ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC BÁCH KHOA
KHOA KHOA HỌC - KỸ THUẬT MÁY TÍNH

KHAI PHÁ DỮ LIỆU

Thực hành phân cụm dữ liệu

Áp dụng K-Means và công cụ Weka

GVHD:	Lê Hồng Trang	
Lớp:	L01	
Nhóm:	Undefine	
Thành viên:	Bùi Quốc Khải	1711726
	Nguyễn Văn Hoàng	1711400
	Huỳnh Thi Trường Duy	1710779
	Võ Quý Giang	1711130

TP. HỒ CHÍ MINH, THÁNG 05/2021

Mục lục

1 Cở sở kỹ thuật

1.1 Phân cụm dữ liệu

1.1.1 Phân cụm

Phân cụm là kỹ thuật rất quan trọng trong khai phá dữ liệu, nó thuộc lớp các phương pháp Unsupervised Learning trong Machine Learning. Có rất nhiều định nghĩa khác nhau về kỹ thuật này, nhưng về bản chất ta có thể hiểu phân cụm là các qui trình tìm cách nhóm các đối tượng đã cho vào các cụm (clusters), sao cho các đối tượng trong cùng 1 cụm tương tự (similar) nhau và các đối tượng khác cụm thì không tương tự (Dissimilar) nhau.

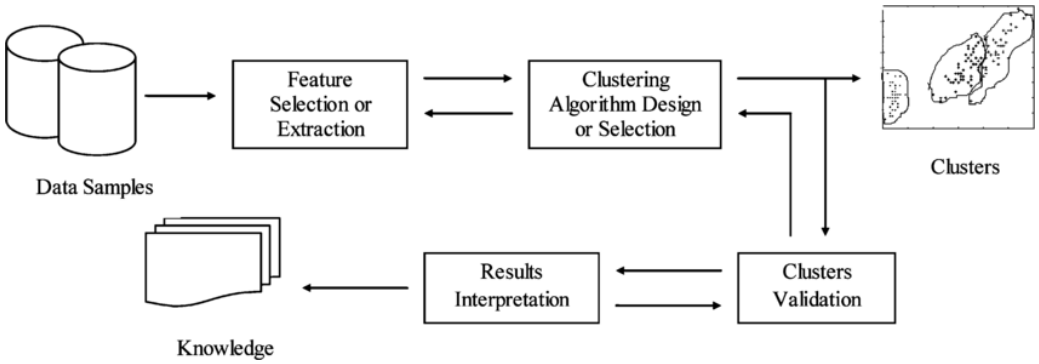
Mục đích của phân cụm là tìm ra bản chất bên trong các nhóm của dữ liệu. Các thuật toán phân cụm (Clustering Algorithms) đều sinh ra các cụm (clusters). Tuy nhiên, không có tiêu chí nào là được xem là tốt nhất để đánh hiệu của của phân tích phân cụm, điều này phụ thuộc vào mục đích của phân cụm như: data reduction, "natural clusters", "useful" clusters, outlier detection.

1.1.2 Ứng dụng thực tế

Kỹ thuật phân cụm có thể áp dụng trong rất nhiều lĩnh vực như:

1. Marketing: Xác định các nhóm khách hàng (khách hàng tiềm năng, khách hàng giá trị, phân loại và dự đoán hành vi khách hàng,...) sử dụng sản phẩm hay dịch vụ của công ty để giúp công ty có chiến lược kinh doanh hiệu quả hơn.
2. Biology: Phận nhóm động vật và thực vật dựa vào các thuộc tính của chúng.
3. Libraries: Theo dõi độc giả, sách, dự đoán nhu cầu của độc giả...
4. Insurance, Finance: Phân nhóm các đối tượng sử dụng bảo hiểm và các dịch vụ tài chính, dự đoán xu hướng (trend) của khách hàng, phát hiện gian lận tài chính (identifying frauds).
5. WWW: Phân loại tài liệu (document classification); phân loại người dùng web (clustering weblog).

1.1.3 Quá trình phân cụm

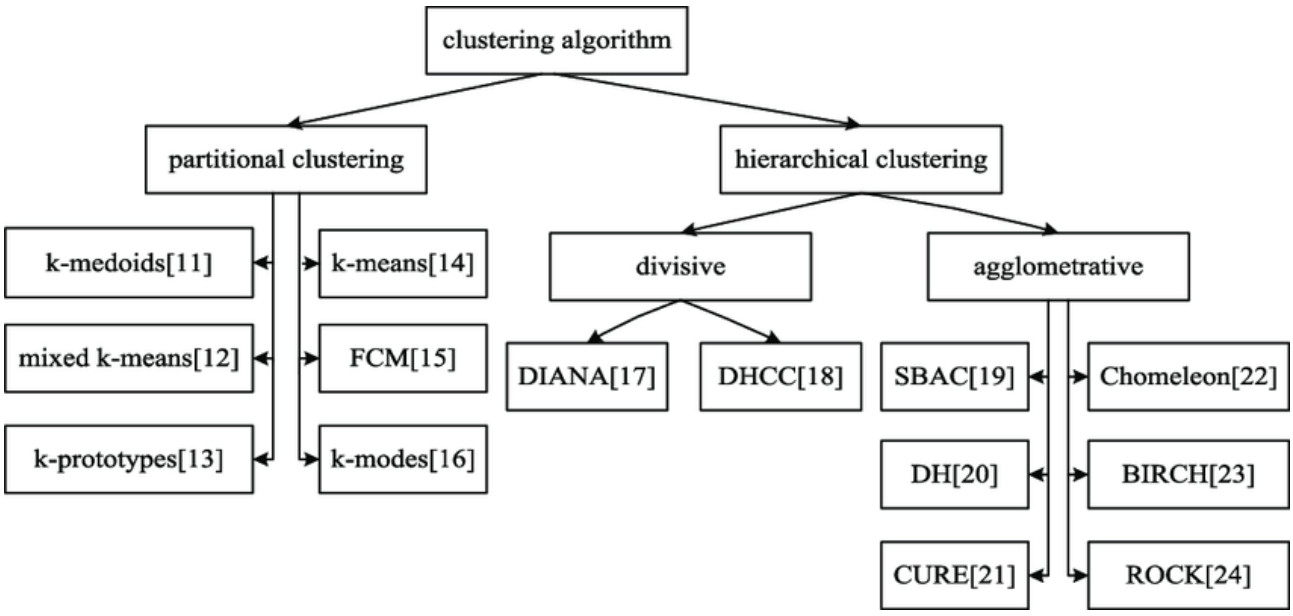


Hình 1: Quá trình phân cụm dữ liệu (www.researchgate.net)

1.1.4 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 gom 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 gom cụm dựa trên ràng buộc (constraint-based clustering).
- Khả diễn và khả dụng (interpretability and usability).

1.1.5 Phân loại các kỹ thuật phân cụm



Hình 2: Các loại phân cụm dữ liệu (www.researchgate.net)

1.2 Kỹ thuật K-Means

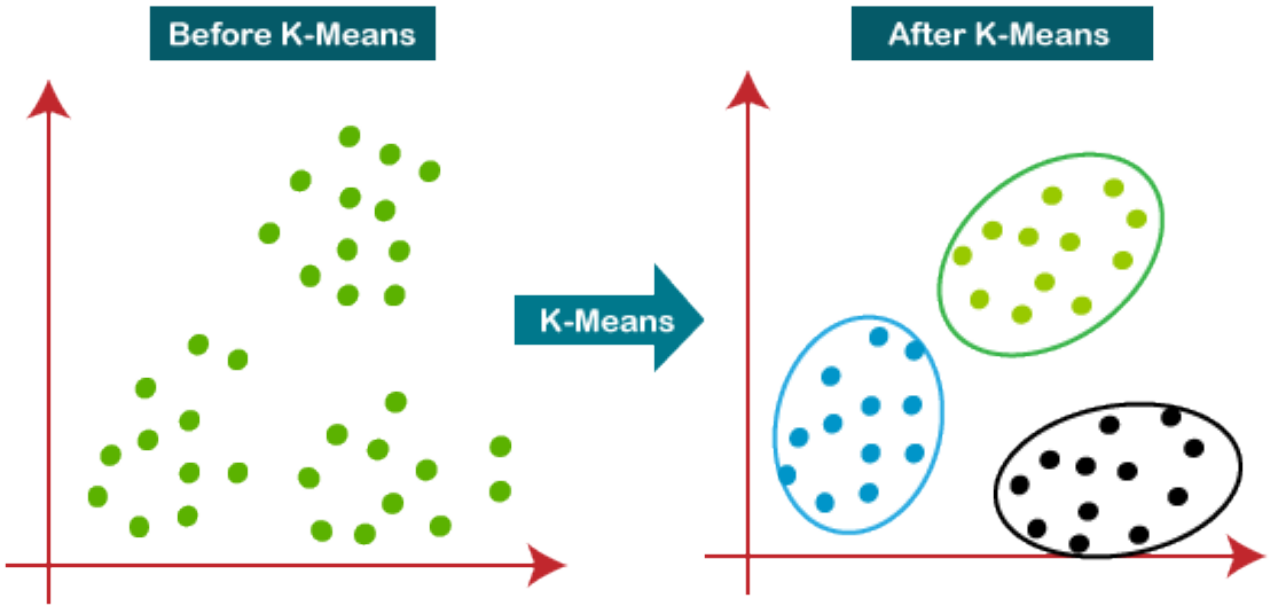
1.2.1 Thuật toán K-Means

K-Means Clustering là một thuật toán rất quan trọng và được sử dụng phổ biến trong kỹ thuật phân cụm, thuộc nhóm thuật toán học tập không giám sát (Unsupervised Learning algorithm). Tư tưởng chính của thuật toán K-Means là tìm cách phân nhóm các đối tượng (objects) đã cho vào K cụm (K là số các cụm được xác đinh trước, K nguyên dương). Dựa trên centroid, trong đó mỗi cụm được liên kết với một centroid. Mục đích chính của thuật toán này là giảm thiểu tổng khoảng cách giữa điểm dữ liệu và các cụm tương ứng.

Thuật toán lấy tập dữ liệu không được gắn nhãn làm đầu vào, chia tập dữ liệu thành K-số cụm và lặp lại quá trình cho đến khi không tìm thấy cụm tốt nhất. Giá trị của K là giá trị được xác định trước.

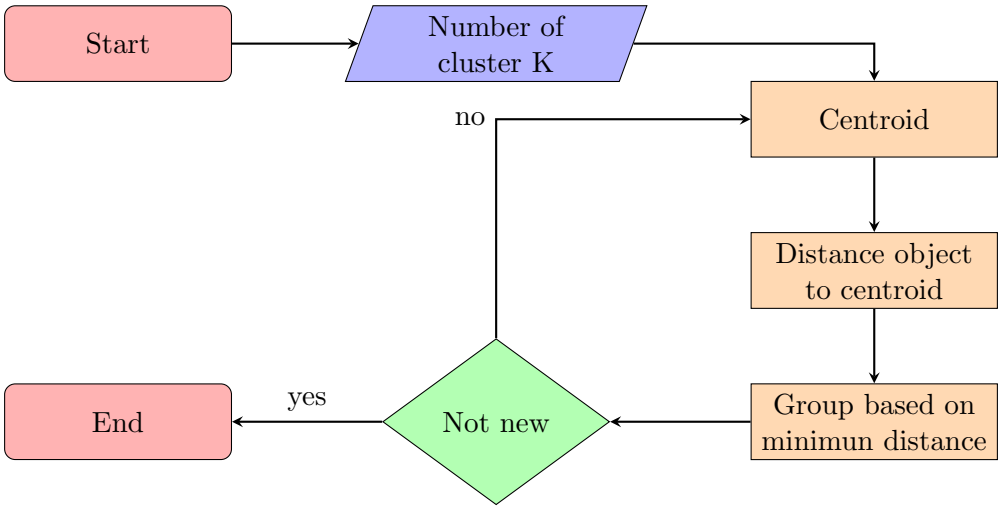
Thuật toán phân cụm k- mean chủ yếu thực hiện hai tác vụ:

- Xác định giá trị tốt nhất cho K điểm trung tâm bằng một quy trình lặp lại.
- Gán mỗi điểm dữ liệu cho trung tâm K gần nhất. Những điểm dữ liệu gần trung tâm K cụ thể sẽ tạo ra một cụm.



Hình 3: Kết quả phân cụm với K-Means

1.2.2 Sơ đồ hoạt động



1.2.3 Mô tả hoạt động

Hoạt động của thuật toán K-Means tuân theo các bước:

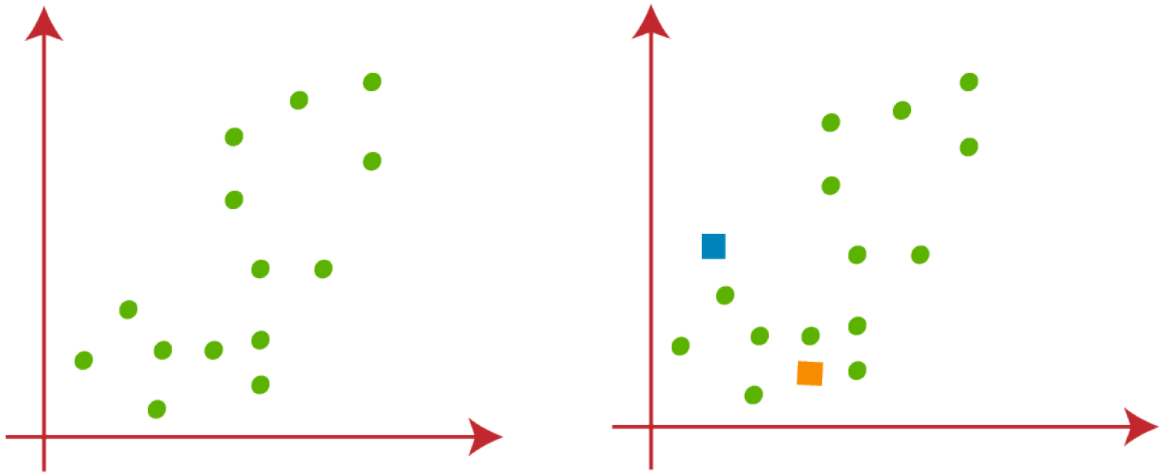
1. Chọn số K để quyết định số lượng cụm.
2. Chọn K điểm hoặc trọng tâm ngẫu nhiên. (Có thể khác với tập dữ liệu đầu vào).
3. Gán mỗi điểm dữ liệu cho trung tâm gần nhất của chúng, sẽ tạo thành các cụm K được xác định trước.
4. Tính toán phương sai và đặt một trung tâm mới của mỗi cụm.
5. Lặp lại các bước thứ ba, có nghĩa là chỉ định lại mỗi điểm dữ liệu cho trung tâm gần nhất mới của mỗi cụm.
6. Nếu có bất kỳ sự phân công lại nào xảy ra, hãy chuyển sang bước-4, sau đó hoàn tất quá trình.

1.2.4 Minh họa trực quan

Giả sử chúng ta có hai biến M1 và M2. Biểu đồ phân tán trục xy của hai biến này được biểu diễn:

Hãy lấy số K của các cụm, ta chọn, K = 2, để xác định tập dữ liệu và đặt chúng vào các cụm khác nhau. Nghĩa là sẽ cố gắng nhóm các tập dữ liệu này thành hai cụm khác nhau.

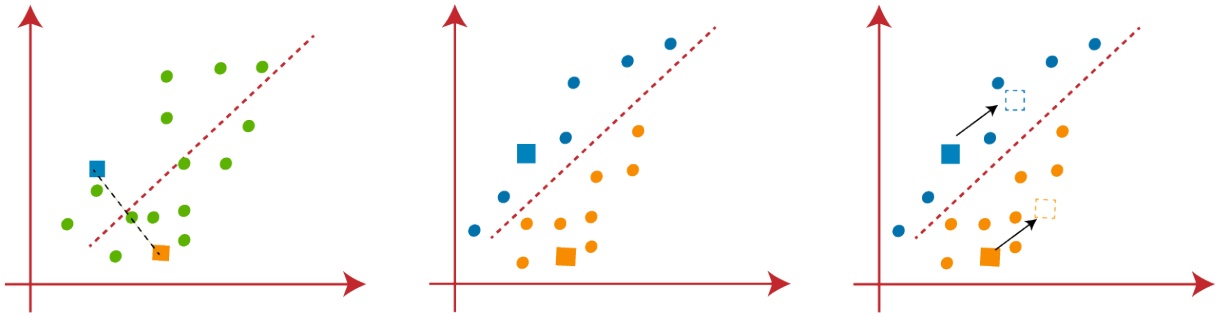
Chọn một số điểm K ngẫu nhiên hoặc centroid để tạo thành cụm. Những điểm này có thể là điểm từ tập dữ liệu hoặc bất kỳ điểm nào khác. Ở đây chúng ta đang chọn hai điểm dưới, không phải là một phần của tập dữ liệu.



Hình 4: Chọn 2 centroid khởi đầu

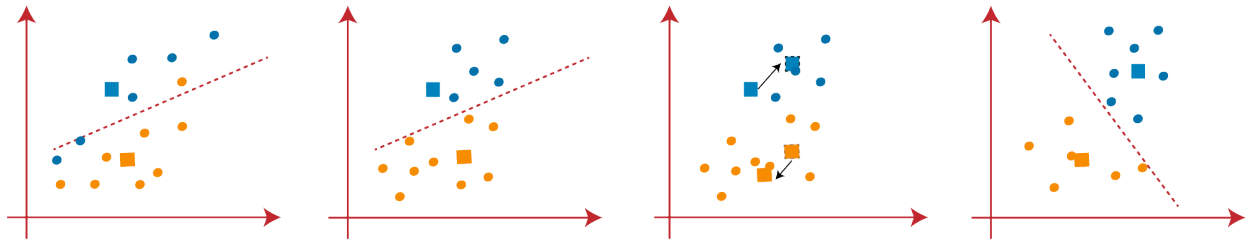
Gán mỗi điểm dữ liệu của biểu đồ phân tán cho điểm K hoặc tâm gần nhất của nó. Chúng ta áp dụng một số toán học để tính khoảng cách giữa hai điểm. Vẽ đường trung trực của đường nối giữa 2 centroid.

Chúng ta nhận thấy đường gạch đỏ vừa tạo chia cách 2 tập điểm về 2 phía khác nhau ứng với 2 centroid hiện tại. Tô lại màu cho 2 cụm vừa được gom.



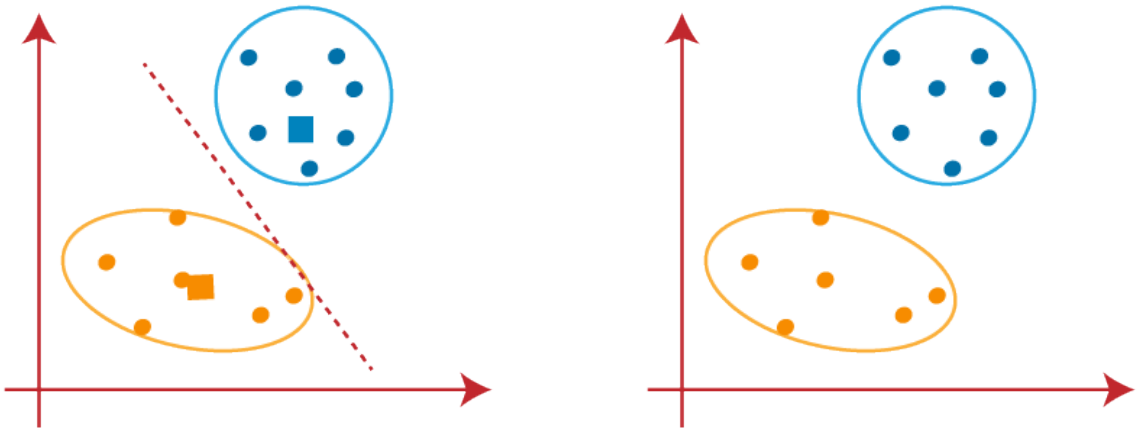
Hình 5: Vẽ trung trực phân cụm dữ liệu xác định centroid mới

Quá trình được lặp lại bằng cách tìm một trung tâm mới cho 2 cụm vừa được gom bằng cách tính toán trọng tâm các điểm của từng cụm.



Hình 6: Lặp lại quá trình phân cụm

Quá trình lặp lại đến khi không còn xuất hiện điểm dữ liệu nào di chuyển giữa các nhóm. Mô hình được hoàn chỉnh và kết thúc.



Hình 7: Cụm cố định và kết thúc thuật toán

1.2.5 Xác định K trong thuật toán K-Means

Hiệu suất của thuật toán phân cụm K-mean phụ thuộc vào số cụm được phân. Vậy vậy xác định số lượng cụm tối ưu là một nhiệm vụ quan trọng. Có nhiều cách khác nhau để tìm số lượng cụm tối ưu, trong đó phương pháp Khuỷu tay (Elbow Method) được xem là hiệu quả và phổ biến nhất trong đánh giá lựa chọn số lượng K cho giải thuật K-Means.

WCSS là viết tắt của Within Cluster Sum of Squares , xác định tổng số các biến thể trong một cụm. Công thức tính giá trị của WCSS được khái quát:

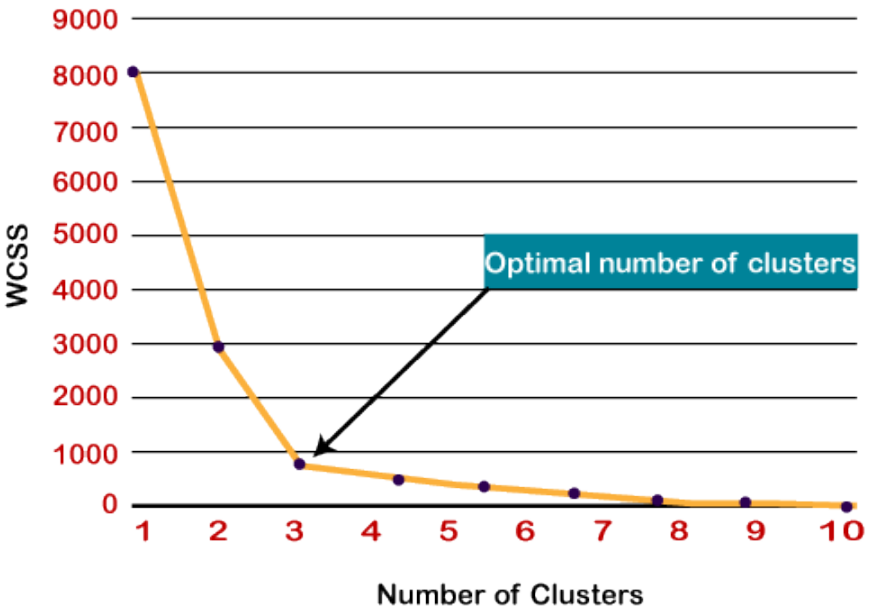
$$WWCS(k) = \sum_{j=1}^{k} \sum_{i \in clusterj} \|x_i - \overline{x_j}\|^2$$

Trong đó $\overline{x_j}$ là trung bình mãu trong cọm j.

Để tìm giá trị tối ưu của các cụm, phương pháp Elbow thực hiện theo các bước:

1. Thực hiện phân cụm K-Means trên một tập dữ liệu nhất định cho các giá trị K khác nhau (phạm vi từ 1-10).
2. Đối với mỗi giá trị của K tính giá trị WCSS.

3. Vẽ đồ thị một đường cong giữa các giá trị WCSS được tính toán và số lượng các cụm K.
4. Điểm uốn cong hoặc một điểm của đồ thị có dạng như một cánh tay thì điểm đó được coi là giá trị tốt nhất của K.



Hình 8: Đồ thị xác định K trong K-Means

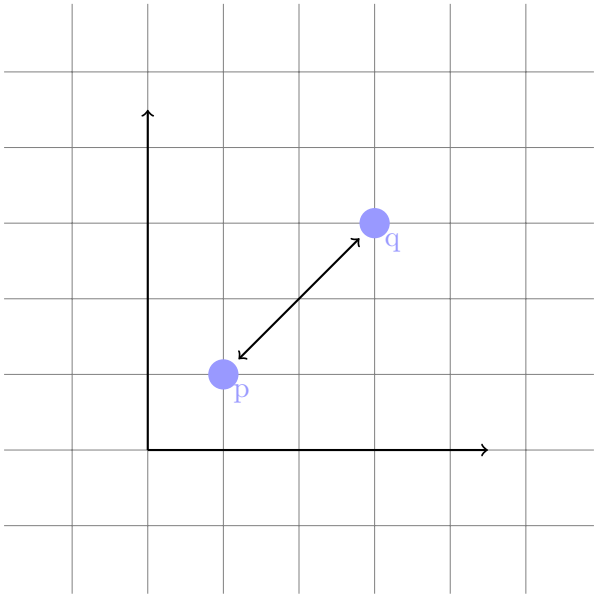
1.2.6 Các phương pháp tính khoảng cách trong cụm

Có nhiều phương pháp được áp dụng để tính toán khoảng cách giữa cách điểm trong tập phân cụm:

- Phương pháp thước đo Euclidean.
- Phương pháp thước đo Manhattan.
- Phương pháp thước đo Euclidean về bình phương khoảng cách.
- Phương pháp thước đo Cosine.

1.2.6.a Euclidean Distance Measure

Trường hợp phổ biến nhất là xác định khoảng cách giữa hai điểm. Nếu ta có điểm P và điểm Q thì khoảng cách euclidean là một đường thẳng thông thường. Đó là khoảng cách giữa hai điểm trong không gian euclidean.



$$d = \sqrt{\sum_{i=1}^{n}(q_j - p_i)^2}$$

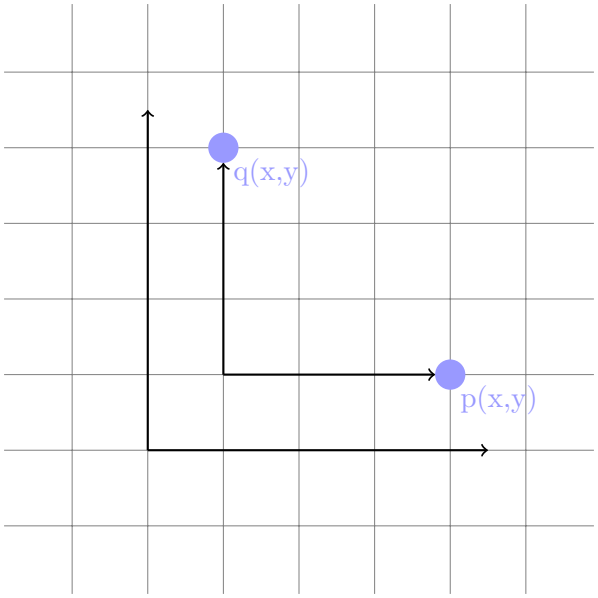
1.2.6.b Euclidean Distance Measure

Nếu không áp dụng căn của tổng ta được phương pháp đo bình phương khoảng cách:

$$d = \sum_{i=1}^{n}(q_j - p_i)^2$$

1.2.6.c Manhattan Distance Measure

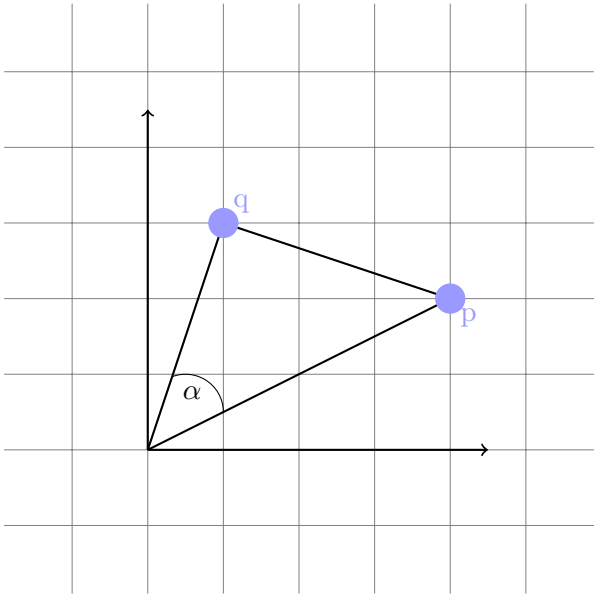
Khoảng cách Manhattan được tính theo tổng khoảng cách giưa hai điểm được đo theo chiều ngang và chiều dọc.



$$d = \sum_{i=1}^{n} |q_x - p_x| + |q_y - p_y|$$

1.2.6.d Cosine Distance Measure

Dựa vào góc giữa hai đoạn thẳng nối điểm với gốc tọa độ.



$$d = \frac{\sum_{i=1}^{n-1} q_i - p_x}{\sum_{i=1}^{n-1}(q_i)^2 \times \sum_{i=1}^{n-1}(p_i)^2}$$

2 Các biến thể cải tiến của K-Means

2.1 K-Means++

Vấn đề đặt ra đối với thuật toán K-Means trong phân cụm dữ liệu là tìm ra trung tâm của cụm (tối thiểu hóa phương sai đến các tập điểm còn lại trong cụm), đó là tổng khoảng cách từ mỗi điểm

trong một cụm đến trung tâm cụm (gần nhất với nó).Mặc dù, việc tìm kiếm lời giải chính xác cho vấn đề K-Means này là một bài toán khó, phương pháp tiêu chuẩn được sử dụng để tìm ra kết quả xấp xỉ (còn được gọi là thuật toán Lloyd) được chấp nhận rộng rãi và thường xuyên.

Tuy nhiên, thuật toán K-Means gặp 2 thiếu sót lớn:

- Thời gian chạy trong trường hợp xấu nhất.
- Việc chọn ngẫu nhiên các điểm k-centroid ban đầu dẫn đến vấn đề về độ nhạy khởi tạo. Có xu hướng ảnh hưởng đến các cụm được hình thành cuối cùng.

Thuật toán K-Means++ giải quyết trở ngại thứ hai bằng cách chỉ định một thủ tục để khởi tạo các k-centroid trước khi tiếp tục với các lần lặp tối ưu hóa K-Means tiêu chuẩn. Với việc khởi tạo K-Mean ++, thuật toán được đảm bảo tìm độ tối ưu O(log k) so với giải pháp K-Means tối ưu.

K-Means++ khởi tạo centroid thông minh và phần còn lại của thuật toán giống như của K-Means. Các bước cần làm để khởi tạo centroid là:

1. Chọn ngẫu nhiên điểm tâm đầu tiên C_i
2. Tính toán khoảng cách của tất cả các điểm trong tập dữ liệu từ trung tâm đã chọn. Khoảng cách của điểm x_i từ tâm xa nhất có thể được tính bằng:
$$d = \max_{j:1 \mapsto m} \|x_i - C_j\|^2$$
3. Lặp lại cho đến khi tìm thấy centroid

2.2 K-Medoids

Ý tưởng của việc phân cụm K-Medoids là tạo ra các trung tâm cuối cùng dưới dạng các điểm dữ liệu thực tế. Kết quả này làm cho các trung tâm có thể giải thích được.

Thuật toán phân cụm K-Medoids được gọi là Phân vùng xung quanh Medoids (PAM) gần giống như thuật toán của Lloyd với một chút thay đổi trong bước cập nhật.

Các bước cần thực hiện đối với thuật toán PAM:

1. Khởi tạo: Giống như của K-Mean ++
2. Nhiệm vụ: Giống như của K-Means
3. Cập nhật trung tâm: Trong trường hợp của K-Means, chúng ta tính toán trung bình của tất cả các điểm có mặt trong cụm. Nhưng đối với việc cập nhật thuật toán PAM của centroid nếu có m-point trong một cụm, hoán đổi centroid trước đó với tất cả (m-1) điểm khác từ cụm và chốt lại điểm đó là centroid mới có mức tổn thất tối thiểu. Tổn thất tối thiểu được tính theo hàm chi phí dưới đây:
$$M_1, M_2, ..., M_k = argmin \sum_{i=1}^{k} \sum_{x \in S_i} \|x - M_i\|^2$$
4. Lặp lại tương tự thuật toán K-Means

3 Một số ứng dụng trong phát hiện bất thường

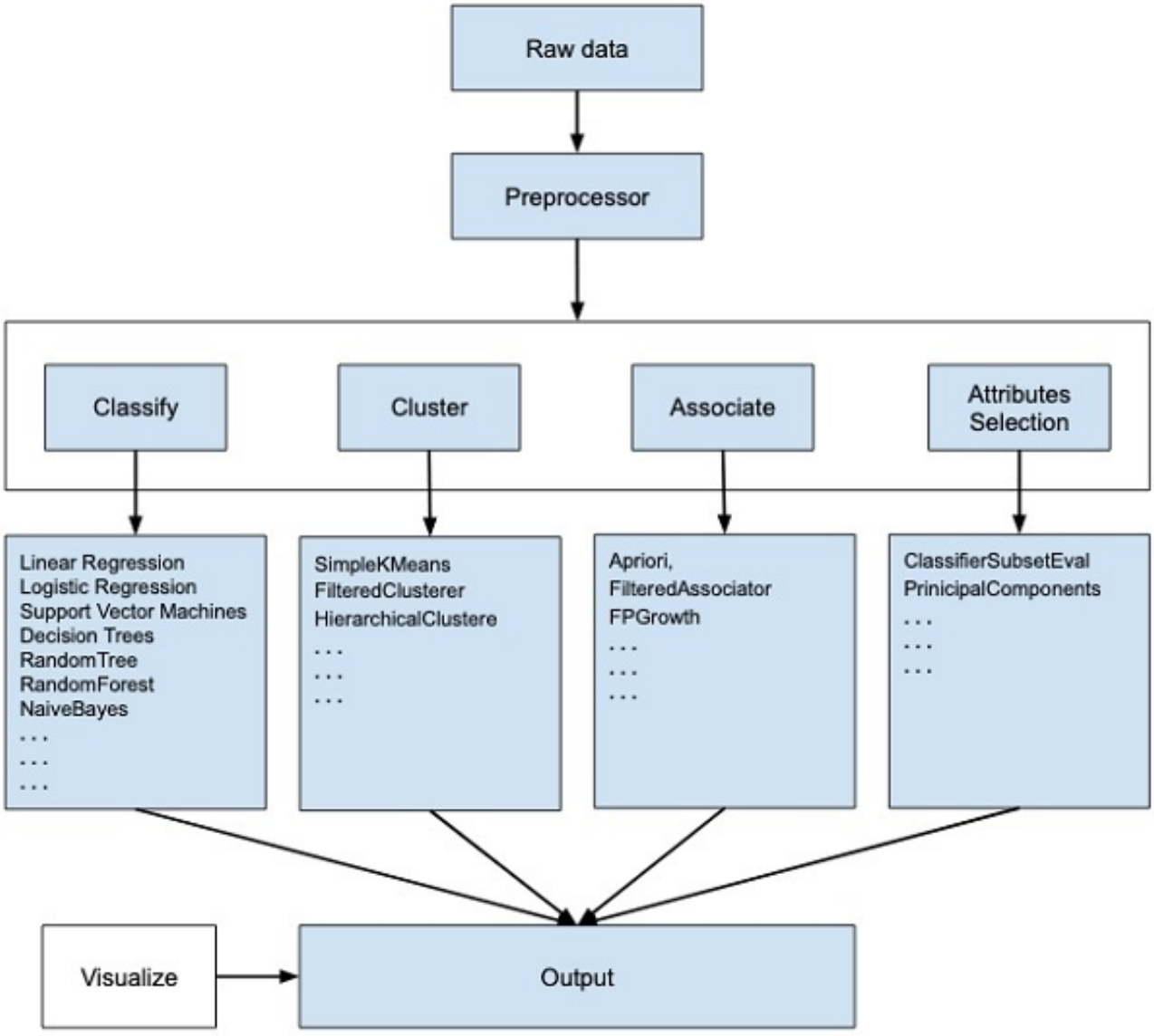
K-Means được áp dụng phổ biến trong các yêu cầu phân cụm, xác định các giá trị bất thường trong tập dữ liệu lịch sử được thống kê. Điều này được ứng dụng hỗ trợ các hoạt động:

- Với dữ liệu liên quan đến tội phạm có sẵn ở các địa phương cụ thể trong thành phố, loại tội phạm, khu vực phạm tội và mối liên hệ giữa hai loại tội phạm có thể cung cấp thông tin chi tiết về các khu vực dễ xảy ra tội phạm trong thành phố hoặc địa phương.
- Tách các vật thể trong ảnh, dựa vào cách giá trị bất thời về màu sắc của vật thể khi được biểu diễn trên hình ảnh, thực hiện gom cụm tách các vùng màu tương tự nhau tạo ra được ảnh mới được tách biệt bởi các gam màu khác nhau.
- Được sử dụng như một phương pháp hỗ trợ phất hiện các bất thường trong hệ thống mạng.

4 Ứng dụng Weka trong gom cụm dữ liệu

4.1 Công cụ Weka

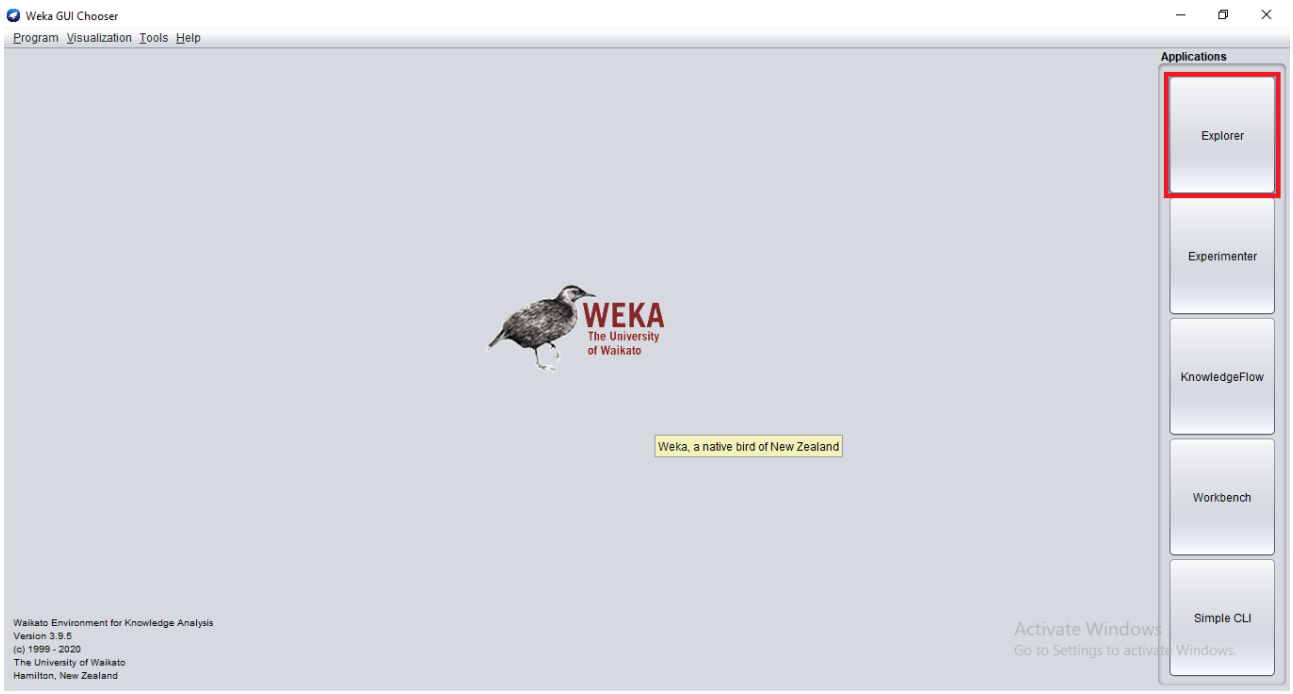
Weka - một phần mềm mã nguồn mở cung cấp các công cụ để xử lý trước dữ liệu, triển khai một số thuật toán Máy học và trực quan hóa giúp phát triển các kỹ thuật máy học và áp dụng chúng vào các vấn đề khai thác dữ liệu trong thế giới thực:



Hình 9: Ứng dụng Weka

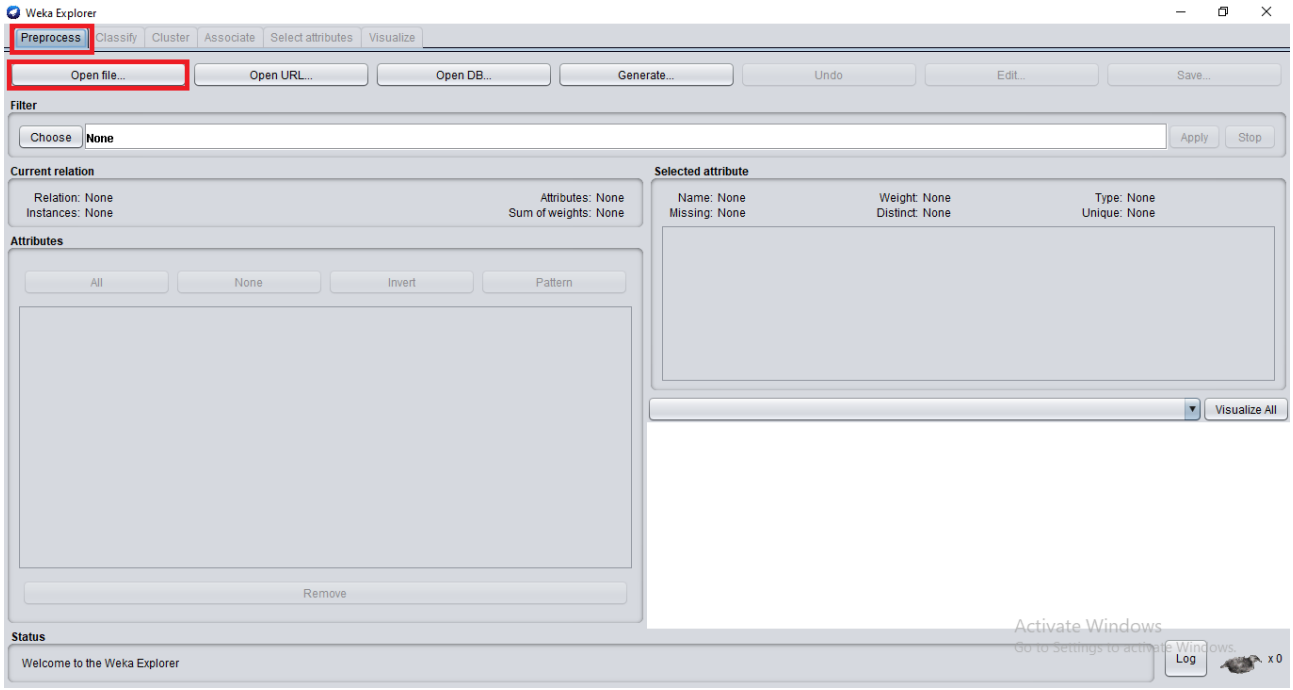
4.2 Sử dụng Weka trong gom cụm dữ liệu

Bước 1: Giao diện chính Weka cung cấp các lựa chọn xử lí dữ liệu và các phương thức học máy khác nhau. Sử dụng Explorer cho yêu cầu.



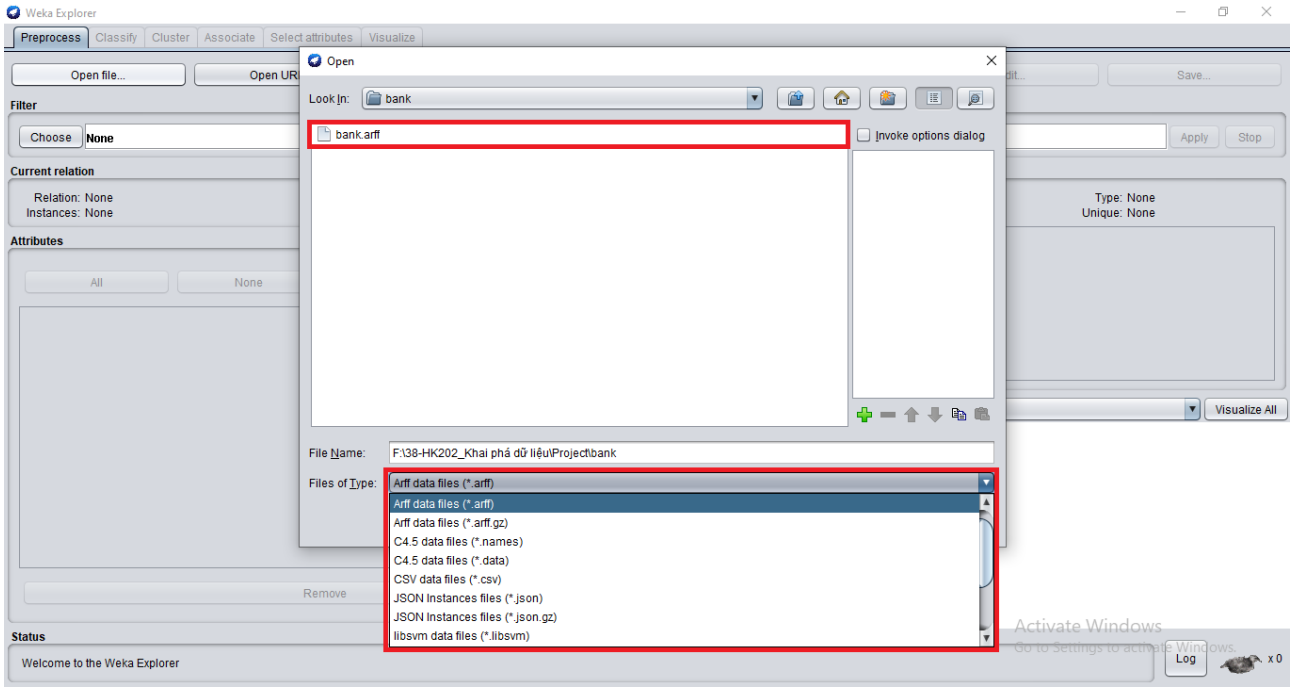
Hình 10: Lựa chọn chế độ Explorer

Bước 2: Ở tab **Preprocess** chọn **Open file**



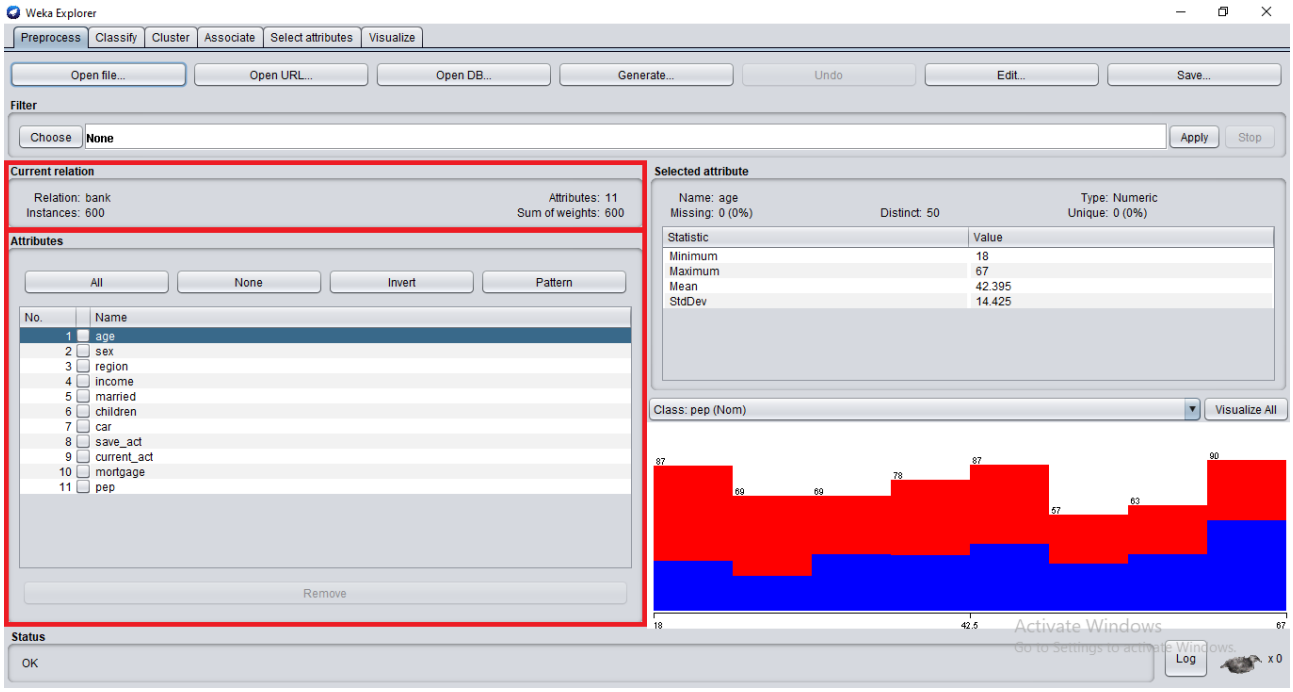
Hình 11: Trong tab preprocess chọn mở file để import file dataset vào Weka

Bước 3: Từ thư mục của máy tính lựa chọn file dataset muốn thực thi. Weka hỗ trợ nhiều loại định dạng tệp khác nhau bao gồm (***.arff, *.arff.gz, *.names, *.data, *.csv, *.json, *.json.gz, *.libsvm, *.m, *.dat, *.bsi, *.xrff, *.xrff.gz**).



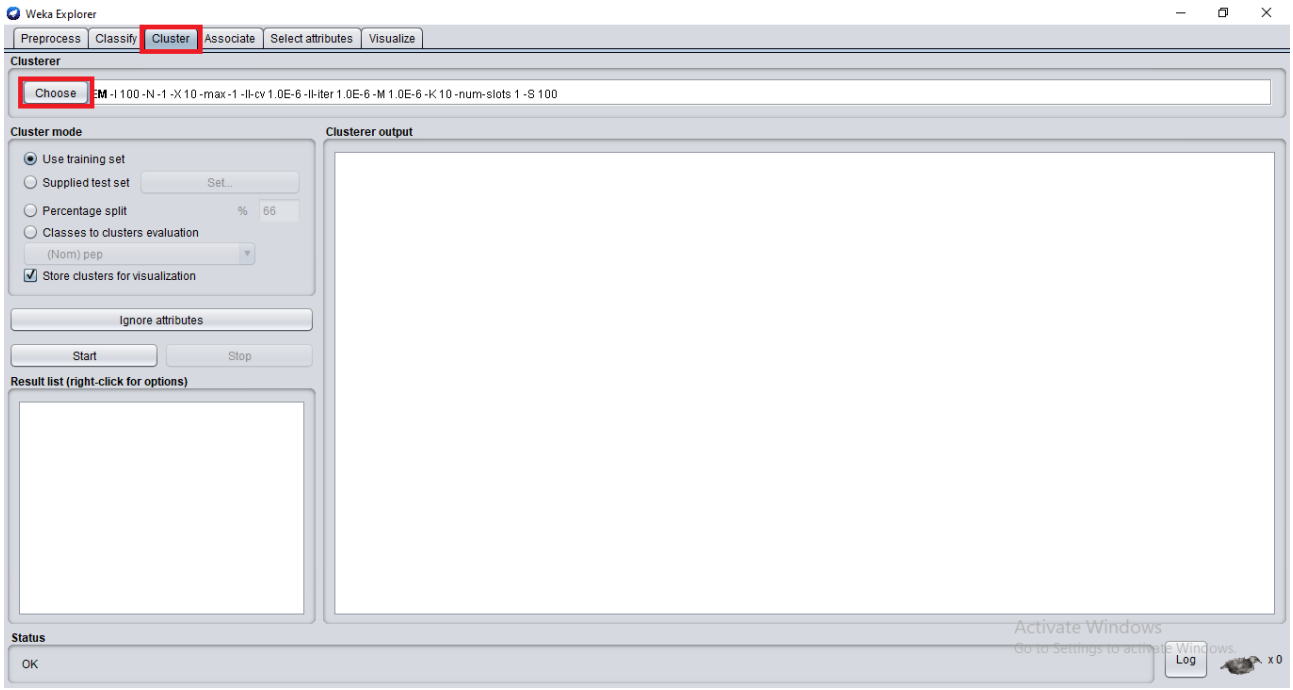
Hình 12: Dẫn file từ thư mục với các định dạng được Weka hỗ trợ sẵng

Bước 4: Sau khi dữ liệu được nhập vào Weka, trình hiển thị các thông số dữ liệu của file dataset. Bao gồm **instance** (số lượng dòng dữ liệu trong tập dataset), **attributes** (các thuộc tính - cột trong tập dataset).



Hình 13: Thông tin về dataset

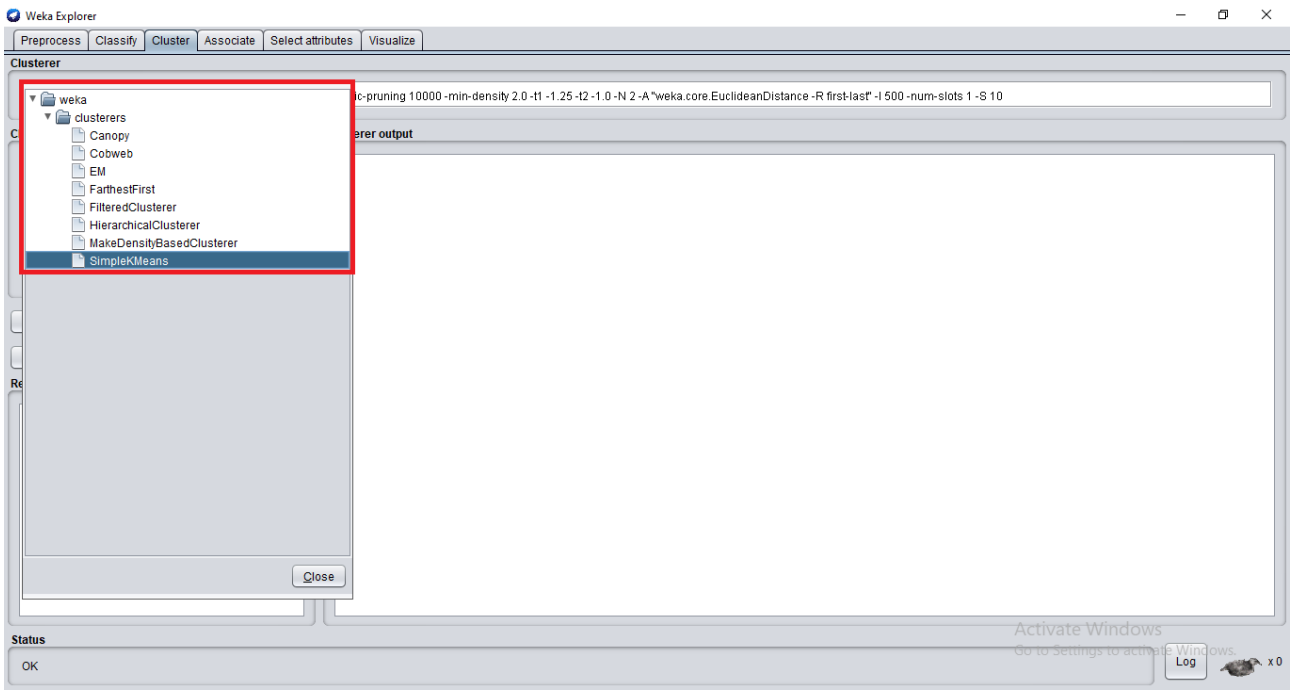
Bước 5: Chọn tab **Cluster** thực thi tác vụ gom nhóm dữ liệu. Chọn **Choose** để chỉ định giải thuật áp dụng.



Hình 14: Giao diện cho tác vụ cluster

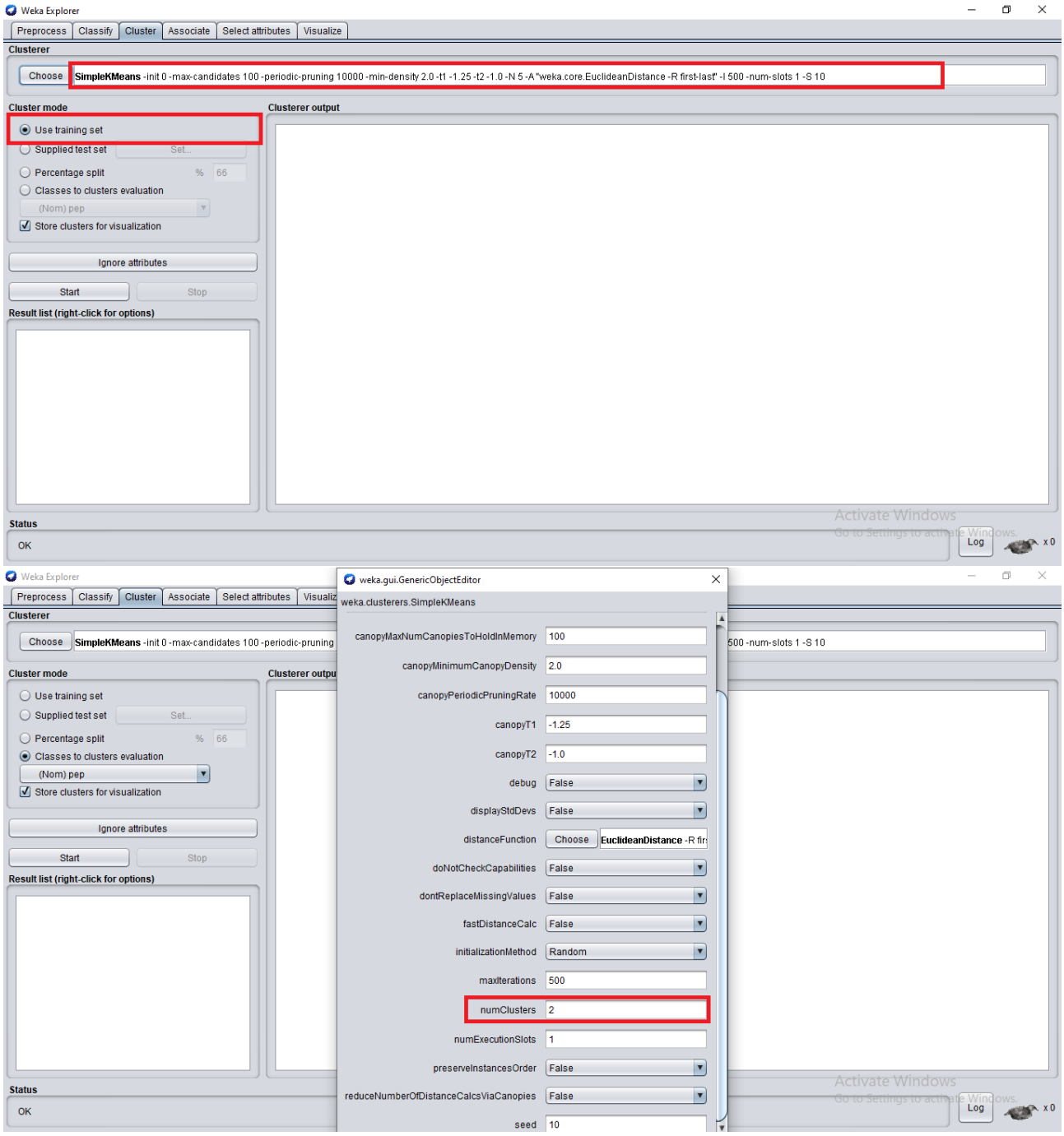
Bước 6: Weka cung cấp nhiều giải thuật gom cụm khác nhau:

- Canopy
- Cobweb
- EM
- FarthesFirst
- FilteredClusterer
- HireachicalClusterer
- MakeDensityBasedClusterer
- **SimepleKMeans**



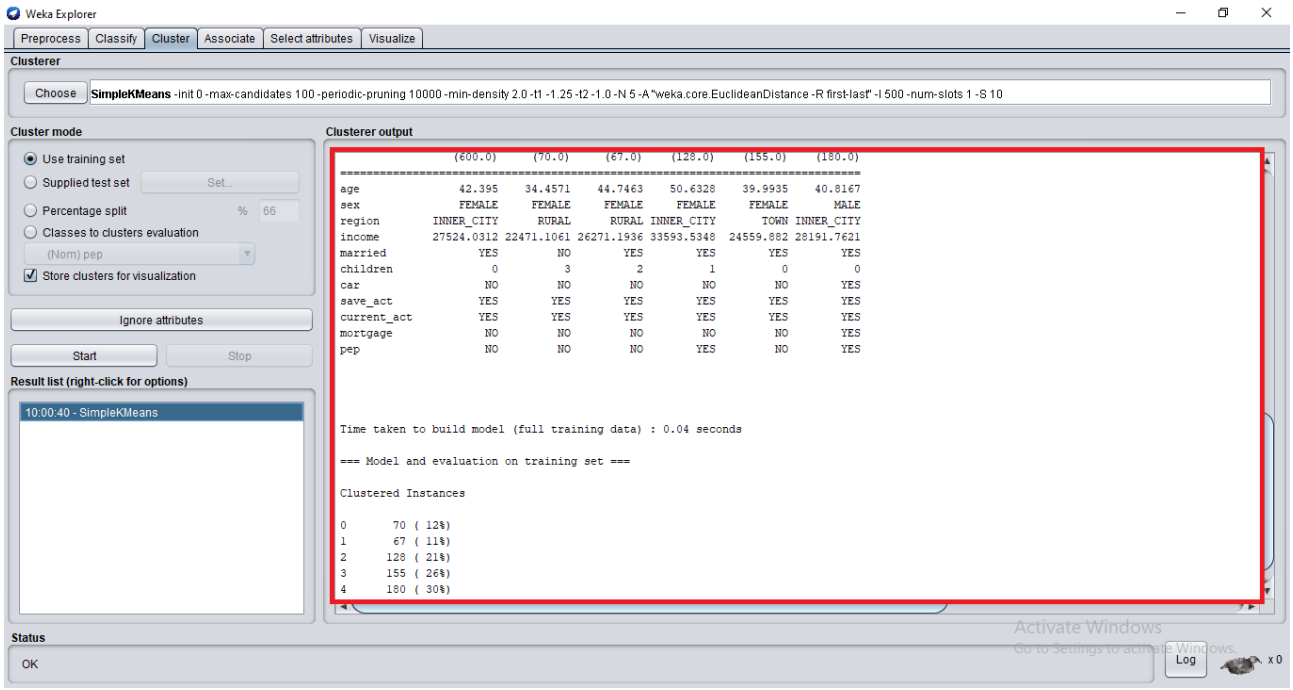
Hình 15: Chỉ định giải thuật SimpleKMeans

Bước 7: Lựa chọn chế độ mặc định cho toàn bộ dataset. Thay đổi tham số chi tiết cho giải thuật K-Means (chỉ định số lượng cụm K).



Hình 16: Lựa chọn K cho giải thuật K-Means

Bước 8: Chọn **Start** bắt đầu quá trình phân cụm. Kết quả phân cụm được hiển thị bên cạnh sau khi quá trình phân cụm hoàn tất.



Hình 17: Kết quả phân cụm dữ liệu

5 Phân tích thực nghiệm

5.1 Dataset

Lựa chọn tập dữ liệu phân tích đối tượng khách hàng của một ngân hàng chứa thông tin cơ bản của các khách hàng trong 600 đối tượng khảo sát thu được dataset: http://bis.net.vn/files/storage/bank.rar

Dữ liệu bao gồm các thuộc tính: **age, sex, region, income, married, children, car, save-act, current-act, mortgage, pep**.

5.2 Thống kê kết quả

Number of K	Iterations	WCSS
1	1	2363

age	sex	region	income	married	children	car	save-act	crr-act	mortgage	pep
25	F	RURAL	14505.3	NO	3	NO	YES	YES	NO	NO

Number of K	**Iterations**	**WCSS**
2	4	2016

age	sex	region	income	married	children	car	save-act	crr-act	mortgage	pep
25	F	RURAL	14505.3	NO	3	NO	YES	YES	NO	NO
61	F	RURAL	22942.9	YES	2	NO	YES	YES	NO	NO

Number of K	**Iterations**	**WCSS**
3	5	1833

age	sex	region	income	married	children	car	save-act	crr-act	mortgage	pep
25	F	RURAL	14505.3	NO	3	NO	YES	YES	NO	NO
61	F	RURAL	22942.9	YES	2	NO	YES	YES	NO	NO
54	F	CITY	31095.6	YES	2	NO	NO	YES	NO	YES

Number of K	**Iterations**	**WCSS**
4	7	1710

age	sex	region	income	married	children	car	save-act	crr-act	mortgage	pep
25	F	RURAL	14505.3	NO	3	NO	YES	YES	NO	NO
61	F	RURAL	22942.9	YES	2	NO	YES	YES	NO	NO
54	F	CITY	31095.6	YES	2	NO	NO	YES	NO	YES
36	F	TOWN	26920.8	YES	0	NO	NO	YES	NO	NO

Number of K	**Iterations**	**WCSS**
5	5	1702

age	sex	region	income	married	children	car	save-act	crr-act	mortgage	pep
25	F	RURAL	14505.3	NO	3	NO	YES	YES	NO	NO
61	F	RURAL	22942.9	YES	2	NO	YES	YES	NO	NO
54	F	CITY	31095.6	YES	2	NO	NO	YES	NO	YES
36	F	TOWN	26920.8	YES	0	NO	NO	YES	NO	NO
42	M	INNER	15499.9	YES	0	YES	NO	YES	YES	YES

Number of K	**Iterations**	**WCSS**
6	6	1604

age	sex	region	income	married	children	car	save-act	crr-act	mortgage	pep
25	F	RURAL	14505.3	NO	3	NO	YES	YES	NO	NO
61	F	RURAL	22942.9	YES	2	NO	YES	YES	NO	NO
54	F	CITY	31095.6	YES	2	NO	NO	YES	NO	YES
36	F	TOWN	26920.8	YES	0	NO	NO	YES	NO	NO
42	M	INNER	15499.9	YES	0	YES	NO	YES	YES	YES
50	M	TOWN	40972.9	NO	2	YES	YES	YES	YES	YES

Number of K	**Iterations**	**WCSS**
7	12	1531

age	sex	region	income	married	children	car	save-act	crr-act	mortgage	pep
25	F	RURAL	14505.3	NO	3	NO	YES	YES	NO	NO
61	F	RURAL	22942.9	YES	2	NO	YES	YES	NO	NO
54	F	CITY	31095.6	YES	2	NO	NO	YES	NO	YES
36	F	TOWN	26920.8	YES	0	NO	NO	YES	NO	NO
42	M	INNER	15499.9	YES	0	YES	NO	YES	YES	YES
50	M	TOWN	40972.9	NO	2	YES	YES	YES	YES	YES
48	F	INNER	42603.9	YES	0	NO	YES	YES	NO	NO

Number of K	**Iterations**	**WCSS**
8	7	1496

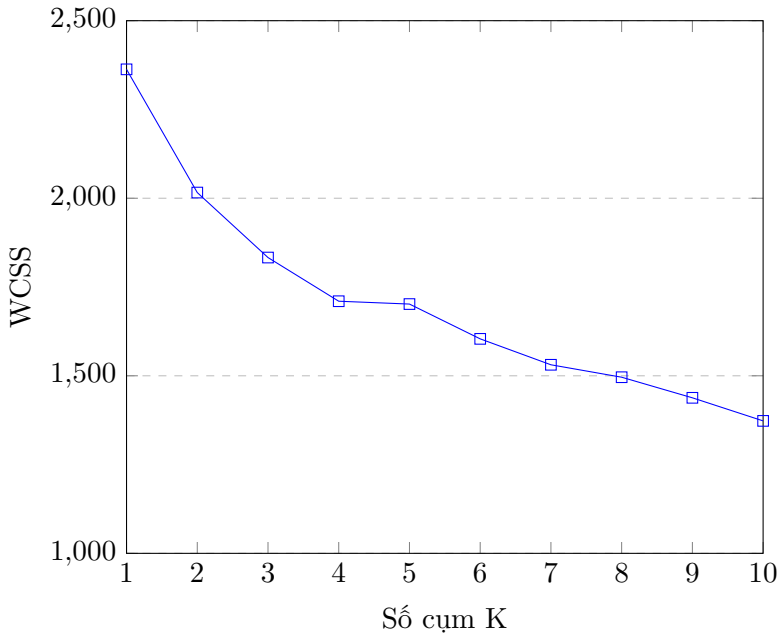
age	sex	region	income	married	children	car	save-act	crr-act	mortgage	pep
25	F	RURAL	14505.3	NO	3	NO	YES	YES	NO	NO
61	F	RURAL	22942.9	YES	2	NO	YES	YES	NO	NO
54	F	CITY	31095.6	YES	2	NO	NO	YES	NO	YES
36	F	TOWN	26920.8	YES	0	NO	NO	YES	NO	NO
42	M	INNER	15499.9	YES	0	YES	NO	YES	YES	YES
50	M	TOWN	40972.9	NO	2	YES	YES	YES	YES	YES
48	F	INNER	42603.9	YES	0	NO	YES	YES	NO	NO
64	F	INNER	34513.6	YES	1	NO	YES	YES	NO	YES

Number of K	**Iterations**	**WCSS**
9	8	1438

age	sex	region	income	married	children	car	save-act	crr-act	mortgage	pep
25	F	RURAL	14505.3	NO	3	NO	YES	YES	NO	NO
61	F	RURAL	22942.9	YES	2	NO	YES	YES	NO	NO
54	F	CITY	31095.6	YES	2	NO	NO	YES	NO	YES
36	F	TOWN	26920.8	YES	0	NO	NO	YES	NO	NO
42	M	INNER	15499.9	YES	0	YES	NO	YES	YES	YES
50	M	TOWN	40972.9	NO	2	YES	YES	YES	YES	YES
48	F	INNER	42603.9	YES	0	NO	YES	YES	NO	NO
64	F	INNER	34513.6	YES	1	NO	YES	YES	NO	YES
59	F	RURAL	51284.3	NO	0	YES	YES	YES	YES	NO

Number of K	**Iterations**	**WCSS**
10	7	1373

age	sex	region	income	married	children	car	save-act	crr-act	mortgage	pep
25	F	RURAL	14505.3	NO	3	NO	YES	YES	NO	NO
61	F	RURAL	22942.9	YES	2	NO	YES	YES	NO	NO
54	F	CITY	31095.6	YES	2	NO	NO	YES	NO	YES
36	F	TOWN	26920.8	YES	0	NO	NO	YES	NO	NO
42	M	INNER	15499.9	YES	0	YES	NO	YES	YES	YES
50	M	TOWN	40972.9	NO	2	YES	YES	YES	YES	YES
48	F	INNER	42603.9	YES	0	NO	YES	YES	NO	NO
64	F	INNER	34513.6	YES	1	NO	YES	YES	NO	YES
59	F	RURAL	51284.3	NO	0	YES	YES	YES	YES	NO
23	M	SUBURBAN	11073	YES	2	NO	YES	NO	NO	NO



Tài liệu

[1] Thuật toán K-Means với bài toán phân cụm dữ liệu , nguồn tham khảo: http://bis.net.vn/forums/t/374.aspx, truy cập: 05/24/2021.

[2] K-Means Clustering Algorithm, nguồn tham khảo: https://www.javatpoint.com/k-means-clustering-algorithm-in-machine-learning, truy cập: 05/24/2021.

[3] K-Means Clustering Algorithm: Applications, Types, Demos and Use Cases, nguồn tham khảo: https://www.simplilearn.com/tutorials/machine-learning-tutorial/k-means-clustering-algorithm, truy cập: 05/24/2021.

[4] Hiểu các thuật toán phân cụm K-mean, K-means ++ và K-medoids, nguồn tham khảo: https://ichi.pro/vi/hieu-cac-thuat-toan-phan-cum-k-mean-k-means-va-k-medoids, truy cập: 05/24/2021.

[5] Weka - Clustering, nguồn tham khảo: https://www.tutorialspoint.com/weka/weka_clustering.htm, truy cập: 05/24/2021.