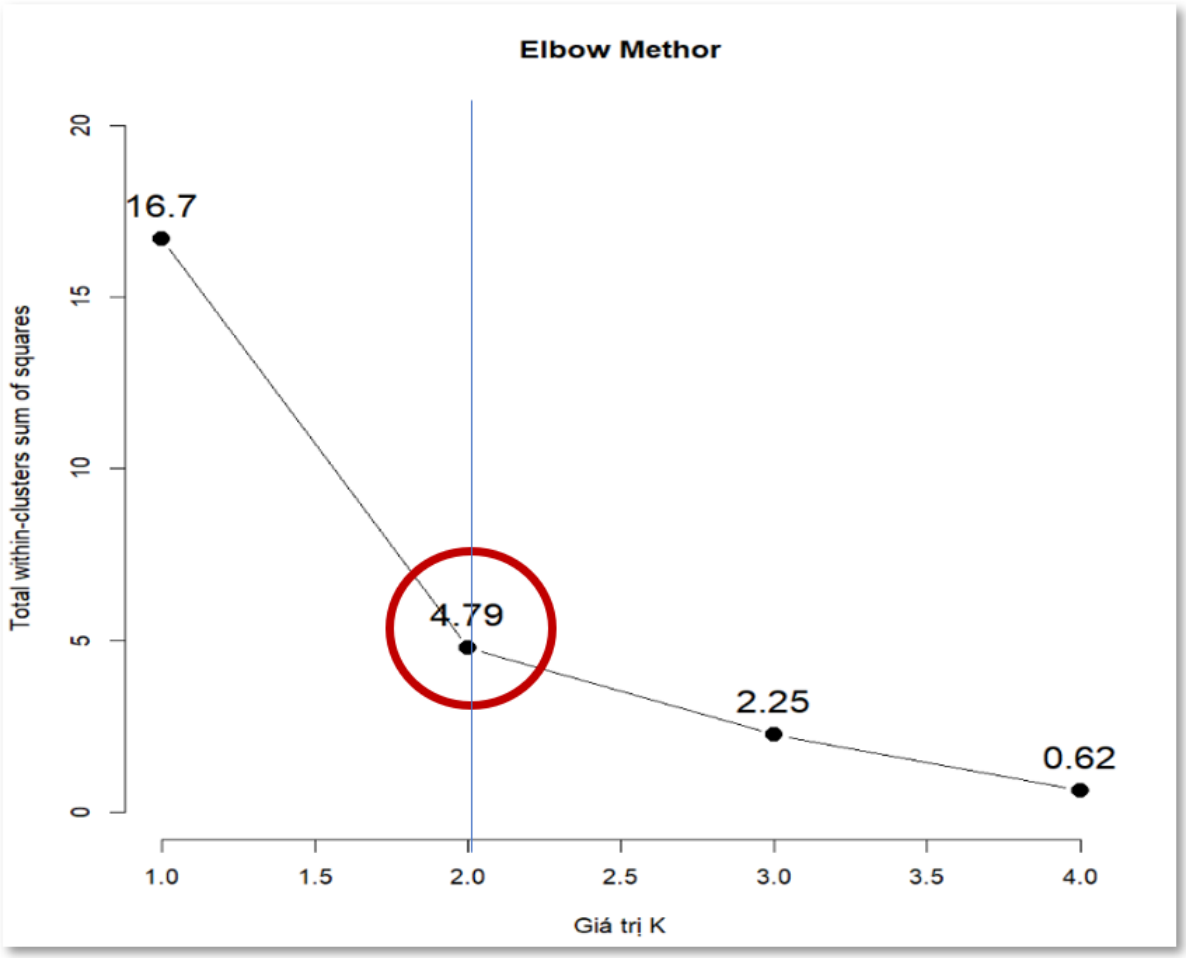








Bước 2: Tính các giá trị **TSS** tương ứng với mỗi giá trị k
(không xét $TSS_{k=5}$ vì khi k=5, thì mỗi điểm dữ liệu được coi là cụm của chính nó, phân cụm trở thành vô nghĩa)
$TSS_{k=1} = 16,7$;
$TSS_{k=2} = 4,79$;
$TSS_{k=3} = 2,25$;
$TSS_{k=4} = 0,62$.
Bước 3+4: Dựa vào đồ thị, tại vị trí k=2, phát hiện cánh tay đóng, khuỷu tay là điểm(2; 4,79)
⇒ **Chọn k=2 là số cụm hợp lý để dùng cho phân cụm**:



Nhận Xét:
Có nhiều phương pháp tìm giá trị k phù hợp khác, nhưng không có phương pháp nào được gọi là phương pháp cho giá trị k tối ưu nhất. Chúng ta có thể nhờ những chuyên gia nắm rõ bộ dữ liệu, từ đó đưa ra lời khuyên chọn giá trị k thích hợp.
Tốt nhất là nên chạy thuật toán với một vài lựa chọn ngẫu nhiên khác nhau.

4.4 Thực hành phân tích số liệu thực tế trên R Studio

Nhóm em lấy bộ dữ liệu các cầu thủ Fifa20 trên Kaggle, gồm 18483 quan sát (Cầu thủ).
Trong tài liệu này, em nhóm các cầu thủ khác nhau thành các nhóm dựa trên các thuộc tính cơ bản của họ. Em sẽ tận dụng thuật toán k-means trong R để phân nhóm người chơi thành các nhóm khác nhau.

Bước 1: Chuẩn bị dữ liệu

Nhập dữ liệu từ file Excel:

A tibble: 18483 × 110

sofifa_id	player_url	short_name	long_name	player_positions	overall	potential
<dbl>	<chr>	<chr>	<chr>	<chr>	<dbl>	<dbl>
158023	https://sofifa.com/player/158023/lionel-messi/200002	L. Messi	Lionel Andrés Messi Cuccittini	RW, CF, ST	94	94
20801	https://sofifa.com/player/20801/c-ronaldo-dos-santos-aveiro/200002	Cristiano Ronaldo	Cristiano Ronaldo dos Santos Aveiro	ST, LW	93	93
190871	https://sofifa.com/player/190871/neymar-da-silva-santos-jr/200002	Neymar Jr	Neymar da Silva Santos Júnior	LW, CAM	92	92
183277	https://sofifa.com/player/183277/eden-hazard/200002	E. Hazard	Eden Hazard	LW, CF	91	91

Tệp có nhiều thông tin không cần để phân cụm, chỉ các thuộc tính của cầu thủ như tốc độ, sức mạnh, chuyền bóng, dứt điểm, đánh đầu,v.v. mới được tính.

	attacking_crossing	attacking_finishing	attacking_heading_accuracy	attacking_short_passing
1	88	95	70	92
2	84	94	89	83
3	87	87	62	87
4	81	84	61	89
5	93	82	55	92
6	13	11	15	43
7	86	72	55	92
8	18	14	11	61
9	53	52	86	78
10	79	90	59	84

Ta được bộ dữ liệu gồm **18483 quan sát** (Cầu thủ) và **34 biến** (cùng đơn vị: thang đo 100 điểm) được giải thích dưới đây:

TẤN CÔNG	1	attacking_crossing	Tạt cánh
	2	attacking_finishing	Dứt điểm
	3	attacking_heading_accuracy	Đánh đầu
	4	attacking_short_passing	Chuyền ngắn
	5	attacking_volleys	Sút Vô Lê
KỸ NĂNG	6	skill_dribbling	Rê bóng
	7	skill_curve	Chuyền cong
	8	skill_fk_accuracy	Đánh đầu
	9	skill_long_passing	Chuyền dài
	10	skill_ball_control	Kiểm soát bóng
DI CHUYỂN	11	movement_acceleration	Tăng tốc
	12	movement_sprint_speed	Tốc độ nước rút
	13	movement_agility	Sự nhanh nhẹn
	14	movement_reactions	Phản ứng/ Phản xạ
	15	movement_balance	Giữ thăng bằng
THỂ CHẤT	16	power_shot_power	Lực sút
	17	power_jumping	Nhảy cao
	18	power_stamina	Sức bền
	19	power_strength	Sức mạnh
	20	power_long_shots	Sút xa
TINH THẦN	21	mentality_aggression	Hiếu chiến
	22	mentality_interceptions	Chặn đường chuyền
	23	mentality_positioning	Định vị bóng
	24	mentality_vision	Tầm nhìn
	25	mentality_penalties	Đá phạt đền
	26	mentality_composure	Mức độ điềm tĩnh
PHÒNG THỦ	27	defending_marking_awareness	Nhận thức phòng thủ
	28	defending_standing_tackle	Tắc bóng/ Lấy lại bóng từ đối thủ
	29	defending_sliding_tackle	Xoạc/ Trượt bóng
THỦ MÔN	30	goalkeeping_diving	Bổ nhào
	31	goalkeeping_handling	Bắt bóng
	32	goalkeeping_kicking	Phát bóng xa
	33	goalkeeping_positioning	Định vị bóng
	34	goalkeeping_reflexes	Phản xạ

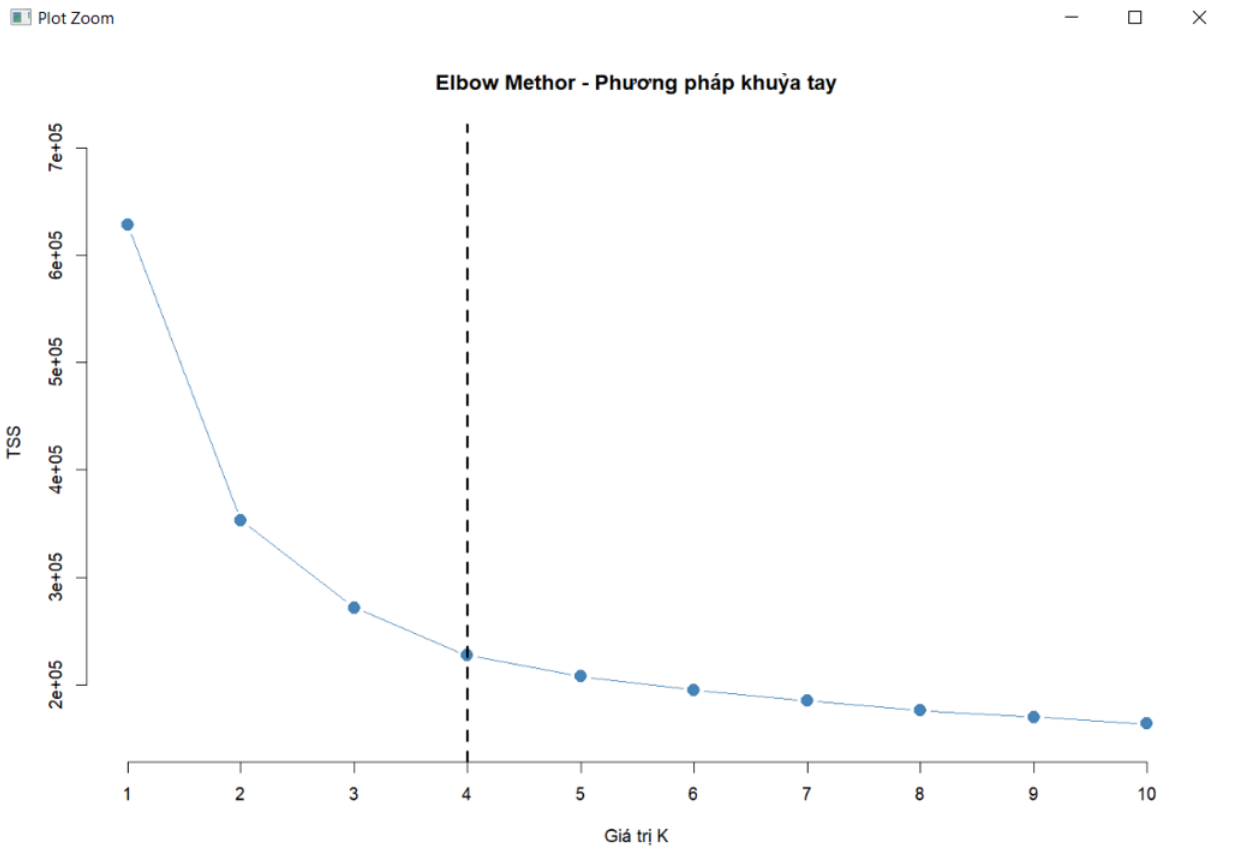
Kiểm tra xem dữ liệu có đầy đủ chưa, có ô nào trống không vì K-Mean không đọc được dữ liệu chứa ô trống ⇒ Kết quả: không có ô trống nào.

```
> names(which(colSums(is.na(k.Fifa))>0))
character(0)
```

Vì K-mean dùng phương pháp tính khoảng cách Euclide và phương sai mẫu hiệu chỉnh giữa các biến lệch nhau rất lớn, nên dù các biến cùng đơn vị đo (thang điểm 100), chúng ta vẫn nên chuẩn hóa dữ liệu. Ta có:

	attacking_crossing	attacking_finishing	attacking_heading_accuracy	attacking_short_passing
1	2.0926108	2.5247564	1.0231518	2.2688679
2	1.8742083	2.4736909	2.1140811	1.6555249
3	2.0380102	2.1162321	0.5638132	1.9281218
4	1.7104064	1.9630354	0.5063958	2.0644202
5	2.3656140	1.8609043	0.1618918	2.2688679
6	-2.0024369	-1.7647493	-2.1348014	-1.0704438
7	1.9834096	1.3502489	0.1618918	2.2688679
8	-1.7294337	-1.6115526	-2.3644708	0.1562421
9	0.1815885	0.3289380	1.9418291	1.3147788
10	1.6012051	2.2694287	0.3915612	1.7236741

Bước 2: Tìm chỉ số k tối ưu bằng phương pháp Elbow Method
Nhìn vào đồ thị, nhóm em chọn k = 4:



Bước 3: Thực hiện phân cụm dữ liệu bằng hàm kmeans có sẵn trên R-Studio.

```
kmeans(kk.Fifa,4, nstart=20)
```

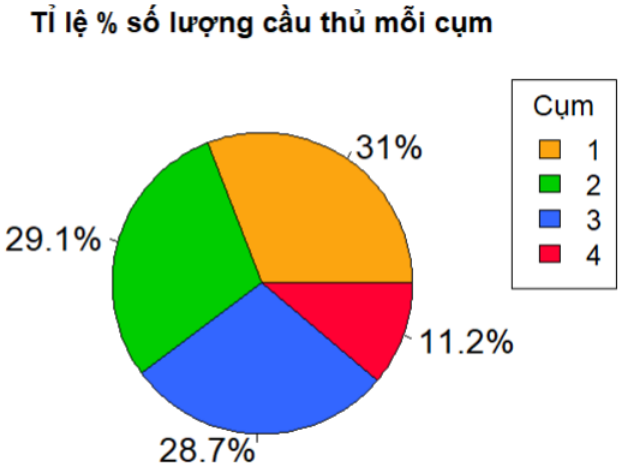
Kết quả phân cụm:

- Số lần lặp là 4, trong khi số lần lặp tối đa mặc định là 10
  $\Rightarrow$ thuật toán trên bộ dữ liệu này dừng lại khi không có sự phân phối lại.

```
iter                integer [1]                4
```

- Số lượng cầu thủ mỗi cụm:

```
  So.Luong
1     5730
2     5383
3     5309
4     2061
```



- Tâm các cụm:

	cluster	attacking_crossing	attacking_finishing	attacking_heading_accuracy	attacking_short_passing	attacking_volleys
1	1	0.8174334	0.5950670	0.401230925	0.81328146	0.6817676
2	2	0.1281949	0.7271316	0.008732731	0.05551851	0.5529799
3	3	-0.2584973	-0.7063392	0.405934424	-0.09503325	-0.6208263
4	4	-1.9415841	-1.7340753	-2.183972515	-2.16129423	-1.7405396

- Chất lượng phân cụm = 63.8% tức là chất lượng ở mức tốt vừa phải.

```
(between_SS / total_SS =  63.8 %)
```

- Kết quả phân cụm K-mean:

Hình 4.2: top 10 cầu thủ của cụm 1

	short_name	player_positions	cluster
1	L. Messi	RW, CF, ST	1
2	Cristiano Ronaldo	ST, LW	1
3	Neymar Jr	LW, CAM	1
4	E. Hazard	LW, CF	1
5	K. De Bruyne	CAM, CM	1
7	L. Modrić	CM	1
9	V. van Dijk	CB	1
10	M. Salah	RW, ST	1
11	G. Chiellini	CB	1
12	S. Agüero	ST	1

Hình 4.3: top 10 cầu thủ của cụm 2

	short_name	player_positions	cluster
412	S. Mandíquez	ST	2
541	K. Piątek	ST	2
565	Loren	ST	2
682	A. Iwobi	LM, LW, RM	2
689	L. Alario	ST	2
691	A. Saint-Maximin	RW, LW, ST	2
837	D. Ginczek	ST, RW	2
894	L. Pavoletti	ST	2
935	A. Mitrović	ST	2
947	M. Diagne	ST	2

Hình 4.4: top 10 cầu thủ của cụm 3

	short_name	player_positions	cluster
19	K. Koulibaly	CB	3
73	M. Škriniar	CB	3
86	K. Manolas	CB	3
99	N. Süle	CB	3
102	J. Giménez	CB	3
114	Sokratis	CB	3
144	Felipe	CB	3
163	Raúl Albiol	CB	3
214	A. Romagnoli	CB	3
219	J. Tah	CB	3

Hình 4.5: top 10 cầu thủ của cụm 4

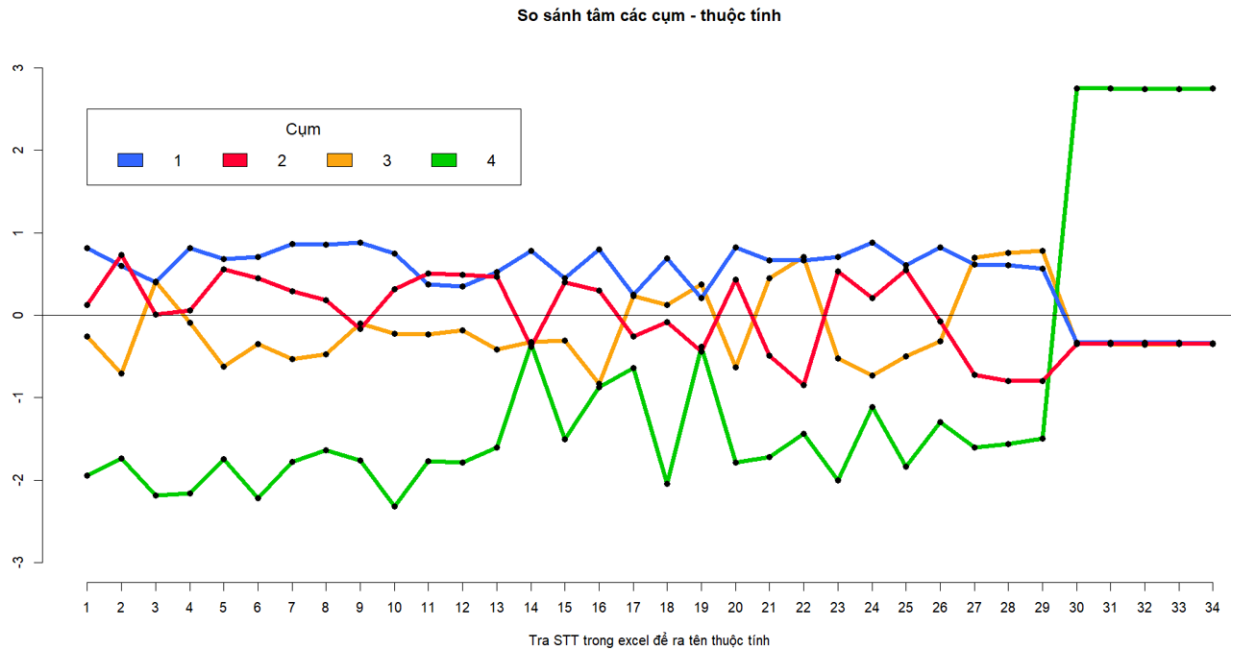
	short_name	player_positions	cluster
6	J. Oblak	GK	4
8	M. ter Stegen	GK	4
17	De Gea	GK	4
21	Alisson	GK	4
25	S. Handanovič	GK	4
26	M. Neuer	GK	4
27	H. Lloris	GK	4
35	T. Courtois	GK	4
39	Ederson	GK	4
50	K. Navas	GK	4

Nhận xét:
Trong bóng đá, có 4 nhóm cầu thủ cơ bản là Thủ môn, Hậu vệ, Tiền vệ, Tiền đạo.

Nếu nhìn sơ qua top 10 cầu thủ ở mỗi cụm, ta dễ dàng thấy, Cụm 4 là nhóm THủ môn, Cụm 3 là nhóm Hậu vệ, Cụm 2 là nhóm Tiền đạo.

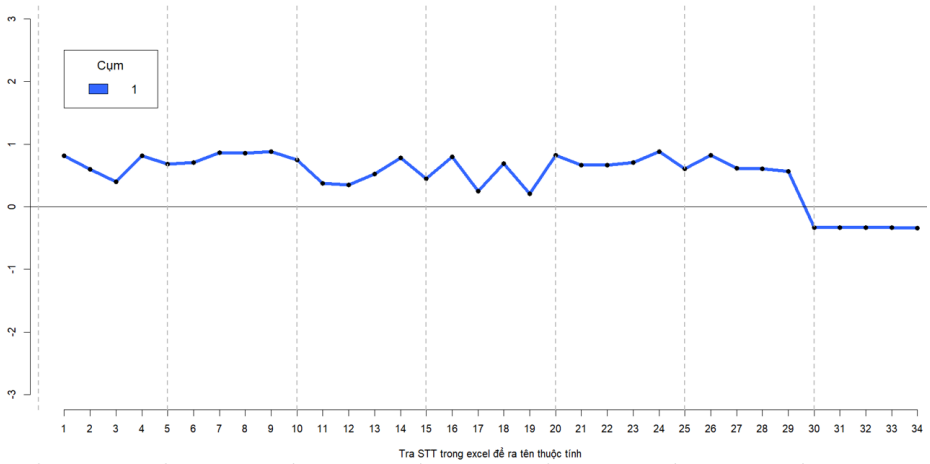
Xét riêng 10 cầu thủ đầu tiên trong cụm 1, không thể nói đây thuộc nhóm Tiền vệ, vì còn có cả các cầu thủ giữ vị trí Tiền đạo và Hậu vệ. Điều này cho thấy bộ dữ liệu này không phù hợp để chạy phương pháp K-Mean.

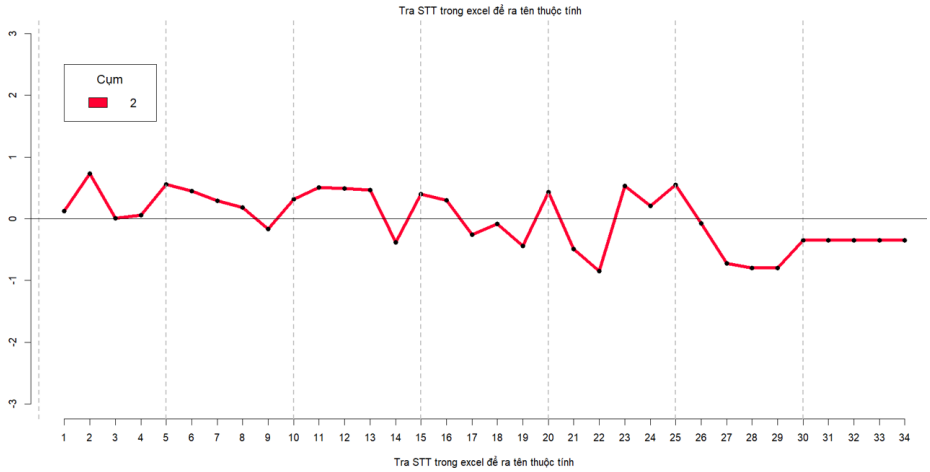
Tuy nhiên các đặc điểm của từng cụm vẫn rất rõ ràng. Ta có thể biểu diễn các tâm cụm qua những biểu đồ dưới đây.

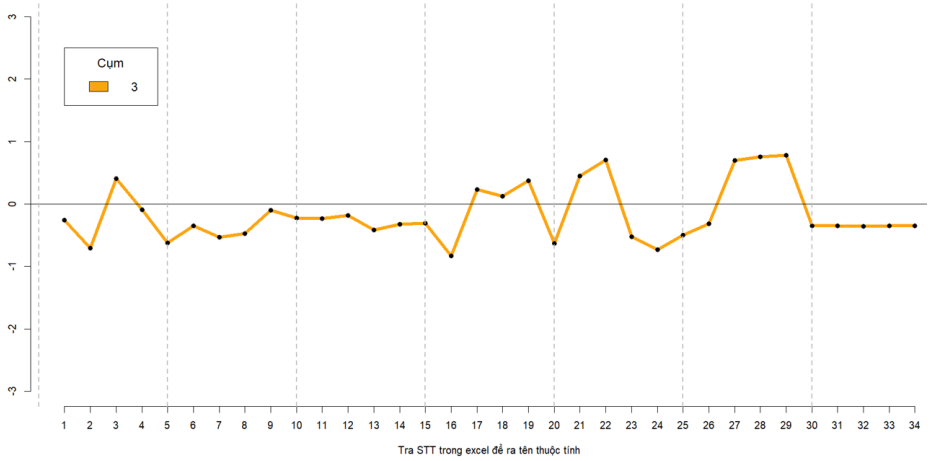


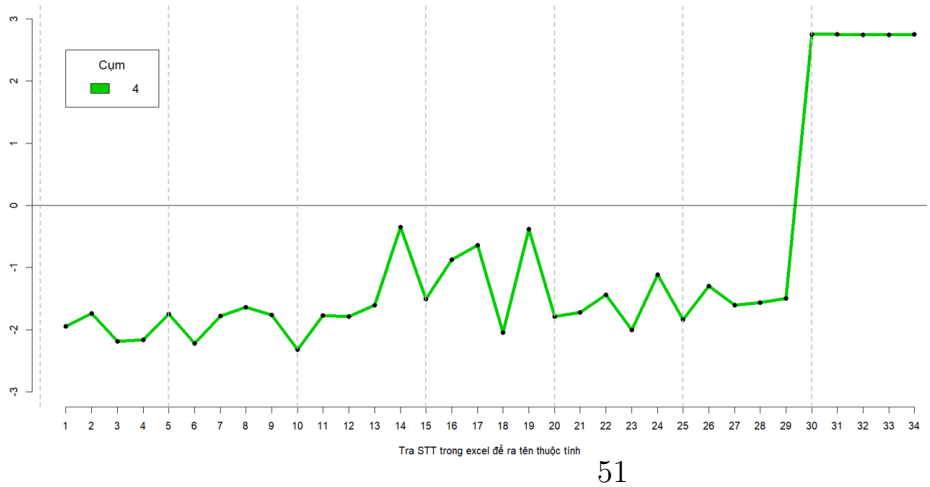
Dễ thấy nhất là đặc điểm của các cầu thủ **Cụm 4 (Xanh lá)**. Các kỹ năng dùng đến chân (1->29) của họ thấp hơn hẳn so với 3 cụm còn lại. Nhưng các kỹ năng chuyên môn của Thủ thành (30->34) lại cao vượt trội. Ngoài các kỹ năng cao vượt trội ra, các kỹ năng như Phản xạ(14) Nhảy cao(17) và Sức mạnh(19) cũng rất quan trọng với họ.

Trái ngược hoàn toàn với Cụm 4, Các kỹ năng từ 1->29 của các cầu thủ **Cụm 1 (Xanh da trời):** đều cao hơn hẳn so với trung bình của tổng thể (trung bình sau khi chuẩn hóa dữ liệu bằng 0). Ta có thể nói, họ mạnh cả về tấn công lẫn phòng thủ.
Cụm 2(Đỏ): chỉ số phòng thủ (27->29) của họ khá thấp. Nhưng khả năng dứt điểm(2) và tốc độ(11+12+13) lại vô cùng tốt, cao nhất so với các nhóm còn lại.
Cụm 3(Cam): chỉ số phòng thủ (27->29) của họ lại là cao nhất trong 4 cụm. Khả năng đánh đầu(3), nhảy cao(17), sức mạnh(19), tinh thần hiếu chiến(21) và chặn đường chuyền(22) cũng rất nổi bật


Cụm
1
Tra STT trong excel để ra tên thuộc tính


Cụm
2
Tra STT trong excel để ra tên thuộc tính


Cụm
3
Tra STT trong excel để ra tên thuộc tính


Cụm
4
Tra STT trong excel để ra tên thuộc tính

4.5 Ưu điểm & Nhược điểm của thuật toán K-mean

Ưu điểm:

- Nó rất dễ hiểu và dễ thực hiện.
- Nếu chúng ta có số lượng biến lớn thì K-mean sẽ nhanh hơn so với phân cụm phân cấp.
- Các cụm chặt chẽ hơn được hình thành với K-means so với phân cụm theo thứ bậc.

Nhược điểm:

- Hiện nay có nhiều phương pháp dùng trong Phân cụm, Phân tích số liệu. Phương pháp K-mean tuy đơn giản, dễ hiểu nhưng không phải phương pháp tốt nhất, phù hợp nhất với từng bộ dữ liệu.
- Đối với các tập dữ liệu có số chiều lớn dữ liệu có nhiều phần tử nhiễu như các tập dữ liệu biểu hiện gen thì giải thuật K-means thực hiện không đạt hiệu quả cao.
- Hiệu quả của thuật toán phụ thuộc vào việc chọn số nhóm K. Dễ thấy rằng điều kiện đầu vào của giải thuật bắt buộc phải chỉ rõ giá trị của K. Trong thực tế không phải lúc nào ta cũng có thể biết trước được có bao nhiêu nhóm cả và chi phí cho thực hiện vòng lặp tính toán khoảng cách lớn khi số cụm K và dữ liệu phân cụm lớn.
- Sự tồn tại của một phần tử ngoại vi có thể dẫn tới có ít nhất một nhóm với rất nhiều đối tượng bị phân tán.
- Thậm chí nếu tập được biết là có tồn tại K nhóm, sự ép buộc dữ liệu thành K nhóm sẽ dẫn đến các cụm vô nghĩa.
- Thường trong các thuật toán chạy đơn cần có K do người sử dụng định rõ. Tốt nhất là nên chạy thuật toán với một vài lựa chọn ngẫu nhiên khác nhau. Như vậy bạn hoàn toàn có thể thử xem dữ liệu của mình tốt với giá trị K nào.

Chương 5

Phân cụm dựa trên mô hình thống kê

5.1 Tổng quan

Những phương pháp phân cụm phổ biến đã được thảo luận trước trong chương này, bao gồm phân cụm liên kết đơn, liên kết hoàn chỉnh, liên kết trung bình, phương pháp Ward và phân cụm K-means, chúng đều là những quy trình hợp lý về mặt trực giác. Nhưng ta cũng chỉ có thể nói tới chúng như vậy khi mà không có mô hình để giải thích các quan sát đã được sinh ra như thế nào. Những lợi ích chính của phương pháp phân cụm đã được làm rõ qua phần giới thiệu về mô hình thống kê, điều đó chỉ ra bộ (p x 1) các phép đo x_j từ N đối tượng đã được tạo ra như thế nào. Mô hình phổ biến nhất là mô hình mà tại đó cụm k có tỉ lệ dự kiến p_k của các đối tượng và có các phép đo tương ứng được tạo ra bởi hàm mật độ xác suất $f_k(x)$. Ngoài ra, nếu có K cụm thì vector quan sát cho một đối tượng đơn lẻ được điều chỉnh theo sự phát sinh từ *phân phối hỗn hợp*

$$f_{mix}(x) = \sum_{k=1}^{K} p_k f_k(x) \tag{5.1.1}$$

Trong đó $p_k \geq 0$ và $\sum_{k=1}^{K} p_k = 1$. Phân phối $f_{mix}(x)$ được gọi là một hỗn hợp của K phân phối $f_1(x), f_2(x), ..., f_K(x)$ bởi vì quan sát được tạo ra từ phân phối thành phần $f_k(x)$ với xác suất p_k. Bộ các vector N quan sát được tạo ra từ phân phối này sẽ là một hỗn hợp các quan sát từ các phân phối thành phần.
Mô hình hỗn hợp phổ biến nhất là hỗn hợp của các phân phối chuẩn đa biến trong đó thành phần thứ k của $f_k(x)$ là hàm mật độ $N_p(\mu_k, \Sigma_k)$.
Mô hình hỗn hợp thông thường cho một quan sát $\mathbf{x}$:

$f_{mix}(x|\mu_1, \Sigma_1, ..., \mu_k, \Sigma_k)$

$$= \sum_{k=1}^{K} p_k \frac{1}{(2\pi)^{p/2}|\Sigma_k|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_k)' \Sigma_k^{-1}(x-\mu_k)\right) \quad (5.1.2)$$

Trong đó các tỉ lệ $p_1, ..., p_k$, các vector trung bình $\mu_1, ..., \mu_k$, các ma trận hiệp phương sai $\Sigma_1, ..., \Sigma_k$ chưa biết. Các phép đo cho đối tượng khác nhau được coi là độc lập và các quan sát phân bố giống hệt nhau từ phân phối hỗn hợp. Thông thường có quá nhiều tham số chưa biết để có thể thực hiện suy luận khi mà số lượng đối tượng được phân cụm vừa phải. Tuy nhiên, một số kết luận nhất định có thể được đưa ra liên quan đến các tình huống mà phương pháp phân cụm phân cấp sẽ hoạt động tốt hơn. Cụ thể, quy trình dựa trên khả năng xảy ra theo mô hình hỗn hợp thông thường với tất cả Σ_k có cùng bội số của ma trận đơn vị, $\eta\mathbf{I}$, xấp xỉ giống như phân cụm K-means và phương pháp Ward. Cho đến nay, không có mô hình thống kê nào được nâng cao mà ở đó quy trình hình thành cụm là xấp xỉ giống liên kết đơn, liên kết hoàn chỉnh hay liên kết trung bình.

Quan trọng hơn cả, theo trình tự các mô hình hỗn hợp (5.0.2) cho các K khác nhau, bài toán chọn số lượng cụm và phương pháp phân cụm thích hợp đã được rút gọn thành bài toán chọn một mô hình thống kê thích hợp. Một cách tiếp cận tốt để chọn một mô hình, trước tiên là thu được các ước lượng khả năng lớn nhất $\hat{p_1}, ..., \hat{p_K}, \hat{\mu}_1, \hat{\Sigma}_1, ..., \hat{\mu_K}, \hat{\Sigma_K}$ cho một số cụm K cố định. Những ước tính này phải được lấy bằng cách sử dụng những phần mềm chuyên dụng đặc biệt. Giá trị cực đại của hàm hợp lý:

$$L_{max} = L(\hat{p_1}, ..., \hat{p_K}, \hat{\mu}_1, \hat{\Sigma}_1, ..., \hat{\mu_K}, \hat{\Sigma_K}) \quad (5.1.3)$$

Giá trị này cung cấp nền tảng cho việc chọn lựa mô hình. Để có thể so sánh các mô hình với số lượng tham số khác nhau thì một hàm phạt (Penalty) đã được ghi lại:

$$-2\ln L_{max} - \text{Penalty}.$$

Tại đó hàm phạt phụ thuộc vào số tham số đã ước lượng và số quan sát N. Vì xác suất p_k có tổng bằng 1, nên chỉ có $K-1$ xác suất cần được ước tính, K x p giá trị trung bình và K x $p(p+1)/2$ phương sai và hiệp phương sai. Đối với tiêu chí thông tin Akaike (AIC), hàm phạt là $2N$ x (số tham số) nên ta có:

$$AIC = 2\ln L_{max} - 2N\left(K\frac{1}{2}(p+1)(p+2) - 1\right) \quad (5.1.4)$$

Tương tự với tiêu chí thông tin Bayesian (BIC) nhưng sử dụng hàm logarit số các tham số trong hàm phạt:

$$BIC = 2\ln L_{max} - 2\ln(N)\left(K\frac{1}{2}(p+1)(p+2) - 1\right) \tag{5.1.5}$$

Vẫn có rất nhiều khó khăn khi mà mô hình hỗn hợp có quá nhiều tham số cho nên các cấu trúc đơn giản đã được giả định cho Σ_k. Đặc biệt, các cầu trúc phức tạp dần được xác định như trong bảng dưới đây:

Giả định Σ_k	Số tham số	BIC
$\Sigma_k = \eta\mathbf{I}$	$K(p+1)$	$\ln L_{max} - 2\ln(N)\,(K(p+1))$
$\Sigma_k = \eta_k\mathbf{I}$	$K(p+2)-1$	$\ln L_{max} - 2\ln(N)\,(K(p+2)-1)$
$\Sigma_k = \eta_k\mathbf{Diag}(\lambda_1, \lambda_2, ..., \lambda_p)$	$K(p+2)+p-1$	$\ln L_{max} - 2\ln(N)\,(K(p+2)+p-1)$

Ngay cả đối với một số cụm cố định thì việc ước lượng mô hình hỗn hợp là rất phức tạp. Một gói phần mềm, MCLUST, đã được phát triển và có sẵn trong thư viện phần mềm R. Thư viện này kết hợp phương pháp phân cụm phân cấp, thuật toán EM và tiêu chí BIC để chọn ra được mô hình phù hợp với phân cụm. Ở bước 'E' của thuật toán EM, một ma trận (N x K) được tạo ra với hàng thứ j chứa các ước lượng của xác suất có điều kiện (dựa trên những ước lượng tham số hiện tại) mà quan sát x_j thuộc vào cụm $1, 2, ..., K$. Vì vậy, khi hội tụ, quan sát thứ j được gán cho cụm k mà tại đó xác suất có điều kiện của chúng là lớn nhất:

$$p(k|x_j) = \hat{p}_j f(x_j|k) / \sum_{i=1}^{K}(\hat{p}_i f(x_i|k)) \tag{5.1.6}$$

5.2 Ví dụ

Ta sẽ sử dụng bộ dữ liệu về bệnh nhân bị tiểu đường có sẵn ở trên R để xây dựng mô hình dự đoán tình trạng của bệnh nhân dựa trên những chỉ số đo lường được.

```
library(mclust)
data(diabetes)
dt <- diabetes
```

Bộ dữ liệu này bao gồm 145 quan sát và 4 tính chất.

	class	glucose	insulin	sspg
1	Normal	80	356	124
2	Normal	97	289	117
3	Normal	105	319	143
4	Normal	90	356	199
5	Normal	90	323	240
6	Normal	86	381	157
7	Normal	100	350	221
8	Normal	85	301	186
9	Normal	97	379	142
10	Normal	97	296	131
11	Normal	91	353	221
12	Normal	87	306	178
13	Normal	78	290	136

Showing 1 to 13 of 145 entries, 4 total columns

Tiếp đó, để phân cụm dữ liệu ta sẽ tính toán chỉ số BIC dựa trên 4 mô hình thống kê như hình dưới đây:

```
fit <- Mclust(dt, modelnames = c('EEI','VVI','EEE','VVV'))
```

Ý nghĩa của các mô hình:

- EEI: Phương sai bằng nhau và hiệp phương sai cố định bằng 0
- VVI: Phương sai biến thiên và hiệp phương sai cố định bằng 0
- EEE: Phương sai bằng nhau và hiệp phương sai bằng nhau
- VVV: Phương sai biến thiên và hiệp phương sai biến thiên

Chỉ số BIC càng lớn thì mô hình đó sẽ càng phù hợp. Dưới đây là top 3 chỉ số BIC lớn nhất sau khi tính toán.

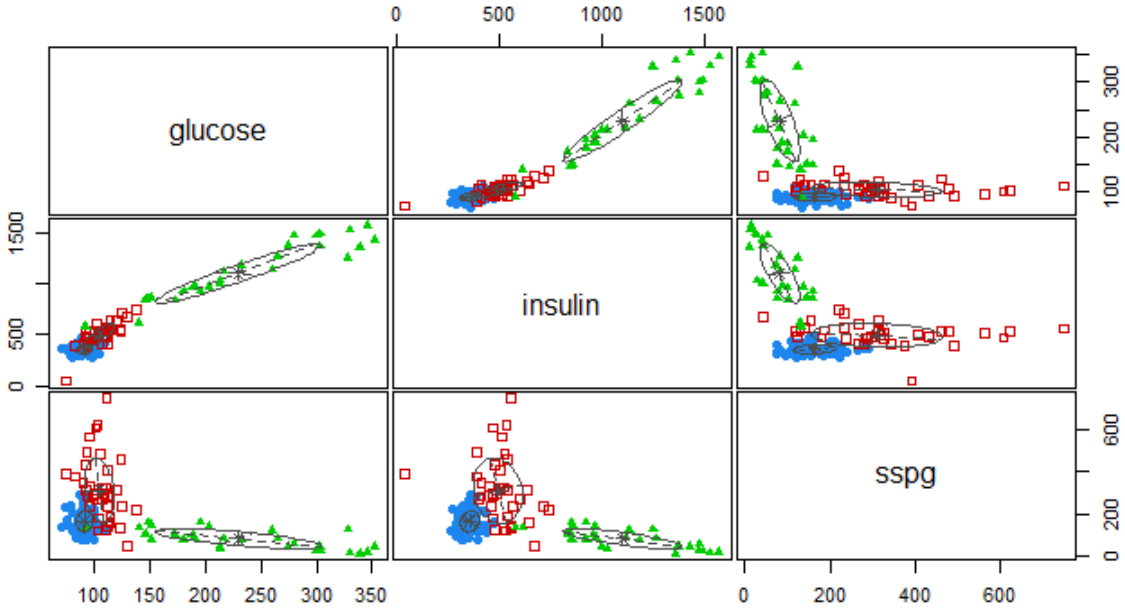
```
Top 3 models based on the BIC criterion:
    VVV,3     VVV,4     EVE,6
-4751.316 -4784.322 -4785.246
```

Ta thấy rằng mô hình VVV cho ra chỉ số BIC lớn nhất khi phân ra làm 3 cụm nên ta sẽ phân dữ liệu ra làm 3 cụm.
Kết quả sau khi phân cụm dữ liệu và biểu diễn đồ thị của chúng:

```
[[1]]
 [1] Normal   Normal   Normal   Normal   Normal   Normal   Normal   Normal
 [9] Normal   Normal   Normal   Normal   Normal   Normal   Normal   Normal
[17] Normal   Normal   Normal   Normal   Normal   Normal   Normal   Normal
[25] Normal   Normal   Normal   Normal   Normal   Normal   Normal   Normal
[33] Normal   Normal   Normal   Normal   Normal   Normal   Normal   Normal
[41] Normal   Normal   Normal   Normal   Normal   Normal   Normal   Normal
[49] Normal   Normal   Normal   Normal   Normal   Normal   Normal   Normal
[57] Normal   Normal   Chemical Normal   Normal   Chemical Chemical Normal
[65] Chemical Chemical Normal   Normal   Normal   Normal   Normal   Normal
[73] Chemical Normal   Normal   Normal   Normal   Chemical Normal   Chemical
[81] Chemical
Levels: Chemical Normal Overt

[[2]]
 [1] Normal   Chemical Normal   Normal   Normal   Chemical Chemical Chemical
 [9] Chemical Chemical Chemical Chemical Chemical Chemical Chemical Chemical
[17] Chemical Chemical Chemical Chemical Chemical Chemical Chemical Chemical
[25] Chemical Chemical Chemical Chemical Chemical Chemical Overt    Overt
[33] Overt    Overt    Overt    Overt
Levels: Chemical Normal Overt

[[3]]
 [1] Chemical Overt    Overt    Overt    Overt    Overt    Overt    Overt
 [9] Overt    Overt    Overt    Overt    Overt    Overt    Overt    Overt
[17] Overt    Overt    Overt    Overt    Overt    Overt    Overt    Overt
[25] Overt    Overt    Overt    Overt
Levels: Chemical Normal Overt
```



Từ dữ kiện trên khi đem so sánh với dữ liệu gốc ta có thể thấy là mô hình này dự đoán tương đối chính xác tình trạng của bệnh nhân dựa trên các chỉ số cho trước.

Chương 6

Thuật toán chia tỷ lệ đa chiều

Trong khi sử dụng các phương pháp phân cụm, dữ liệu có chiều càng lớn thì độ chính xác càng giảm do các thành phần nhiễu, vì vậy người ta xây dựng các phương pháp giảm chiều dữ liệu để xử lý dữ liệu trước khi chạy phân cụm đưa ra các kết quả ổn định hơn. Đồng thời để hiển thị dữ liệu đa biến trong không gian chiều thấp giúp cho việc quan sát đánh giá tương quan giữa các đối tượng dễ dàng hơn.

Một số phương pháp:

- Phương pháp PCA (Phân tích các thành phần chính)
- Phương pháp MDS (Chia tỷ lệ đa chiều)

6.1 Giới thiệu thuật toán chia tỷ lệ đa chiều

Chia tỷ lệ đa chiều (Multidimentional Scaling) là một phương pháp kỹ thuật dùng để giảm chiều dữ liệu hay nén dữ liệu, kỹ thuật này chuyển dữ liệu từ không gian nhiều chiều về không gian ít chiều hơn để xử lý.

Đặc biệt phương pháp chia tỷ lệ đa chiều chỉ dựa vào dữ liệu đầu vào là khoảng cách giữa các đối tượng hoặc các điểm (có nhiều loại khoảng cách).

Mục tiêu chính của chúng ta là biểu diễn dữ liệu ban đầu vào một hệ tọa độ có số chiều thấp sao cho giảm thiểu sự biến dạng khi chúng ta giảm số chiều gây ra, đưa ra khoảng cách rõ nhất giữa các giá trị để phục vụ cho các thuật toán phân cụm ở trước.

Các kỹ thuật chia tỷ lệ được phát triển bởi Shepard, Kruskal, và những người khác.

6.2 Ý tưởng thuật toán

Đối với N đối tượng, có $M = N(N-1)/2$ điểm tương đồng (khoảng cách) giữa các cặp đối tượng khác nhau. Những điểm tương đồng này tạo thành dữ liệu cơ bản. (Trong trường hợp không thể dễ dàng định lượng được các điểm tương đồng, chẳng hạn như sự giống nhau giữa hai màu, thứ tự xếp hạng của các điểm tương đồng là dữ liệu cơ bản). Giả sử không có mối liên hệ nào, các điểm tương đồng có thể được sắp xếp theo một thứ tự tăng dần như sau:

$$s_{i_1k_1} < s_{i_2k_2} < \cdots < s_{i_Mk_M} \tag{6.2.1}$$

Ở đây $s_{i_1k_1}$ là điểm tương đồng nhỏ nhất. Chỉ số phụ i_1k_1 chỉ ra các cặp đối tượng ít tương đồng nhất - nghĩa là các đối tượng có xếp hạng đầu tiên trong thứ tự tương đồng. Các chỉ số phụ khác được giải thích theo cách tương tự. Ta cần tìm 1 cấu hình q - chiều của N phần tử sao cho khoảng cách, $d_{ik}^{(q)}$, giữa các cặp phần tử khớp với thứ tự trong (6.2.1). Nếu các khoảng cách được sắp xếp tương ứng với thứ tự đó, một kết hợp hoàn hảo xảy ra khi

$$d_{i_1k_1}^{(q)} > d_{i_2k_2}^{(q)} > \cdots > d_{i_Mk_M}^{(q)} \tag{6.2.2}$$

Với $d_{ik} = \sqrt{s_{ii} + s_{kk} - 2s_{ik}}$
d_{ik}: điểm khác biệt (dissimilarity),
s_{ik}: điểm tương đồng (similarity).

Có nghĩa là, thứ tự giảm dần của các khoảng cách trong q-chiều hoàn toàn tương tự với thứ tự tăng dần của các điểm tương đồng ban đầu. Miễn là thứ tự trong (6.2.2) được giữ nguyên, độ lớn của các khoảng cách là không quan trọng.

Đối với một giá trị cho trước của q, có thể không tìm được cấu hình của các điểm có khoảng cách theo cặp phần tử là đơn điệu liên quan đến các điểm tương đồng lúc đầu. Kruskal đã đề xuất một thước đo cho mức độ một biểu diễn hình học thiếu phù hợp. Phép đo này được định nghĩa là:

Stress

$$\text{Stress}(q) = \left\{ \frac{\sum\sum_{i<k} \left(d_{ik}^{(q)} - \widehat{d}_{ik}^{(q)}\right)^2}{\sum\sum_{i<k} \left[d_{ik}^{(q)}\right]^2} \right\}^{1/2} \tag{6.2.3}$$

Một số giá trị *Stress* có thể được sử dụng để đánh giá độ phù hợp hay ảnh hưởng của kỹ thuật MDS được sử dụng. Một *Stress* có giá trị nhỏ chứng tỏ

rằng phương pháp mà ta sử dụng là hợp lý (tốt). Kruskal đề nghị rằng *Stress* được giải thích như sau:

Stress	Goodness of fit
20%	Poor
10%	Fair
5%	Good
2.5%	Excellent
0%	Perfect

Goodness of fit (Mức độ phù hợp): đề cập đến mối quan hệ đơn điệu giữa các điểm tương đồng và khoảng cách cuối cùng.

Một thước đo khác thứ hai, được giới thiệu bởi Takane và cộng sự , trở thành tiêu chí được ưa thích. Đối với một kích thước q nhất định, số đo này, ký hiệu là $SStress$, thay thế các d_{ik} và $\widehat{d}_{ik}$ trong (6.2.3) bằng các bình phương của chúng và được cho bởi:

SStress

$$\text{SStress}(q) = \left\{ \frac{\sum\sum_{i<k}\left(d_{ik}^2 - \widehat{d}_{ik}^2\right)^2}{\sum\sum_{i<k} d_{ik}^4} \right\}^{1/2} \tag{6.2.4}$$

Giá trị của SStress luôn nằm trong khoảng từ 0 đến 1. Bất kỳ giá trị điểm của đối tượng nào nhỏ hơn 0.1 thường được coi là có biểu hiện tốt.

Khi các phần tử được định vị trong q - chiều, vectơ tọa độ $q \times 1$ của chúng có thể được coi là các quan sát đa biến. Với mục đích hiển thị, thật thuận tiện để biểu diễn biểu đồ phân tán q - chiều này theo các trục thành phần chính của nó.

Với mỗi q, cấu hình sẽ dẫn đến theo giá trị $Stress$ nhỏ nhất mà nó có thể đạt được. Khi q tăng, $Stress$ nhỏ nhất trong phạm vi sai số làm tròn sẽ giảm và sẽ bằng 0 đối với $q = N - 1$. Bắt đầu với $q = 1$, có thể xây dựng đồ thị của các số $Stress(q)$ này so với q. Giá trị của q mà biểu đồ này bắt đầu chững lại có thể được chọn là giá trị "tốt nhất" của kích thước. Và việc biểu diễn ở số chiều cao hơn q là không cần thiết.

6.3 Các bước thuật toán

Toàn bộ thuật toán MDS được tóm tắt trong các bước sau:

Bước 1. Với N đối tượng, thu được $M = N(N-1)/2$ điểm tương đồng (khoảng cách) giữa các cặp đối tượng khác nhau. Thứ tự các điểm tương đồng như trong (6.2.1). (Khoảng cách được sắp xếp từ lớn nhất đến nhỏ nhất). Nếu các điểm tương đồng (khoảng cách) không thể tính toán được, thứ tự xếp hạng phải được chỉ định.

Bước 2. Sử dụng cấu hình thử nghiệm trong q - chiều, xác định khoảng cách $d_{ik}^{(q)}$ và số $\widehat{d}_{ik}^{(q)}$ (thỏa mãn (6.2.2) và giảm thiểu *Stress* (6.2.3) hoặc *SStress* (6.2.4)). ($\widehat{d}_{ik}^{(q)}$ thường được xác định trong các chương trình chia tỷ lệ trên máy tính bằng cách sử dụng các phương pháp hồi quy được thiết kế để tạo ra các khoảng cách đơn điệu "phù hợp").

Bước 3. Sử dụng $\widehat{d}_{ik}^{(q)}$, di chuyển các điểm xung quanh để có được cấu hình cải tiến. (Với q cố định, cấu hình cải tiến được xác định bằng quy trình giảm thiểu hàm tổng quát áp dụng cho $Stress$. Lúc này $Stress$ được coi là hàm của tọa độ $N \times q$ của N đối tượng. Một cấu hình mới sẽ có $d_{ik}^{(q)}$ mới, $\widehat{d}_{ik}^{(q)}$ mới và $Stress$ nhỏ hơn. Quá trình được lặp lại cho đến khi đạt được biểu diễn tốt nhất ($Stress$ tối thiểu).

Bước 4. Vẽ đồ thị $Stress(q)$ nhỏ nhất so với q và chọn số chiều tốt nhất, $q*$, từ việc kiểm tra biểu đồ này.

Giả định rằng các giá trị tương đồng lúc đầu là đối xứng ($s_{ik} = s_{ki}$), không có ràng buộc và không có quan sát nào bị thiếu. Kruskal đã đề xuất các phương pháp để xử lý sự bất đối xứng, ràng buộc và các quan sát bị thiếu. Ngoài ra, hiện nay có các chương trình máy tính chia tỷ lệ đa chiều sẽ xử lý không chỉ khoảng cách Euclide, mà bất kỳ khoảng cách nào thuộc loại Minskowki.

6.4 Ví dụ Excel

Dữ liệu đầu vào

Dữ liệu là một bảng gồm các giá trị khoảng cách giữa các cặp thành phố ở nước Mỹ.

	Atlanta	Boston	Chicago	Washington	Denver	Los Angeles	Miami	NYC	Seattle	San Francisco	New Orleans
Atlanta	0	934	585	542	1209	1942	605	751	2181	2139	424
Boston	934	0	853	392	1769	2601	1252	183	2492	2700	1356
Chicago	585	853	0	598	918	1748	1187	720	1736	1857	830
Washington	542	392	598	0	1493	2305	922	209	2328	2442	964
Denver	1209	1769	918	1493	0	836	1723	1636	1023	951	1079
Los Angeles	1942	2601	1748	2305	836	0	2345	2461	957	341	1679
Miami	605	1252	1187	922	1723	2345	0	1092	2733	2594	669
NYC	751	183	720	209	1636	2461	1092	0	2412	2577	1173
Seattle	2181	2492	1736	2328	1023	957	2733	2412	0	681	2101
San Francisco	2139	2700	1857	2442	951	341	2594	2577	681	0	1925
New Orleans	424	1356	830	964	1079	1679	669	1173	2101	1925	0

Chạy MDS

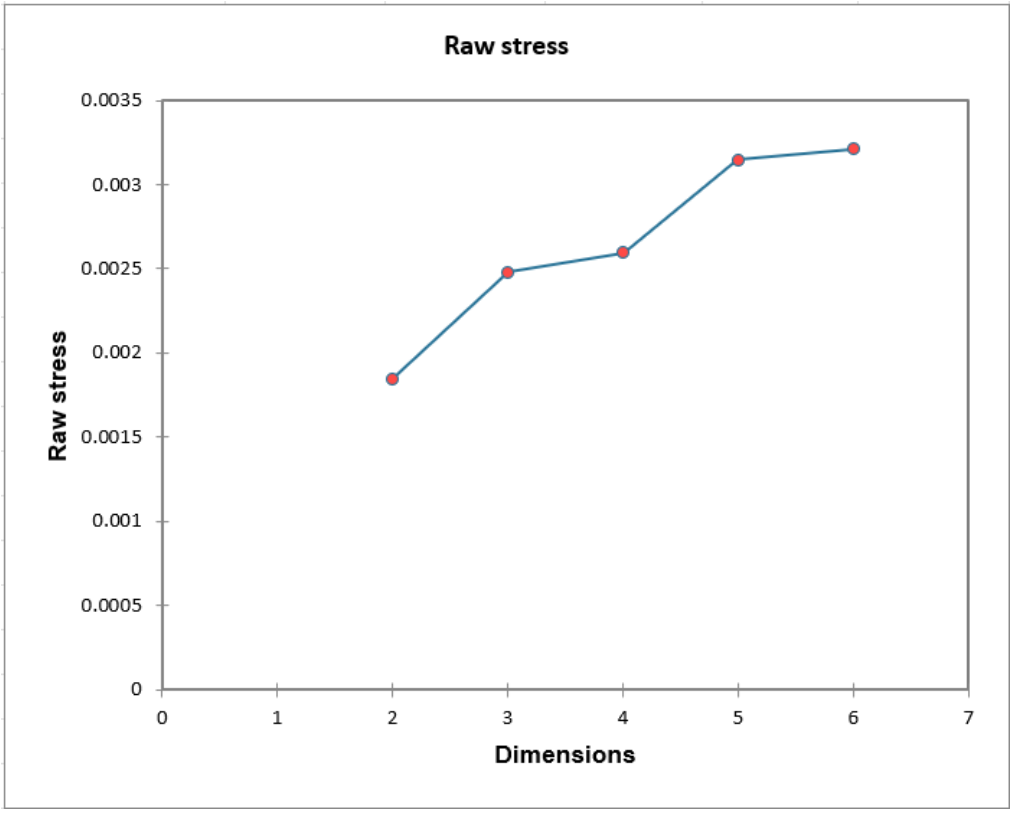
Sử dụng chức năng MDS có sẵn trên công cụ XLSTAT của Excel.
Sau khi chạy dữ liệu với số chiều từ 2 đến 6 ta có những đánh giá sau đây:

Dimensions	2	3	4	5	6
Kruskal's stress	0.002	0.002	0.003	0.003	0.003
Iterations	81	145	158	206	214

Theo bảng kết quả trên ta thấy:

- Với 2 chiều, số lần lặp là 81 và Stress(2) = 0.002 (Perfect).
- Với 3 chiều, số lần lặp là 145 và Stress(3) = 0.002 (Perfect).
- Với 4 chiều, số lần lặp là 158 và Stress(4) = 0.003 (Excellent).
- Với 5 chiều, số lần lặp là 206 và Stress(5) = 0.003 (Excellent).
- Với 6 chiều, số lần lặp là 214 và Stress(6) = 0.003 (Excellent).

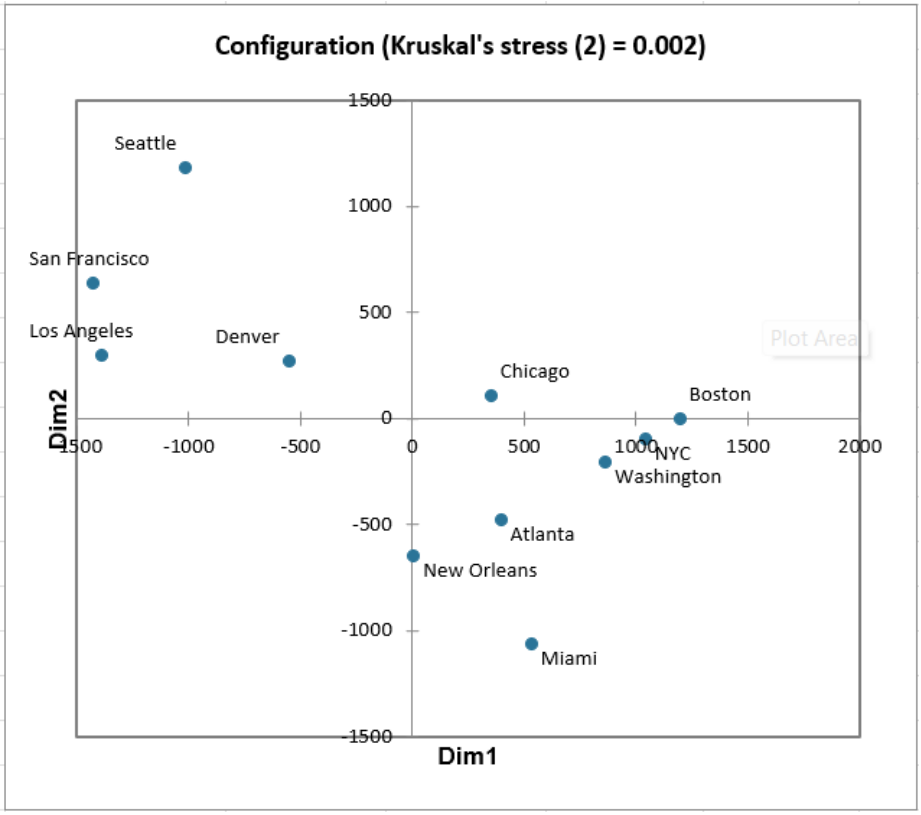
Biểu đồ Elbow:



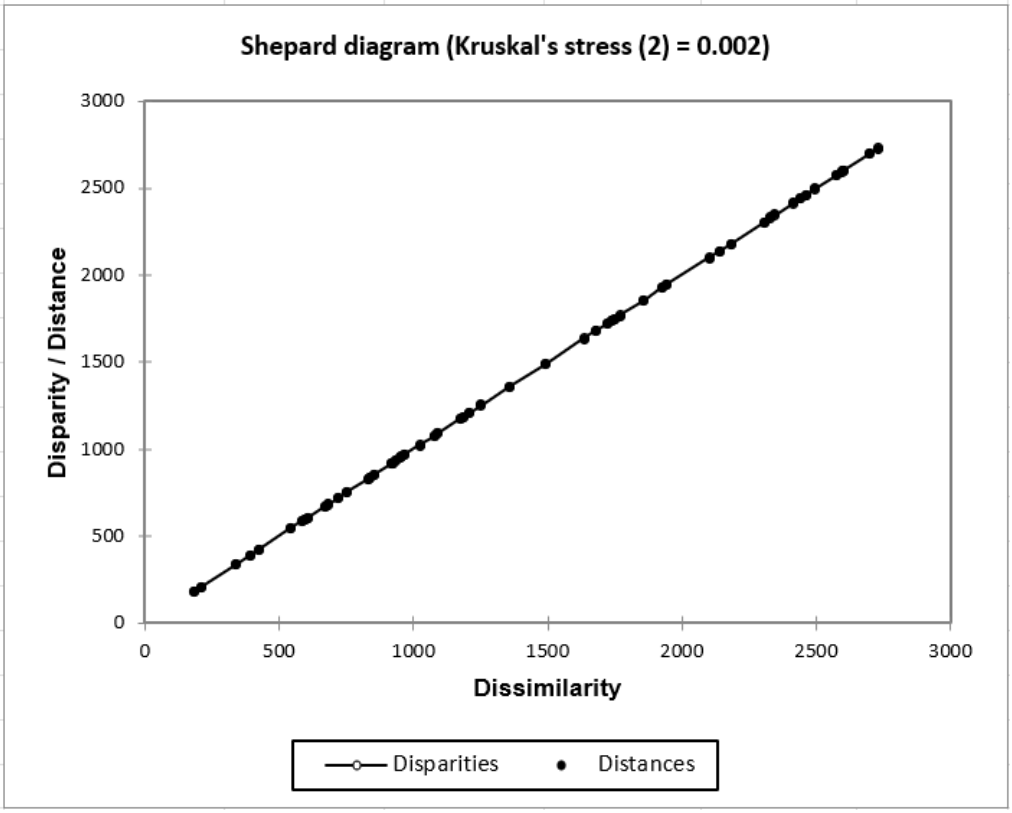
Bảng tọa độ dữ liệu tốt nhất khi đưa về 2 chiều:

Configuration:		
	Dim1	Dim2
Atlanta	394.606	-479.329
Boston	1194.301	1.632
Chicago	350.167	105.610
Washington	861.769	-205.026
Denver	-551.585	273.354
Los Angeles	-1385.723	297.631
Miami	531.152	-1065.525
NYC	1040.410	-97.163
Seattle	-1011.728	1182.062
San Francisco	-1428.950	636.867
New Orleans	5.581	-650.115

Đồ thị biểu diễn các thành phố trên toạ độ 2 chiều:



Sơ đồ Shepard:



Sơ đồ Shepard hiển thị mối quan hệ giữa các khoảng cách lân cận (proximities) và khoảng cách giữa các điểm thiết lập (đánh giá khoảng cách trước và sau khi chạy). Một đường hồi qui sẽ hiển thị thể hiện của khoảng cách lân cận (proximities) và khoảng cách xấp xỉ. Nếu có nhiều điểm nằm ngoài đường hồi qui nên suy ra tỷ lệ chênh lệch là lớn.

Trong hình trên các điểm này nằm hết trên đường hồi quy thì phương pháp được đánh giá là phù hợp.

Tiếp theo là bảng tọa độ dữ liệu tốt nhất khi đưa về 3 chiều:

Configuration:

	Dim1	Dim2	Dim3
Atlanta	-270.094	-555.914	82.620
Boston	137.336	-780.463	892.114
Chicago	173.370	-213.836	253.914
Washington	-57.189	-674.151	569.139
Denver	188.687	476.458	-354.653
Los Angeles	13.025	1108.761	-881.244
Miami	-751.669	-920.433	17.777
NYC	37.421	-726.939	747.638
Seattle	717.130	1351.750	-276.677
San Francisco	245.101	1336.614	-772.658
New Orleans	-433.118	-401.847	-277.968

Bảng tọa độ dữ liệu khi đưa về 4 chiều:

	Dim1	Dim2	Dim3	Dim4
Atlanta	418.429	442.513	-125.226	-3.804
Boston	238.041	549.093	-1001.211	-249.734
Chicago	-13.721	112.929	-354.816	-7.300
Washington	300.649	526.601	-630.726	-135.854
Denver	-295.306	-416.539	326.655	139.865
Los Angeles	-458.710	-832.943	1035.273	176.961
Miami	953.221	708.789	-25.333	32.868
NYC	259.829	550.793	-824.151	-200.254
Seattle	-1098.912	-1033.449	377.570	-11.018
San Francisco	-765.152	-961.457	959.011	118.515
New Orleans	461.631	353.671	262.955	139.753

Nhận xét

- Với tập dữ liệu Input đã cho, phương pháp MDS đều đưa ra các kết quả rất tốt cho từng chiều.

6.5 Ứng dụng của Multi-Dimensional Scale

Như đã tìm hiểu ở trên, Multidimensional Scaling (MDS) là một phương pháp thống kê được sử dụng để biểu diễn dữ liệu đa chiều trong không gian ít chiều hơn.
Một số ứng dụng của MDS trong thực tế bao gồm:

1. Khảo sát mối quan hệ giữa các biến: MDS có thể giúp xác định mức độ tương đồng giữa các biến bằng cách sử dụng các thông tin về khoảng cách hoặc độ tương tự giữa chúng. Ví dụ, MDS có thể được sử dụng để khảo sát mối quan hệ giữa các loại sản phẩm trong một thị trường cụ thể.

2. Phân tích mối quan hệ giữa các đối tượng: MDS cũng có thể được sử dụng để phân tích mối quan hệ giữa các đối tượng dựa trên các thông tin về khoảng cách hoặc độ tương tự giữa chúng. Ví dụ, MDS có thể được sử dụng để khảo sát mối quan hệ giữa các quốc gia trong một lĩnh vực cụ thể như kinh tế, chính trị hoặc văn hóa.

3. Xác định vị trí trong không gian đa chiều: MDS có thể được sử dụng để giảm số chiều của dữ liệu và tạo ra một biểu đồ hoặc một hệ thống các vị trí trong không gian ít chiều hơn. Ví dụ, MDS có thể được sử dụng để biểu diễn mối quan hệ giữa các trang web hoặc các sản phẩm trên một trang web dựa trên các thông tin về độ tương tự hoặc khoảng cách giữa chúng.

4. Phân tích đa biến: MDS có thể được sử dụng để phân tích các biến đa chiều trong không gian ít chiều hơn. Ví dụ, MDS có thể được sử dụng để phân tích các biến về sở thích của người tiêu dùng về các sản phẩm khác nhau và đánh giá mức độ tương đồng giữa các sản phẩm.

Tài liệu tham khảo

[1] ThS. Lê Xuân Lý, 'Bài giảng Xác Suất Thống Kê'.

[2] Johnson Wichern, 'Applied multivariate statistical analysis'.

[3] Jing Gao, 'Clustering - Hierarchical Methods'.

[4] Rebecca C. Steorts, 'K-means Clustering'.

[5] Bettina Grün, 'Model-based Clustering'.