Ảnh có chứa quảng trường

Mô tả được tạo tự động**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG**

-----🙞🙜🕮🙞🙜-----

****

**TRÍ TUỆ NHÂN TẠO**

*Đề tài: Tìm hiểu về phân cụm dữ liệu*

|  |  |
| --- | --- |
| Giảng viên hướng dẫn: | Thầy Đỗ Tiến Dũng |
| Sinh viên thực hiện: | Mai Trần Duy |
|  | Sái Văn Lượng |
|  | Nguyên Công Hoàng |

Hà Nội, Ngày 19/12/2021

**MỤC LỤC**

[**LỜI NÓI ĐẦU** 3](#_Toc91191112)

[**PHẦN I: GIỚI THIỆU** 3](#_Toc91191113)

[1.1 Phân cụm. 3](#_Toc91191114)

[1.1.1 Định nghĩa. 3](#_Toc91191115)

[1.1.2 Quá trình phân cụm. 4](#_Toc91191116)

[1.2 Các phương pháp phân cụm. 6](#_Toc91191117)

[1.2.1 Các yêu cầu tiêu biểu về việc phân cụm dữ liệu. 6](#_Toc91191118)

[1.2.2 Các Phương pháp phân cụm dữ liệu tiêu biểu. 6](#_Toc91191119)

[**PHẦN II: THUẬT TOÁN K-MEANS** 9](#_Toc91191120)

[2.1 Giới thiệu thuật toán K-means. 9](#_Toc91191121)

[2.2 Thuật toán K-means. 9](#_Toc91191122)

[2.2.1 Ví dụ 1. 10](#_Toc91191123)

[2.2.2 Ví dụ 2. 11](#_Toc91191124)

[2.3 Xây dựng và demo thuật toán K-Means. 18](#_Toc91191125)

[2.4 Ứng dụng của thuật toán K-Means 20](#_Toc91191126)

[**TÀI LIỆU THAM KHẢO** 24](#_Toc91191127)

# **LỜI NÓI ĐẦU**

Đây là tài liệu dùng để biểu diễn cơ sở bản thiết kế và giải quyết bài toán phân cụm dữ liệu sử dụng thuật toán K-means do nhóm tôi thiết kế và lập trình. Tài liệu này giúp ta có thể có cài nhìn toạn vẹn về các chức năng của phần mềm cũng như ứng dụng hiệu quả thuật toán K-means để giải quyết bài toán này. Do thời gian hạn chế nên đồ án không thể tối ưu được toàn bộ không gian trạng thái bài toán. Tuy nhiên, nhóm sẽ nghiêm cứu hoàn thiện trong thời gian sớm nhất.

Nhóm thực hiện đề toàn nhằm mục đích xây dựng phần mềm giải quyết bài toán thực tế về phân cụm dữ liệu dựa trên thuật toán học máy đơn giản K-means. Trong quá trình thực hiện đề tài không tránh khỏi những sai sót, nhóm tôi mong sẽ nhận được sự góp ý và đánh giá của thầy.

**PHÂN CÔNG THÀNH VIÊN TRONG NHÓM**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Họ và tên** | **Email** | **Điện thoại** | **Công việc** | **Đánh giá** |
| Mai Trần Duy | Duy.mt198223@sis.hust.edu.vn | 0359093620 | Viết chương trình, kiểm thử | Tốt |
| Sái Văn Lượng | Luong.sv198244@sis.hust.edu.vn | 0969274295 | Thiết kế giao diện, tham gia hoàn thiện báo cáo | Tốt |
| Nguyễn Công Hoàng | Hoang.nc198226@sis.hust.edu.vn | 0966429132 | Tham Khảo tài liệu, tham gia hoàn thiện báo cáo | Tốt |

# **PHẦN I: GIỚI THIỆU**

* 1. **Phân cụm.**
     1. **Định nghĩa.**

Phân cụm nhìn từ góc độ tự nhiên là một việc hết sức bình thường mà chúng ta vẫn làm và thực hiện hàng ngày ví dụ như phân loại học sinh khá, giỏi trong lớp, phân loại đất đai, phân loại tài sản, phân loại sách trong thư viện… Việc phân loại này là thực hiện gom các đối tượng có cùng tính chất hay có các tính chất gần giống nhau thành nhóm. Để thực hiện phân loại các đối tượng nào đó, chúng ta bao giờ cũng đặt câu hỏi, chúng ta phân nhóm dựa trên yếu tố nào? Hoặc chúng ta định phân thành bao nhiêu nhóm?

Một sô ví dụ minh họa:

Ví dụ về phân cụm ảnh

Ảnh có chứa cỏ, khác nhau, ngoài trời, bó

Mô tả được tạo tự động

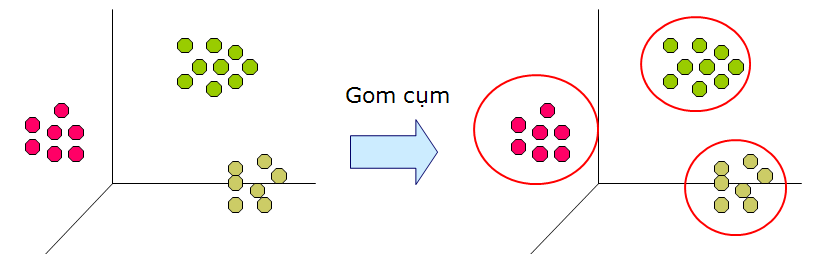
Hay trường hợp tổng quát

Ta phân hoạch các nhóm phần tử trong 1 tập hợp xác định vào các cụm khác nhau theo thuộc tính chung của các phần tử.

Phân cụm

### **1.1.2 Quá trình phân cụm.**

Là quá trình ta phân hoạch các đối tượng vào các cụm hoặc nhóm. Các đối tượng trong một cụm, nhóm có đặc điểm giống nhau nhất so với các phần tử của nhóm,cụm khác.

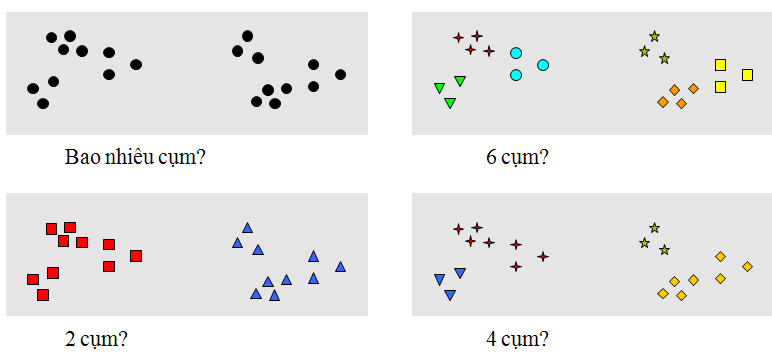


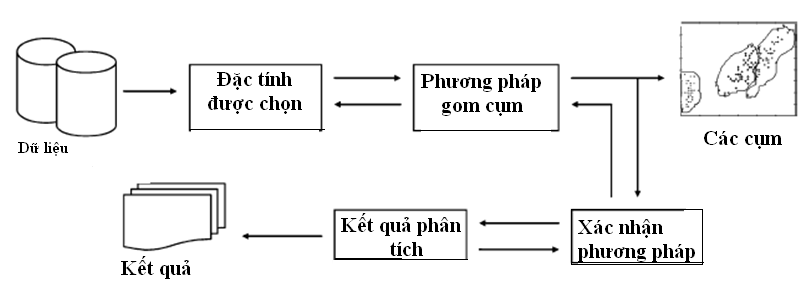
Trước khi thực hiện một quá trình phân cụm thì ta cần phải trả lời những câu hỏi sau:

- Mỗi cụm nên có bao nhiêu phần tử?

- Các phần tử nên được phân vào bao nhiêu cụm?

- Bao nhiêu cụm nên được tạo ra?



****Quá trình phân cụm có thể được minh họa qua sơ đồ sau:

## **1.2 Các phương pháp phân cụm.**

### **1.2.1 Các yêu cầu tiêu biểu về việc phân cụm dữ liệu.**

- Khả năng co giãn về tập dữ liệu (scalability).

- Khả năng xử lý nhiều kiểu thuộc tính khác nhau (different types of attributes).

- Khả năng khám phá các cụm với hình dạng tùy ý (clusters with arbitrary shape).

- Tối thiểu hóa yêu cầu về tri thức miền trong việc xác định các thông số nhập (domain knowledge for input parameters).

- Khả năng xử lý dữ liệu có nhiễu (noisy data).

- Khả năng phân cụm cụm tăng dần và độc lập với thứ tự của dữ liệu nhập (incremental clustering and insensitivity to the order of input records).

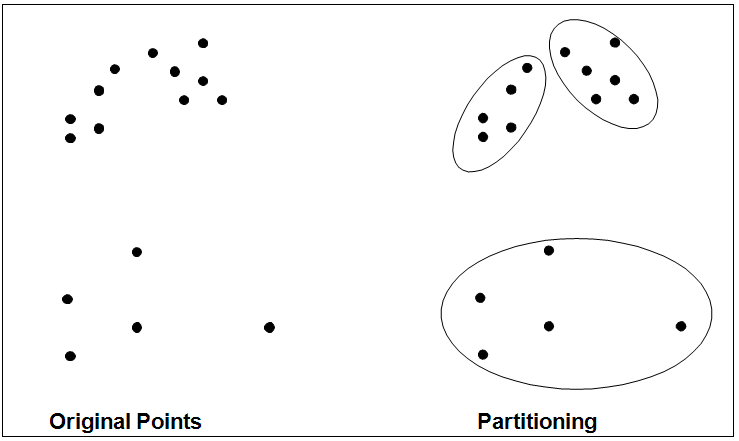
- Khả năng xử lý dữ liệu đa chiều (high dimensionality).

- Khả năng phân cụm dựa trên ràng buộc (constraint-based clustering).

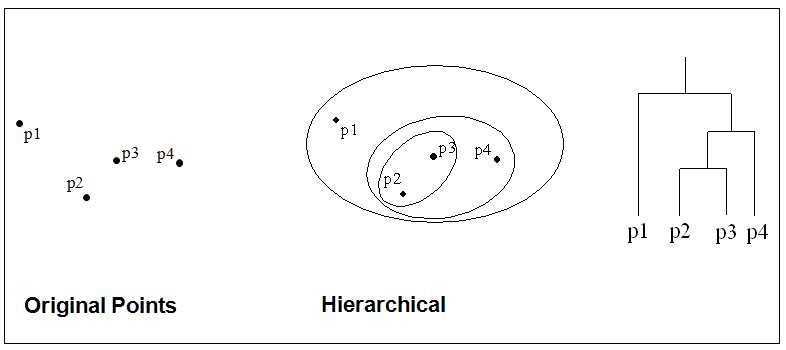
- Khả diễn và khả dụng (interpretability and usability).

### **1.2.2 Các Phương pháp phân cụm dữ liệu tiêu biểu.**

- Phân hoạch (partitioning): các phân hoạch được tạo ra và đánh giá theo một tiêu chí nào đó.

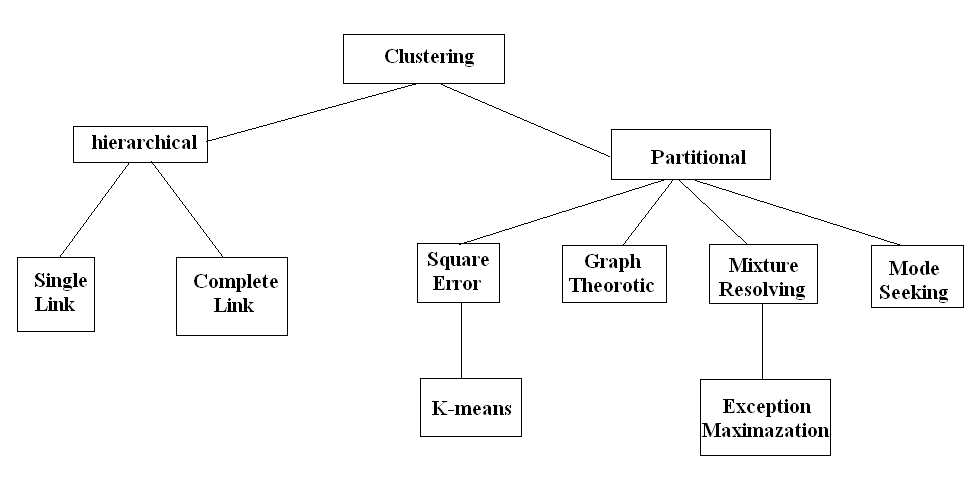


- Phân cấp (hierarchical): phân rã tập dữ liệu/đối tượng có thứ tự phân cấp theo một tiêu chí.

- Dựa trên mật độ (density-based): dựa trên connectivity and density functions.

- Dựa trên lưới (grid-based): dựa trên a multiple-level granularity structure.

- Dựa trên mô hình (model-based): một mô hình giả thuyết được đưa ra cho mỗi cụm; sau đó hiệu chỉnh các thông số để mô hình phù hợp với cụm dữ liệu/đối tượng.



# **PHẦN II: THUẬT TOÁN K-MEANS**

## **2.1 Giới thiệu thuật toán K-means.**

Đây là thuật toán nổi tiếng và được sử dụng nhiều nhất trong hướng tiếp cận phân nhóm phân hoạch. Thuật toán này có nhiều biến thể khác nhau nhưng được đưa ra đầu tiên bởi J.B MacQueen vào năm 1967. Đầu vào của thuật toán này là một tập gồm n mẫu và một số nguyên K. Cần phân n đối tượng này thành K cụm sao cho sự giống nhau giữa các mẫu trong cùng một cụm là cao hơn là giữa các đối tượng khác trong cụm khác.

Tư tưởng của thuật toán này như sau: Đầu tiên chọn ngẫu nhiên K mẫu, mỗi mẫu này coi như biểu diễn 1 cụm, như vậy lúc này trong mỗi cụm thì đối mẫu đó cũng là tâm của cụm (hay còn gọi là nhân). Các mẫu còn lại được gán vào một nhóm nào đó trong K nhóm đã có sao cho tổng khoảng cách từ nhóm mẫu đó đến tâm của nhóm là nhỏ nhất. Sau đó tính lại tâm cho các nhóm và lặp lại quá trình đó cho đến khi hàm tiêu chuẩn hội tụ. Hàm tiêu chuẩn hay được dùng nhất là hàm tiêu chuẩn sai-số vuông. Thuật toán này có thể áp dụng được đối với CSDL đa chiều, nhưng để dễ minh họa chúng tôi mô tả thuật toán trên dữ liệu hai chiều.

## **2.2 Thuật toán K-means.**

Thuật toán k-means được mô tả cụ thể như sau:

Input: K, và dữ liệu về n mẫu của 1 CSDL.

Output: Một tập gồm K cluster sao cho cực tiểu về tổng sai-số vuông.

Thuật toán:

Bước 1: Chọn ngẫu nhiên K mẫu vào K cluster. Coi tâm của cluster chính là mẫu có trong cluster.

Bước 2: Tìm tâm mới của cluster.

Bước 3: Gán (gán lại) các mẫu vào từng cluster sao cho khoảng cách từ mẫu đó đến tâm của cluster đó là nhỏ nhất.

Bước 4: Nếu các cluster không có sự thay đổi nào sau khi thực hiện bước 3 thì chuyển sang bước 5, ngược lại sang bước 2.

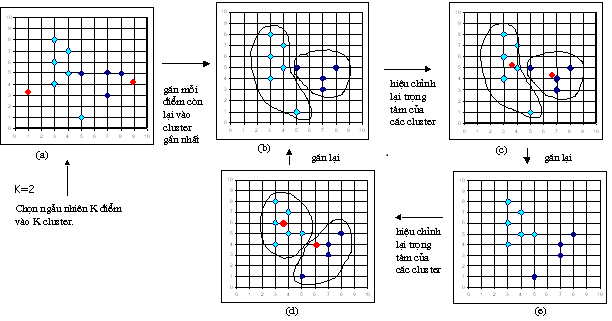
Bước 5: Dừng thuật toán.

### **2.2.1 Ví dụ 1.**

Giả sử trong không gian hai chiều, cho 12 điểm (n = 12) cần phân 12 điểm này thành hai cluster (k=2).

Đầu tiên chọn hai điểm ngẫu nhiên vào hai cluster, giả sử chọn điểm (1,3) và điểm (9,4) (điểm có màu đỏ trên hình 9.a).

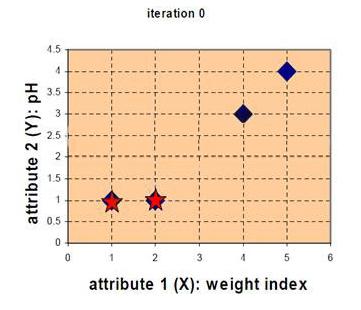
Coi điểm (1,3) là tâm của cluster 1 và điểm (9,4) là tâm của cluster hai. Tính toán khoảng cách từ các điểm khác đến hai điểm này và ta gán được các điểm còn lại này vào một trong hai cluster, những điểm có màu xanh lơ vào cluster 1, những điểm có màu xanh đậm vào cluster 2 (hình 9.b). Hiệu chỉnh lại tâm của hai cluster, điểm màu đỏ trên hình 9.c là tâm mới của hai cluster. Tính lại các khoảng cách các điểm đến tâm mới và gán lại các điểm này, hình 9d. Tiếp tục hiệu chỉnh lại tâm của hai cluster. Cứ như thế lặp lại cho đến khi không còn sự thay đổi nữa thì dừng. Khi đó ta thu được output của bài toán.



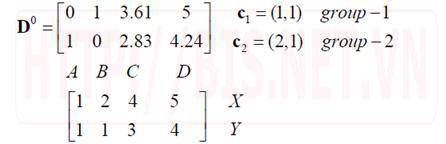
### **2.2.2 Ví dụ 2.**

Giả sử ta có 4 loại thuốc A, B, C, D, mỗi loại thuộc được biểu diễn bởi 2 đặc trưng X và Y như sau. Mục đích của ta là nhóm các thuốc đã cho vào 2 nhóm (K=2) dựa vào các đặc trưng của chúng.

|  |  |  |
| --- | --- | --- |
| object | Feature 1 (X): weight index | Feature 2 (X): pH |
| Medicine A | 1 | 1 |
| Medicine B | 2 | 1 |
| Medicine C | 4 | 3 |
| Medicine D | 5 | 4 |

Bước 1. Khởi tạo tâm (centroid) cho 2 nhóm. Giả sử ta chọn A là tâm của nhóm thứ nhất (tọa độ tâm nhóm thứ nhất c1(1,1)) và B là tâm của nhóm thứ 2 (tạo độ tâm nhóm thứ hai c2 (2,1)).

Bước 2. Tính khoảng cách từ các đối tượng đến tâm của các nhóm (Khoảng cách Euclidean).



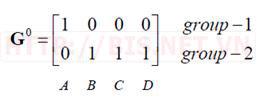
Mỗi cột trong ma trận khoảng cách (D) là một đối tượng (cột thứ nhất tương ứng với đối tượng A, cột thứ 2 tương ứng với đối tượng B,…). Hàng thứ nhất trong ma trận khoảng cách biểu diễn khoảng cách giữa các đối tượng đến tâm của nhóm thứ nhất (c1) và hàng thứ 2 trong ma trận khoảng cách biểu diễn khoảng cách của các đối tượng đến tâm của nhóm thứ 2 (c2).

Ví dụ, khoảng cách từ loại thuốc C=(4,3) đến tâm c1(1,1) là 3.61 và đến tâm c2(2,1) là 2.83 được tính như sau:

http://bis.net.vn/photos/storage/ClusteringTechnique006.jpg

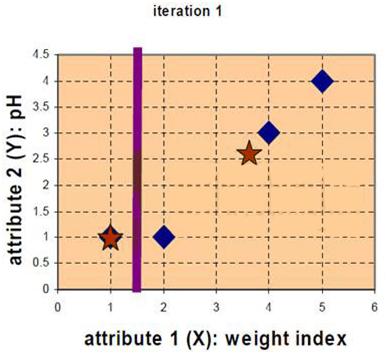
http://bis.net.vn/photos/storage/ClusteringTechnique007.jpg

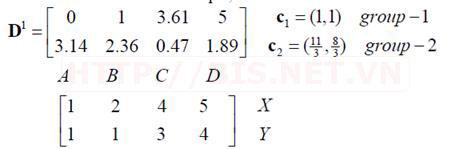
Bước 3. Nhóm các đối tượng vào nhóm gần nhất.



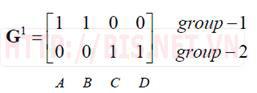
http://bis.net.vn/photos/storage/ClusteringTechnique009.jpgTa thấy rằng nhóm 1 sau vòng lặp thứ nhất gồm có 1 đối tượng A và nhóm 2 gồm các đối tượng còn lại B, C, D.

Bước 4. Tính lại tọa độ các tâm cho các nhóm mới dựa vào tọa độ của các đối tượng trong nhóm. Nhóm 1 chỉ có 1 đối tượng A nên tâm nhóm 1 vẫn không đổi, c1(1,1). Tâm nhóm 2 được tính như sau:



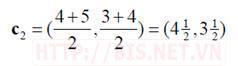
Bước 5. Tính lại khoảng cách từ các đối tượng đến tâm mới.

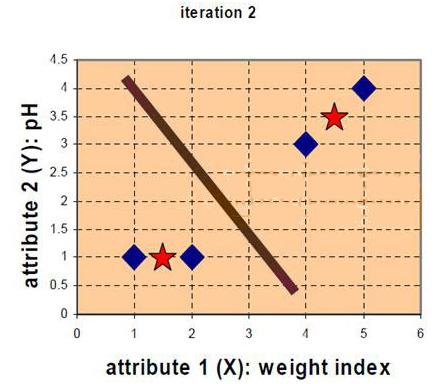
Bước 6. Nhóm các đối tượng vào nhóm.



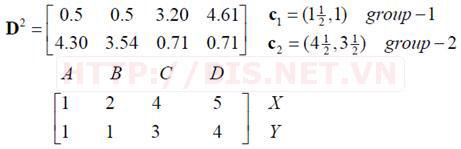
**Bước 7**. Tính lại tâm cho nhóm mới.

http://bis.net.vn/photos/storage/ClusteringTechnique013.jpg

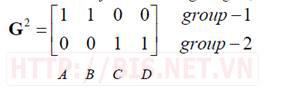




**Bước 9.** Tính lại khoảng cách từ các đối tượng đến tâm mới.



**Bước 10.** Nhóm các đối tượng vào nhóm.



Ta thấy G2 = G1 (Không có sự thay đổi nhóm nào của các đối tượng) nên thuật toán dừng và kết quả phân nhóm như sau:

|  |  |  |  |
| --- | --- | --- | --- |
| Object | Feature 1 (X): weight index | Future 2 (X); pH | Group (result) |
| Medicine A | 1 | 1 | 1 |
| Medicine B | 2 | 1 | 1 |
| Medicine C | 4 | 3 | 2 |
| Medicine D | 5 | 4 | 2 |

Ưu điểm:

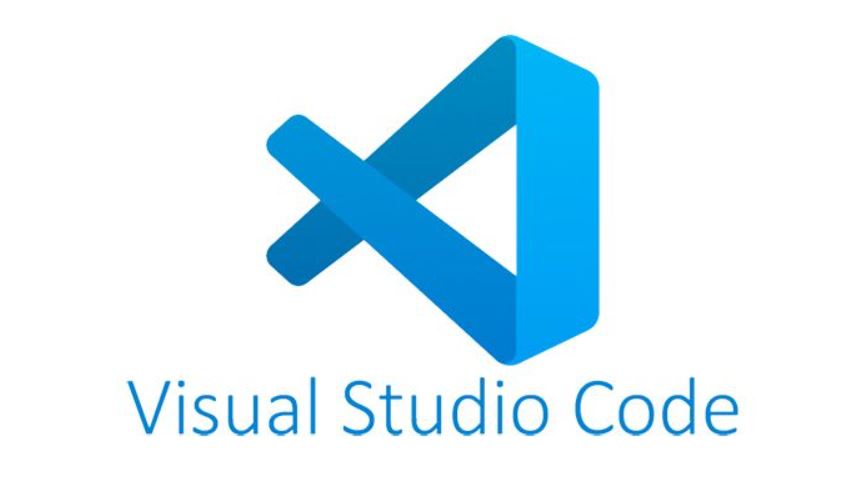
- Dễ hiểu và cài đặt.

Hạn chế:

- Phụ thuộc vào số nhóm K chọn ban đầu.

- Chi phí cho thực hiện vòng lặp tính toán khoảng cách lớn khi số cụm K và dữ liệu phân cụm lớn.

## **2.3 Xây dựng và demo thuật toán K-Means.**

Chương trình được xây dựng bằng ngôn ngữ Python 3.9.9 và sử dụng trình biên tập VSCode.

Vẽ giao diện:

**Ảnh có chứa văn bản

Mô tả được tạo tự động**

**Ảnh có chứa văn bản

Mô tả được tạo tự động**Thực hiện gán các điểm gần với cụm gần nó nhất và cập nhật lại tâm của các cụm:

Giao diện sau khi hoàn thành:

**Ảnh có chứa quảng trường

Mô tả được tạo tự động**

## **2.4 Ứng dụng của thuật toán K-Means**

Trong ví dụ này, tôi sẽ giới thiệu cách xây dựng một KnowledgeFlow để triển khai kỹ thuật phân cụm dựa trên thuật toán K-Means trên Data Mining Software WeKa.

Link download: <http://www.cs.waikato.ac.nz/ml/weka/downloading.html>

Dữ liệu dùng để phân cụm trong ví dụ này là dữ liệu dùng để phân loại khách hàng của ngân hàng gồm có 11 thuộc tính và 600 khách hàng (file bank.arff).Nhiệm vụ của chúng ta là dùng thuật toán K-Means để phân nhóm các khách hàng vào K nhóm (trong ví dụ này K=5) dựa vào sự tương tự (similar) trên 11 thuộc tính của họ.

Ảnh có chứa văn bản, ảnh chụp màn hình, máy tính, trong nhà

Mô tả được tạo tự động

Ta xây dựng một KnowledgeFlow trong WeKa như sau:

Ảnh có chứa văn bản, ảnh chụp màn hình, trong nhà

Mô tả được tạo tự động

Thiết lập các tham số cho thuật toán K-Means như số cụm (trong ví dụ này K=5), Cách tính khoảng cách (trong ví dụ này dùng khoảng cách Euclidean),…

Ảnh có chứa văn bản, trong nhà, ảnh chụp màn hình

Mô tả được tạo tự động

Kết quả phân cụm chi tiết như sau:

Ảnh có chứa văn bản, ảnh chụp màn hình, trong nhà

Mô tả được tạo tự động

Ảnh có chứa văn bản, ảnh chụp màn hình, máy tính, máy tính xách tay

Mô tả được tạo tự động

# **TÀI LIỆU THAM KHẢO**

http://bis.net.vn/forums/t/374.aspx?PageIndex=1

http://en.wikipedia.org/wiki/K-means\_clustering