

TRƯỜNG ĐẠI HỌC GIAO THÔNG VẬN TẢI



BÁO CÁO TỔNG KẾT

ĐỀ TÀI NGHIÊN CỨU KHOA HỌC CỦA SINH VIÊN
NĂM HỌC 2022-2023

THUẬT TOÁN K-MEANS TRONG PHÂN CỤM DỮ LIỆU VÀ ỨNG DỤNG

Sinh viên thực hiện

Đinh Vũ Lớp: Toán ứng dụng K62 Khoa: Khoa học cơ bản
Nguyễn Đức Anh Lớp: Toán ứng dụng K62 Khoa: Khoa học cơ bản

Người hướng dẫn: TS. Đặng Thị Mai

HÀ NỘI, 2023

Mục lục

1	Bài toán phân cụm	2
1.1	Định nghĩa	2
1.2	Độ đo sử dụng trong phân cụm	3
1.3	Các phương pháp phân cụm	4
1.3.1	Phương pháp phân hoạch	5
1.3.2	Phương pháp phân cấp	6
1.4	Ứng dụng của phân cụm	8
2	Thuật toán K-Means và ứng dụng	10
2.1	Thuật toán K-Means	10
2.1.1	Phân tích thuật toán	10
2.1.2	Các bước thuật toán	11
2.1.3	Phương pháp Elbow chọn số cụm	13
2.1.4	Ưu điểm và nhược điểm của thuật toán K-Means	14
2.2	Bài toán phân vùng ảnh	16
2.2.1	Các cách tiếp cận phân vùng ảnh	16
2.2.2	Kỹ thuật phân vùng ảnh	16
2.2.3	Sử dụng thuật toán K-Means	16
2.2.4	Đánh giá chất lượng ảnh sau khi phân vùng	17
2.3	Ứng dụng	18

Lời nói đầu

Phân cụm là kỹ thuật rất quan trọng trong khai phá dữ liệu, nó thuộc lớp các phương pháp Unsupervised Learning (học không giám sát) trong Machine Learning. Trong thuật toán này, ta không biết được nhãn của dữ liệu mà chỉ có dữ liệu đầu vào. Thuật toán unsupervised learning sẽ dựa vào cấu trúc của dữ liệu để thực hiện một công việc nào đó, ví dụ như phân cụm (clustering) hoặc giảm số chiều của dữ liệu (dimension reduction) để dễ dàng lưu trữ và tính toán dữ liệu. Phân cụm là quá trình nhóm các điểm dữ liệu trong cơ sở dữ liệu thành các cụm sao cho những điểm dữ liệu trong cùng một cụm có độ tương đồng lớn và những điểm không cùng một cụm có sự tương đồng là rất nhỏ.

K-Means là thuật toán rất quan trọng và được sử dụng phổ biến trong kỹ thuật phân cụm. Tự tưởng chính của thuật toán K-Means là tìm cách phân nhóm các đối tượng đã cho vào k cụm sao cho tổng bình phương khoảng cách giữa các đối tượng đến tâm nhóm là nhỏ nhất.

Báo cáo này gồm 2 chương, trong đó chương 1 nghiên cứu tổng quan về bài toán phân cụm, một số độ đo thường được sử dụng trong phân cụm đó là khoảng cách Euclidean và khoảng cách Manhattan. Cũng ở chương này chúng em trình bày một số phương pháp phổ biến trong phân cụm đó là phương pháp phân hoạch, phương pháp phân cấp. Cụ thể hơn ở chương 2 chúng em đi nghiên cứu về thuật toán K-Means, đây là một trong những thuật toán thuộc phương pháp phân hoạch.

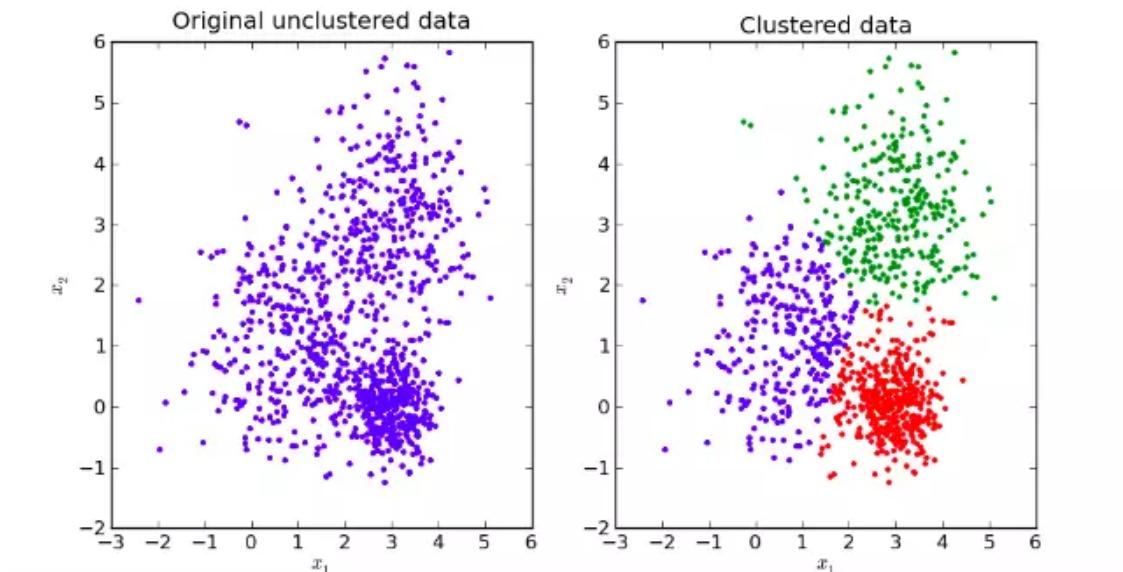
Trong chương 2, chúng em nghiên cứu về thuật toán K-Means và ứng dụng của nó. Trong thuật toán K-Means việc xác định số cụm K là rất khó, do đó chúng em đã tìm hiểu và nghiên cứu phương pháp Elbow, đây là phương pháp giúp xác định số cụm K, sao cho phân cụm đạt kết quả tốt. Từ đó chúng em áp dụng thuật toán K-Means kết hợp với sử dụng thuật toán Elbow vào bài toán phân vùng ảnh. Kết quả thu được thông qua ngôn ngữ lập trình Python.

Chương 1

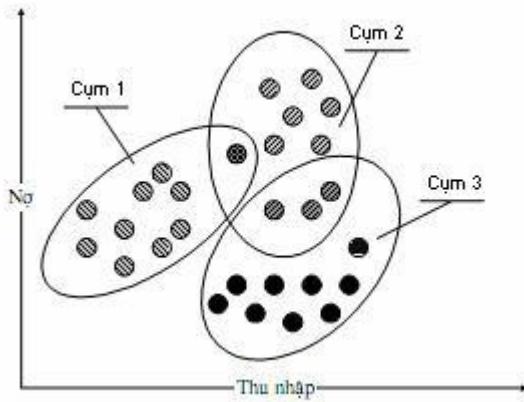
Bài toán phân cụm

1.1 Định nghĩa

Phân cụm là một cách nhóm các điểm dữ liệu nhiều chiều thành các cụm khác nhau sao cho các cụm chứa những điểm dữ liệu có tính tương đồng về đặc điểm (ví dụ: khoảng cách Euclidean). Phân cụm dữ liệu còn được sử dụng nhiều trong trí tuệ nhân tạo, hoặc các ứng dụng như phân vùng ảnh, lượng tử hóa màu sắc, khai phá dữ liệu, học máy, v.v. Một cụm thường được xác định bởi các tâm cụm (centroid). Phân cụm dữ liệu là một vấn đề tương đối khó trong nhận dạng mẫu không giám sát bởi dữ liệu có thể có nhiều hình dạng và kích cỡ. Dưới đây là 2 ví dụ đơn giản về phân cụm:



Hình 1.1.1: Ví dụ về phân cụm



Hình 1.1.2: Mô hình về phân cụm dựa trên tiêu chuẩn thu nhập và số nợ

Bài toán phân cụm được mô tả như sau:

Cho tập dữ liệu $X = \{x_1, x_2, \dots, x_n\}$ trong đó $x_i \in R^d$ là 1 tập gồm n đối tượng chứa đặc tính dữ liệu trong không gian d chiều. Ta cần phân tách tập dữ liệu thành k cụm: C_1, C_2, \dots, C_k rời nhau thỏa mãn điều kiện:

- Tất cả đối tượng phải được phân vào trong các cụm.

$$\bigcup_{i=1}^k C_i = Z$$

- Mỗi cụm có ít nhất một đối tượng.

$$C_i \neq \emptyset, \forall i \in [1, k]$$

- Mỗi đối tượng chỉ được nằm trong một cụm duy nhất.

$$C_i \cap C_j \forall i \neq j$$

1.2 Độ đo sử dụng trong phân cụm

Để xác định mức độ tương đồng trong phân cụm, ta thường xác định thông qua hàm giá trị "khoảng cách" giữa các đối tượng. Các độ đo sau đây được xác định trong không gian metric. Một không gian metric là một không gian được trang bị một hàm đo khoảng cách d nào đó thỏa mãn:

- $d(x, y) > 0; \forall x \neq y$
- $d(x, y) = 0; \forall x = y$
- $d(x, y) = d(y, x); \forall x, y$
- $d(x, z) \leq d(x, y) + d(y, z); \forall x, y, z$

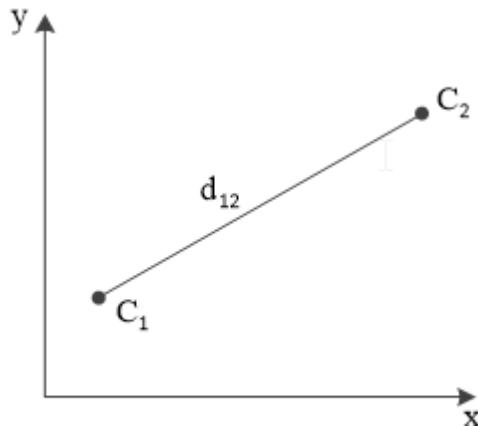
Trên thực tế ta có thể sử dụng nhiều công thức đo khoảng cách khác nhau, và mỗi một công thức cho ra một kết quả cụm khác nhau. Hiện nay khoảng cách Euclidean vẫn đang được sử dụng phổ biến nhất và được định nghĩa như sau:

$$d(a, b) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$$

với x, y là 2 đối tượng n thuộc tính, $x = (x_1, x_2, \dots, x_n)$ và $y = (y_1, y_2, \dots, y_n)$.

Khoảng cách Euclidean là trường hợp đặc biệt (với $\alpha = 2$) của khoảng cách Minkowski. Khoảng cách Minkowski được định nghĩa như sau:

$$d(a, b) = \left(\sum_{i=1}^n |x_i - y_i|^\alpha \right)^{1/\alpha}$$



Hình 1.2.1: Khoảng cách Euclidean

Ngoài ra, ta còn có khoảng cách Manhattan cũng là trường hợp đặc biệt (với $\alpha = 1$):

$$d(a, b) = \left(\sum_{i=1}^n |x_i - y_i| \right)$$

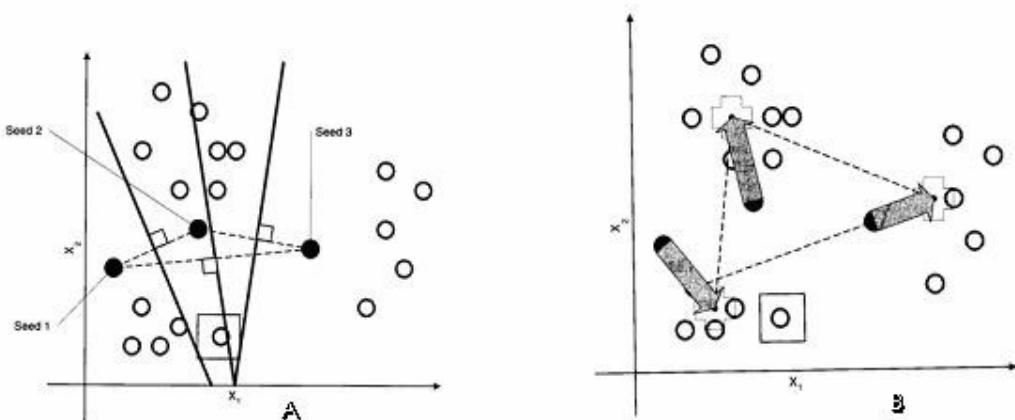
1.3 Các phương pháp phân cụm

Phần lớn các thuật toán phân cụm đều dựa trên 2 phương pháp phổ biến: phân hoạch (partitioning) và phân cấp (hierarchical). Ngoài ra còn một vài phương pháp khác như phương pháp dựa trên mật độ (density-based), dựa trên lưới (grid-based). Ở đây báo cáo này sẽ chỉ đưa ra tổng quan về 2 phương pháp phân cụm phổ biến nhất là phân hoạch và phân cấp.

1.3.1 Phương pháp phân hoạch

Ý tưởng của phương pháp phân hoạch như sau: Cho tập D gồm n đối tượng và một tham số đầu vào k được xác định bởi người dùng. Thuật toán phân hoạch sẽ chọn k đối tượng đại diện cho k cụm (k đối tượng đại diện có thể được chọn ngẫu nhiên hoặc theo một tiêu chuẩn của người sử dụng). Với một đối tượng dữ liệu q sẽ được đưa vào cụm có đối tượng đại diện gần với q nhất. Sau đó, đối tượng đại diện của mỗi cụm sẽ được tính lại dựa vào những điểm dữ liệu thuộc cụm đó. Thông thường thì đối tượng đại diện được xác định sao cho khoảng cách từ đối tượng đại diện đến điểm xa nhất là nhỏ nhất có thể được.

Hình dưới mô tả quá trình phân hoạch với $k = 3$. Khởi tạo bởi hình A với 3 đối tượng đại diện là 3 điểm đậm được lựa chọn ngẫu nhiên. Kế tiếp mỗi đối tượng dữ liệu được đưa vào cụm mà khoảng cách từ điểm đó tới đối tượng đại diện của cụm là nhỏ nhất. Với mỗi cụm tìm đối tượng đại diện cho cụm đó (lấy đối tượng dữ liệu mới là điểm trung bình của tất cả các đối tượng dữ liệu thuộc cụm). Quá trình trên được lặp lại cho đến khi các đối tượng đại diện của tất cả các cụm là không thay đổi.



Hình 1.3.1: Ví dụ quá trình phân hoạch với $k = 3$

Mô hình thuật toán phân cụm phân hoạch:

- **Đầu vào:** Số cụm k và tập D gồm n đối tượng.
- **Đầu ra:** Tập các cụm.

Phát biểu thuật toán **Partition(D, k)**:

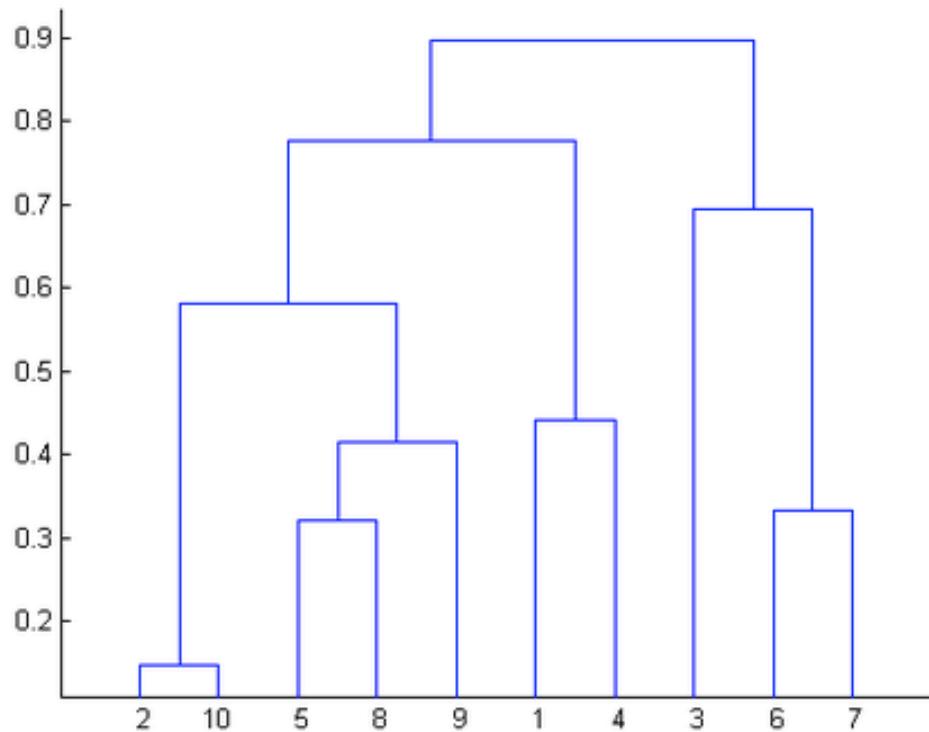
1. Chọn ngẫu nhiên k tâm bất kỳ O . Đặt $i = 0$.
 2. Với mỗi điểm dữ liệu $p \in D$ thì tìm đối tượng đại diện gần nhất và đưa p vào cụm đó.
 3. Tính lại đối tượng đại diện của các cụm O^{i+1} dựa vào các điểm dữ liệu thuộc cụm.
 4. Nếu $O^{i+1} = O^i$ thì dừng lại. Trong trường hợp ngược lại cho và quay lại bước 2.
- $O^i = \{o_1^{(i)}, o_2^{(i)}, \dots, o_k^{(i)}\}$ là tập các đối tượng đại diện của k cụm.

Với phương pháp này, số cụm được thiết lập là đặc trưng được lựa chọn trước. Phương pháp phân hoạch thích hợp với bài toán tìm các cụm trong không gian 2D. Ngoài ra, phương pháp xem xét đến khoảng cách cơ bản giữa các điểm dữ liệu để xác định chúng có quan hệ gần nhau, hoặc không gần nhau hay không có quan hệ. Nhược điểm của phương pháp này là đòi hỏi phải đưa vào tham số k và không xử lý trên bộ dữ liệu thuộc cụm có hình dạng phức tạp hoặc mật độ phân bố dày đặc.Thêm vào đó, thuật toán có độ phức tạp tính toán lớn khi cần xác định kết quả tối ưu.

Các thuật toán trong phương pháp phân hoạch: K-Means, PAM (Partitioning Around Medoids), CLARA (Clustering LARge Application), CLARANS (Clustering Large Applications based upon RANdomized Search),...

1.3.2 Phương pháp phân cấp

Thuật toán phân cấp sẽ sinh ra một đồ thị dạng cây (hay dendrogram) sử dụng chiến lược hợp nhất hoặc chiến lược phân chia. Theo phương pháp này, chúng tạo ra những biểu diễn phân cấp trong đó các cụm ở mỗi cấp của hệ thống phân cấp được tạo bằng cách hợp nhất các cụm ở cấp độ thấp hơn bên dưới. Ở cấp thấp nhất, mỗi cụm chứa một quan sát. Ở cấp cao nhất, chỉ có một cụm chứa tất cả dữ liệu.



Hình 1.3.2: Ví dụ một dendrogram

Các chiến lược phân cụm phân cấp chia thành hai mô hình cơ bản: Hợp nhất (agglomerative) và phân chia (divisive). Trục hoành thể hiện index của các quan sát trong

nhóm được phân vào một cụm, trong khi tục tung là giá trị thước đo sự khác biệt giữa các cụm. Một cụm được đại diện bởi một node mà toàn bộ các quan sát khác nếu thuộc cụm thì đều liên kết tới node đó. Như vậy chúng ta có thể nhận thấy rằng các cụm có sự phân cấp dựa vào level của node. Khi kẻ một đường thẳng nằm ngang cắt toàn bộ các đường thẳng đứng ta sẽ thu được các cụm tương ứng với các node nằm gần nhất bên dưới đường thẳng. Bất kì hai cụm nào trong số chúng sẽ không chồng lấn nhau.

Thuật toán phân cụm phân cấp được xây dựng trên bộ dữ liệu có kích thước N thì sẽ trải qua tổng cộng N bước phân chia. Có hai chiến lược phân chia chính phụ thuộc vào chiều di chuyển trên biểu đồ dendrogram:

- **Chiến lược hợp nhất:** Chiến lược này sẽ đi theo chiều bottom-up (từ dưới lên trên). Quá trình phân cụm bắt đầu ở dưới cùng tại các node lá (còn gọi là leaf node hoặc terminal node). Ban đầu mỗi quan sát sẽ được xem là một cụm tách biệt được thể hiện bởi một node lá. Ở mỗi level chúng ta sẽ tìm cách hợp một cặp cụm thành một cụm duy nhất nhằm tạo ra một cụm mới ở level cao hơn tiếp theo. Cụm mới này tương ứng với các node quyết định (non-leaf node). Như vậy sau khi hợp cụm thì số lượng cụm ít hơn. Một cặp được chọn để hợp nhất sẽ là những cụm trung gian không giao nhau.
- **Chiến lược phân chia:** Chiến lược này sẽ thực hiện theo chiều top-down. Tức là phân chia bắt đầu từ node gốc của đồ thị. Node gốc bao gồm toàn bộ các quan sát, tại mỗi level chúng ta phân chia một cách để qui các cụm đang tồn tại tại level đó thành hai cụm mới. Phép phân chia được tiến hành sao cho tạo thành hai cụm mới mà sự tách biệt giữa chúng là lớn nhất. Sự tách biệt này sẽ được đo lường thông qua một thước đo khoảng cách.

Như vậy đồ thị của chiến lược phân chia và chiến lược hợp nhất đều là cây nhị phân, chúng chỉ khác biệt về chiều thực hiện thuật toán. Node gốc của cây nhị phân sẽ bao gồm toàn bộ các quan sát và cây nhị phân bao gồm N node lá đại diện cho N quan sát từ bộ dữ liệu. Mỗi một node quyết định bao gồm hai node con. Quá trình phân chia thì hai node con thể hiện kết quả được phân chia từ node cha và quá trình hợp nhất thì node cha là thể hiện kết quả sau khi gộp hai node con.

Phương pháp này có một vài ưu điểm:

- không phụ thuộc tham số đầu vào
- không nhạy cảm với nhiễu trong tập dữ liệu
- có thể sinh ra các cụm giống như con người quan sát

Tuy nhiên, phương pháp này cũng có những nhược điểm như:

- Chi phí tính toán cao (độ phức tạp $O(N^2 \log N)$, chi phí lưu trữ $O(N^2)$). Do đó, phương pháp này không được sử dụng cho các tập dữ liệu quá lớn. Đây là một trong những điểm mấu chốt khiến phương pháp phân hoạch được ưu chuộng hơn trong thực tế.
- Đối tượng đưa vào một cụm không thể chuyển sang cụm khác
- Phương pháp này có thể không tác động được các cụm chồng lên nhau do thiếu thông tin về hình dạng tổng quan hoặc kích cỡ các cụm.

1.4 Ứng dụng của phân cụm

Kỹ thuật phân cụm có thể được sử dụng trong rất nhiều lĩnh vực trong thực tế. Lấy một ví dụ về lĩnh vực thương mại là phân khúc thị trường như sau:

Giả sử ta là một nhà phân tích thị trường cho một công ty nghiên cứu thị trường. Công ty của ta đã thu thập dữ liệu từ một số khảo sát trực tuyến về thói quen mua sắm của người tiêu dùng trong một khu vực nhất định. Dữ liệu bao gồm thông tin về tuổi, giới tính, thu nhập, sở thích mua sắm và thông tin về các giao dịch trực tuyến của khách hàng. ta muốn phân cụm dữ liệu để nhận biết các nhóm khách hàng có đặc điểm tương tự về hành vi mua sắm.

Bằng cách sử dụng phân cụm dữ liệu, ta có thể chia nhóm khách hàng thành các cụm dựa trên các yếu tố chung như sở thích mua sắm, mức độ chi tiêu, và tuổi. Ví dụ, ta có thể tìm thấy nhóm khách hàng trẻ tuổi, có thu nhập cao, thích mua sắm hàng hiệu và thường xuyên mua sắm trực tuyến. Hoặc ta có thể nhận ra một nhóm khách hàng lớn hơn, có thu nhập trung bình, yêu thích mua sắm sản phẩm giá rẻ và thường thăm các cửa hàng truyền thống.

Khi ta đã phân cụm dữ liệu, ta có thể áp dụng thông tin này để phân tích thị trường. Ví dụ, ta có thể đưa ra các đề xuất về chiến lược tiếp thị, nhằm mục tiêu quảng cáo đến từng nhóm khách hàng cụ thể, hoặc tìm ra những xu hướng mua sắm mới để dự đoán và đáp ứng nhu cầu của khách hàng.

Tóm lại, phân cụm dữ liệu trong phân tích thị trường có thể giúp ta hiểu rõ hơn về các đặc điểm và hành vi mua sắm của khách hàng, từ đó tạo ra các chiến lược tiếp thị hiệu quả và tối ưu hóa hoạt động kinh doanh

Một ứng dụng khác trong lĩnh vực khác như y tế là chẩn đoán ung thư. Giả sử có một tập dữ liệu với thông tin từ 500 bệnh nhân bao gồm kết quả xét nghiệm máu, thông tin về tế bào ung thư và kết quả xét nghiệm hình ảnh (ví dụ: X-quang, siêu âm). Mục tiêu là chẩn đoán ung thư phổi và phân loại bệnh nhân thành hai nhóm: nhóm ung thư phổi và nhóm không ung thư phổi. Các bước để áp dụng phân cụm dữ liệu vào chẩn đoán ung thư phổi như sau:

- Chuẩn bị dữ liệu: Tiền xử lý dữ liệu bằng cách loại bỏ các giá trị thiếu, chuẩn hóa dữ liệu và chọn các biến quan trọng liên quan đến ung thư phổi (ví dụ: kết quả xét nghiệm CEA, xét nghiệm huyết học).
- Phân cụm dữ liệu: Sử dụng một thuật toán phân cụm như K-means hoặc hierarchical clustering, áp dụng phân cụm vào tập dữ liệu. Thuật toán sẽ phân chia các bệnh nhân thành các cụm dựa trên đặc trưng tương tự trong dữ liệu xét nghiệm.
- Phân tích các nhóm phân cụm: Xem xét các đặc trưng chung và khác biệt trong từng nhóm phân cụm. So sánh các chỉ số ung thư phổi quan trọng như tỷ lệ CEA, kích thước khối u, vị trí và sự lây lan của ung thư trong từng nhóm. Điều này giúp xác định các đặc điểm đáng chú ý và tiềm năng của ung thư phổi.
- Xác nhận kết quả: So sánh kết quả phân cụm với các phương pháp chẩn đoán khác như kết quả biopsy hoặc xét nghiệm hình ảnh. Đánh giá độ chính xác của phân cụm dữ liệu trong việc chẩn đoán ung thư phổi và so sánh với các phương pháp chẩn đoán khác.

- Đưa ra kết luận: Dựa trên phân tích và xác nhận kết quả, đưa ra kết luận về khả năng mắc phải ung thư phổi của từng nhóm phân cụm.

Ngoài ra còn một số ứng dụng khác có thể kể đến như là: phân vùng ảnh, phân tích dữ liệu thống kê, gợi ý tìm kiếm,...

Chương 2

Thuật toán K-Means và ứng dụng

2.1 Thuật toán K-Means

Thuật ngữ K-Means được J. MacQueen giới thiệu vào năm 1967 và phát triển dựa trên ý tưởng của H. Steinhaus đề xuất năm 1956. Thuật toán này sử dụng giá trị trung bình (mean) của các đối tượng trong cụm làm tâm của cụm đó. Tư tưởng chính của thuật toán K-Means là tìm cách phân nhóm các đối tượng đã cho vào k cụm (k là số các cụm được xác định trước, k nguyên dương) sao cho tổng bình phương khoảng cách giữa các đối tượng đến tâm cụm là nhỏ nhất.

Tổng bình phương khoảng cách giữa các đối tượng đến tâm cụm còn gọi là hàm tiêu chuẩn (criterion function) được tính bởi công thức:

$$E = \sum_{i=1}^k \sum_{x \in C_i} |x - m_i|^2$$

Trong đó, x là một điểm, m_i là giá trị trung bình của cụm C_i .

2.1.1 Phân tích thuật toán

Giả sử có N điểm dữ liệu là $X = (x_1, x_2, \dots, x_n) \in \mathbb{R}^{d \times N}$ và $k < N$ là số cụm chúng ta muốn phân chia. Chúng ta cần tìm các tâm cụm $m_1, m_2, \dots, m_k \in \mathbb{R}^{d \times 1}$ và nhãn của mỗi điểm dữ liệu.

Với mỗi điểm dữ liệu x_i đặt $y_i = (y_{i1}, y_{i2}, \dots, y_{ik})$ là nhãn vector của nó, trong đó nếu x_i được phân vào cụm k thì $y_{ik} = 1$ và $y_{ij} = 0$, $\forall j \neq k$. Điều này có nghĩa là có đúng 1 phần tử của vector y_i là bằng 1 (tương ứng với cụm của x_i), các phần tử còn lại bằng 0. Ví dụ: nếu một điểm dữ liệu có nhãn vector là $(1, 0, 0, \dots, 0)$ thì nó thuộc vào cụm 1 và $(0, 1, 0, \dots, 0)$ thì nó thuộc vào cụm 2, v.v

Khi đó sai số của toàn bộ điểm dữ liệu được phân vào cụm k là:

$$\|x_i - m_k\|_2^2$$

Vì x_i được phân vào cụm k nên $y_{ik} = 1, y_{ij} = 0, \forall j \neq k$. Biểu thức trên tương đương:

$$y_{ij}\|x_i - m_k\|_2^2 = \sum_{j=1}^k y_{ij}\|x_i - m_j\|_2^2$$

Do đó sai só cho toàn dữ liệu là:

$$\mathcal{L}(Y, M) = \sum_{i=1}^N \sum_{j=1}^k y_{ij}\|x_i - m_j\|_2^2$$

Trong đó $Y = (y_1, y_2, \dots, y_N)$, $M = (m_1, m_2, \dots, m_k)$.

Như vậy để tìm Y và M ta cần giải bài toán tối ưu sau:

$$Y, M = \operatorname{argmin}_{Y, M} \sum_{i=1}^N \sum_{j=1}^k y_{ij}\|x_i - m_j\|_2^2$$

thỏa mãn $y_{ij} \in \{0, 1\} \forall i, j$ và $\sum_{j=1}^k y_{ij} = 1 \forall i$. Ta có thể giải bài toán tối ưu trên bằng cách xen kẽ giải Y và M khi biến còn lại được cố định. Giả sử khi xác định được nhãn Y của từng điểm dữ liệu thì bài toán tìm tâm cụm M trở thành:

$$m_j = \operatorname{argmin}_{m_j} \sum_{j=1}^k y_{ij}\|x_i - m_j\|_2^2$$

Đặt $l(m_j)$ là hàm bên trong dấu argmin , ta có đạo hàm:

$$\frac{\partial l(m_j)}{\partial m_j} = 2 \sum_{i=1}^N y_{ij}(m_j - x_i)$$

Giải phương trình đạo hàm bằng 0 ta có:

$$\begin{aligned} m_j \sum_{i=1}^N y_{ij} &= \sum_{i=1}^N y_{ij}x_i \\ \Rightarrow m_j &= \frac{\sum_{i=1}^N y_{ij}x_i}{\sum_{i=1}^N y_{ij}} \end{aligned}$$

Như vậy ta thấy rằng: m_j là trung bình cộng của các điểm trong cụm j ;

2.1.2 Các bước thuật toán

Thuật toán K-Means được phát biểu như sau:

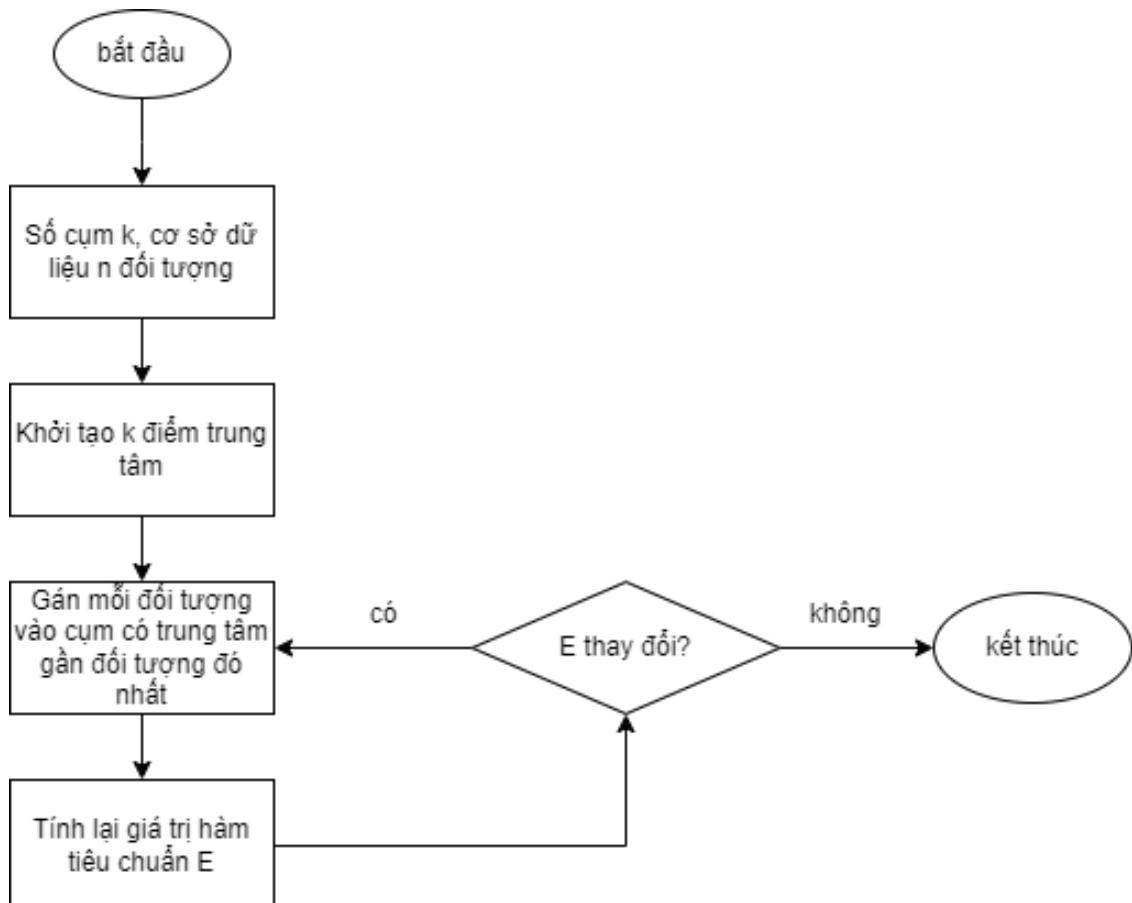
Đầu vào : Số các cụm k , cơ sở dữ liệu gồm n đối tượng.

Đầu ra : Tập k cụm mà có giá trị hàm tiêu chuẩn E nhỏ nhất.

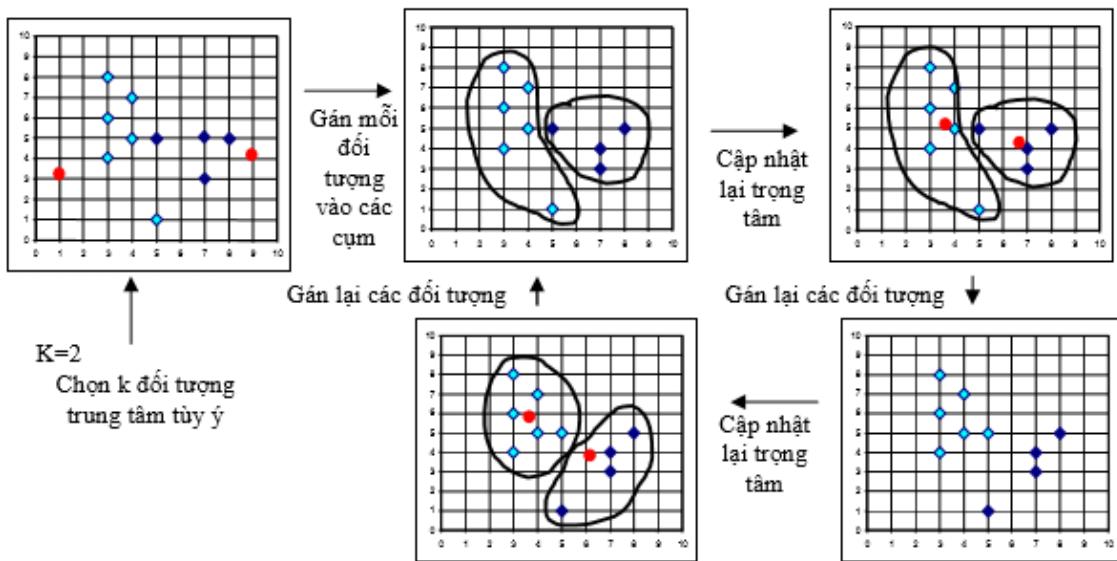
Phương pháp :

1. Khởi tạo k điểm trung tâm cụm bằng cách chọn k đối tượng tùy ý.
2. Lặp các bước:
 - (a) Gán mỗi đối tượng vào cụm có trung tâm gần đối tượng đó nhất, hình thành một tập các cụm mới.
 - (b) Tính lại giá trị E của mỗi cụm theo các đối tượng mới thu được sau bước 2(a).
3. Thuật toán dừng khi giá trị E không thay đổi.

Tại bước 1, thực hiện chọn ngẫu nhiên k điểm từ cơ sở dữ liệu các đối tượng cần phân cụm là điểm trung tâm cho k cụm. Sau đó, thực hiện lần lượt tính khoảng cách từ điểm trung tâm tới các điểm, so sánh xem giá trị nào nhỏ hơn (có nghĩa gần trung tâm hơn) thì gán điểm đó vào cụm chứa điểm trung tâm đó. Tiếp đến tính lại giá trị hàm tiêu chuẩn E , nếu giá trị mới nhỏ hơn giá trị cũ thì thay đổi giá trị E . Thuật toán lặp lại các bước cho đến khi giá trị E không thay đổi nữa. Để tính khoảng cách giữa điểm trung tâm tới các điểm, ta dùng độ đo khoảng cách Euclidean.



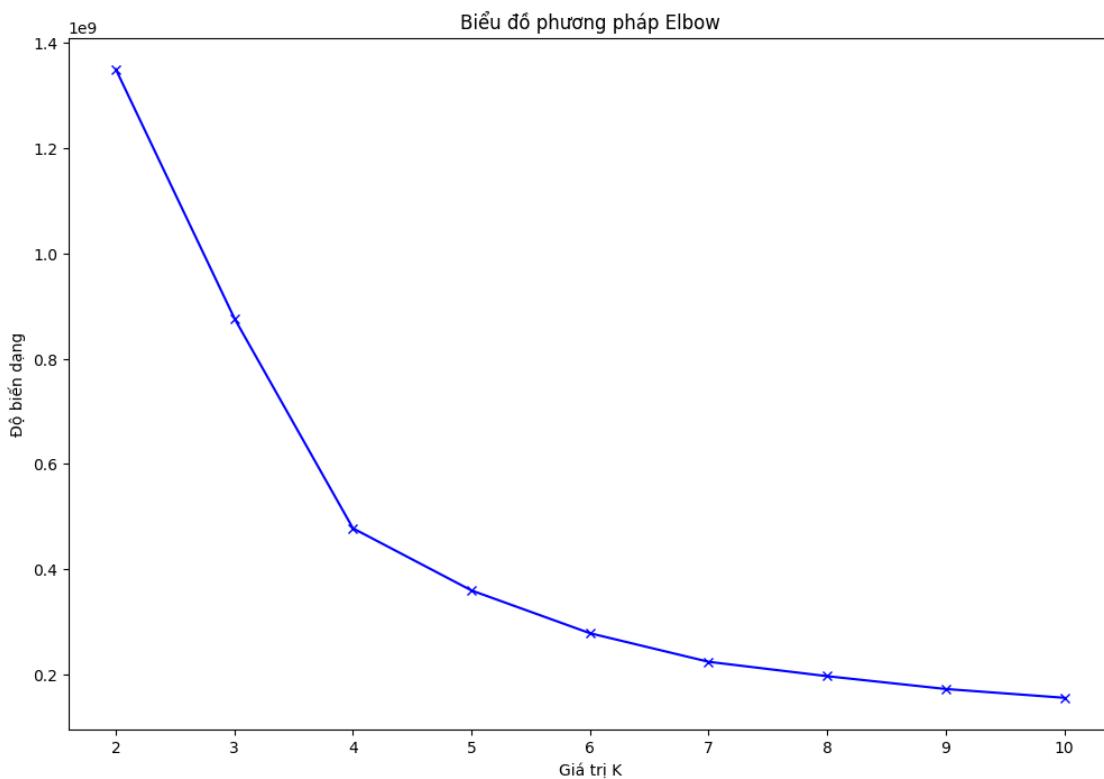
Hình 2.1.1: Sơ đồ khái quát các bước thực hiện thuật toán K-Means



Hình 2.1.2: Minh họa các bước thực hiện thuật toán K-means

2.1.3 Phương pháp Elbow chọn số cụm

Trong thuật toán k-Means thì chúng ta cần phải xác định trước số cụm. Câu hỏi đặt ra là đâu là số lượng cụm cần phân chia tốt nhất đối với một bộ dữ liệu cụ thể? Phương pháp Elbow là một cách giúp ta lựa chọn được số lượng các cụm phù hợp dựa vào đồ thị trực quan hóa bằng cách nhìn vào sự suy giảm của hàm biến dạng và lựa chọn ra điểm khuỷu tay (elbow point). Hàm biến dạng thực chất là tổng bình phương của từng điểm đến tâm cụm, hay chính là hàm tiêu chuẩn.



Hình 2.1.3: Ví dụ đồ thị hàm biến dạng của thuật toán K-Means

Điểm khuỷu tay là điểm mà ở đó tốc độ suy giảm của hàm biến dạng sẽ thay đổi nhiều nhất. Tức là kể từ sau vị trí này thì gia tăng thêm số lượng cụm cũng không giúp hàm biến dạng giảm đáng kể. Nếu thuật toán phân chia theo số lượng cụm tại vị trí này sẽ đạt được tính chất phân cụm một cách tổng quát nhất mà không gặp các hiện tượng vị khớp (overfitting). Trong hình trên thì ta thấy vị trí của điểm khuỷu tay chính là $k = 4$ vì khi số lượng cụm lớn hơn 4 thì tốc độ suy giảm của hàm biến dạng dường như không đáng kể so với trước đó.

Phương pháp Elbow là một phương pháp thường được sử dụng để lựa chọn số lượng cụm phân chia hợp lý dựa trên biểu đồ, tuy nhiên có một số trường hợp chúng ta sẽ không dễ dàng phát hiện vị trí của Elbow, đặc biệt là đối với những bộ dữ liệu mà qui luật phân cụm không thực sự dễ dàng được phát hiện. Nhưng nhìn chung thì phương pháp Elbow vẫn là một phương pháp tốt nhất được ứng dụng trong việc tìm kiếm số lượng cụm cần phân chia.

2.1.4 Ưu điểm và nhược điểm của thuật toán K-Means

Ưu điểm

- Chi phí tính toán thấp nên chúng có thể được sử dụng với bộ dữ liệu lớn.
- Thuật toán hoạt động tốt kể cả với dữ liệu chưa được gán nhãn.
- Kết quả đầu ra dễ hiểu, phù hợp khi ta cần đánh giá nhanh một phân đoạn dữ liệu.

- Thuật toán cũng dễ dàng cài đặt và hiện nay có rất nhiều thư viện Python đã tích hợp sẵn thuật toán để sẵn sàng cho việc sử dụng.

Nhược điểm

- Thuật toán chỉ áp dụng với dữ liệu có thuộc tính số và khám phá các cụm có dạng hình cầu, không thích hợp với việc tìm các cụm có hình dáng không lồi hay các cụm có hình dáng khác xa nhau, nhạy cảm với các phần tử ngoại lai, phần tử nhiễu, phần tử cận biên cụm.
- Việc chọn lựa tập điểm trung tâm ban đầu cũng ảnh hưởng nhiều đến chất lượng cụm sinh ra. Trên thực tế chưa có một giải pháp tối ưu nào để chọn các tham số đầu vào, giải pháp thường được sử dụng nhất là thử nghiệm với các giá trị đầu vào k khác nhau rồi sau đó chọn giải pháp tốt nhất.

2.2 Bài toán phân vùng ảnh

Phân vùng ảnh (Image segmentation) là một phương pháp mà trong đó, hình ảnh kỹ thuật số được chia thành nhiều nhóm con khác nhau được gọi là segments. Mục tiêu của phân vùng ảnh là làm giảm độ phức tạp của hình ảnh, giúp cho quá trình xử lý hoặc phân tích hình ảnh sau đó trở nên đơn giản hơn. Nói một cách dễ hiểu, phân vùng là dán nhãn cho từng pixel. Tất cả các yếu tố hình ảnh hoặc pixel thuộc cùng một danh mục sẽ có chung một nhãn. Ví dụ: Đối với bài toán phát hiện đối tượng, thay vì xử lý toàn bộ hình ảnh, máy có thể chỉ thực hiện trên một đoạn được chọn bởi thuật toán phân vùng. Điều này sẽ ngăn máy xử lý toàn bộ hình ảnh, do đó làm giảm thời gian suy luận.

2.2.1 Các cách tiếp cận phân vùng ảnh

- Cách tiếp cận tương đồng (Similarity approach), có nghĩa là phát hiện sự tương đồng giữa các pixel hình ảnh để tạo thành một phân đoạn, dựa trên một ngưỡng. Các thuật toán học máy như phân cụm thường dựa trên kiểu tiếp cận này để phân vùng một hình ảnh.
- Cách tiếp cận gián đoạn (Discontinuity approach): Cách tiếp cận này dựa trên sự gián đoạn của các giá trị cường độ pixel trong hình ảnh. Các kỹ thuật phát hiện đường, điểm và cạnh sử dụng kiểu tiếp cận gián đoạn để thu được các kết quả phân vùng trung gian. Kết quả này sau đó có thể được xử lý để cho ra hình ảnh được phân vùng cuối cùng.

2.2.2 Kỹ thuật phân vùng ảnh

Một vài kỹ thuật phân vùng ảnh có thể kể đến như: phân vùng dựa trên ngưỡng (Threshold Based Segmentation), phân vùng dựa trên cạnh (Edge Based Segmentation), phân vùng dựa trên khu vực (Region-Based Segmentation), phân vùng dựa trên kỹ thuật phân cụm (Clustering Based Segmentation), phân vùng dựa trên mạng nơron nhân tạo (Artificial Neural Network Based Segmentation).

Một trong những phương pháp hiệu quả nhất và đang được sử dụng hiện nay là phân cụm và cũng là phương pháp được sử dụng trong báo cáo này (cụ thể là thuật toán K-Means).

2.2.3 Sử dụng thuật toán K-Means

Cho một ảnh kích cỡ $x \times y$ và ảnh phải được phân thành k cụm. Gọi $p(x, y)$ là điểm ảnh đầu vào và C_k là các tâm cụm. Bài toán được phát biểu lại như sau:

1. Khởi tạo k tâm cụm.
2. Với mỗi điểm ảnh, ta tính khoảng cách Euclidean d từ tâm đến các điểm ảnh đó.

$$d = \|p(x, y) - C_k\|$$

3. Đưa các điểm ảnh vào các cụm gần nhất dựa theo khoảng cách d .

4. Tính lại tâm cụm theo công thức

$$C_k = \frac{1}{k} \sum_{y \in C_k} \sum_{x \in C_k} p(x, y)$$

5. Lặp lại quá trình đến khi sai số chấp nhận được.

6. Chuyển các cụm điểm ảnh thành ảnh.

2.2.4 Đánh giá chất lượng ảnh sau khi phân vùng

Sai số toàn phương trung bình

Sai số toàn phương trung bình (Mean squared error) là trung bình của bình phương các sai số, tức là sự khác biệt giữa các ước lượng và những gì được đánh giá. MSE được sử dụng như một độ đo tiêu chuẩn trong xử lý hình ảnh cho biết hình ảnh đầu ra bị lệch bao nhiêu so với ảnh gốc. MSE càng nhỏ thì chất lượng ảnh đầu ra càng tốt.

Công thức của MSE với ảnh có kích cỡ $m \times n$, I là ảnh gốc và k là ảnh đầu ra:

$$MSE = \frac{1}{mn} \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} [I(i, j) - K(i, j)]^2$$

Tỉ số tín hiệu cực đại trên nhiễu

Tỉ số tín hiệu cực đại trên nhiễu (Peak signal-to-noise ratio) là tỉ lệ giữa giá trị năng lượng tối đa của một tín hiệu và năng lượng nhiễu ảnh hướng đến độ chính xác của thông tin. PSNR được sử dụng để đo chất lượng tín hiệu khôi phục của các thuật toán nén có mất mát dữ liệu. PSNR càng lớn thì chất lượng ảnh đầu ra càng tốt.

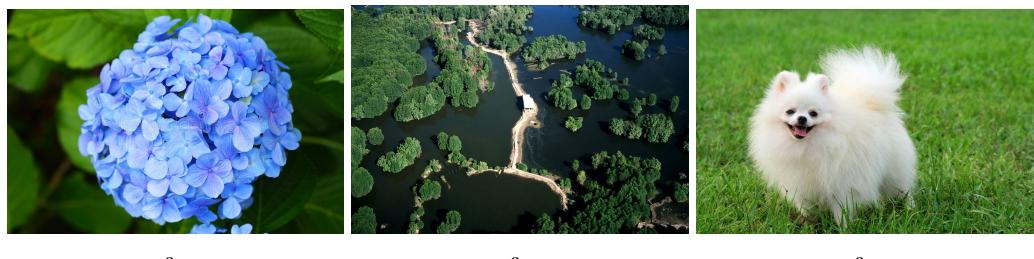
Công thức của PSNR được xác định thông qua MSE với ảnh có kích cỡ $m \times n$, I là ảnh gốc và k là ảnh đầu ra:

$$PSNR = 20 \cdot \log_{10} \left(\frac{255}{\sqrt{MSE}} \right)$$

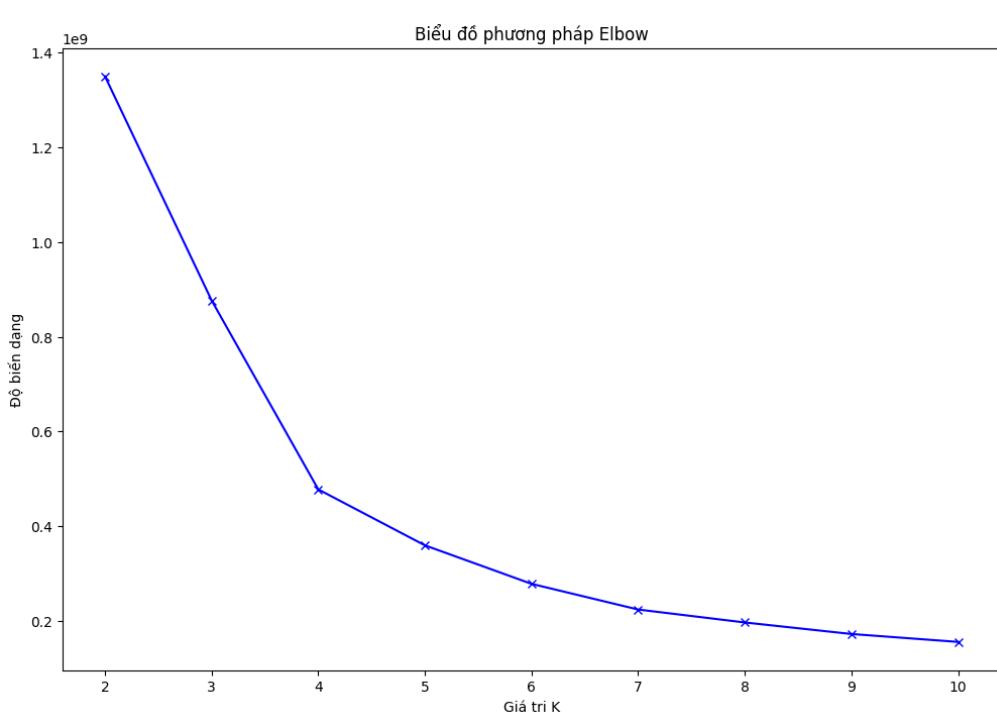
Giá trị thông thường của PSNR nằm từ 30 đến 50 dB, giá trị càng cao thì càng tốt.

2.3 Ứng dụng

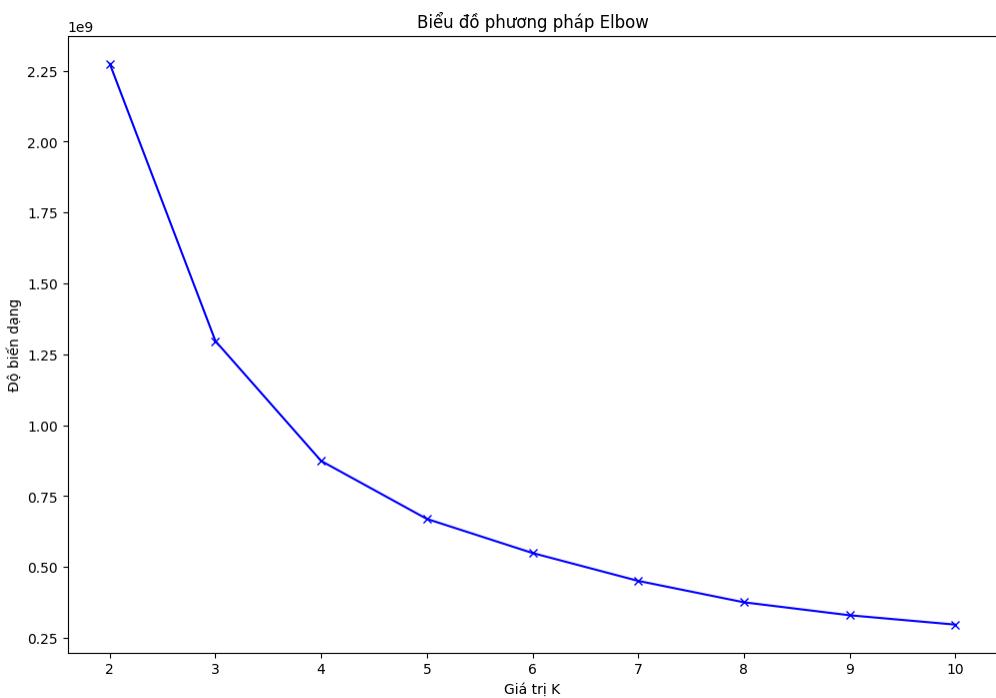
Về cách thức tiến hành, đầu tiên chúng em tiến hành sử dụng phương pháp Elbow để chọn ra giá trị k hợp lý nhất rồi sau đó kiểm tra và so sánh các giá trị MSE và PSNR của vài giá trị k khác.



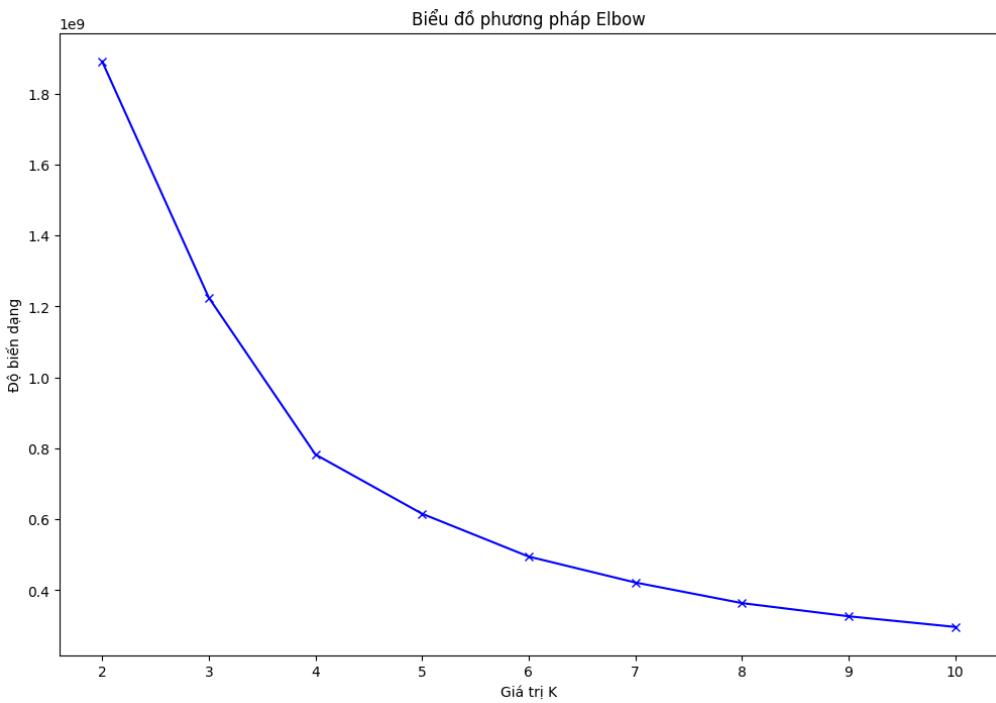
Hình 2.3.1: Ảnh gốc



Hình 2.3.2: Biểu đồ phương pháp Elbow cho biết độ biến dạng theo k (ảnh 1)



Hình 2.3.3: Biểu đồ phương pháp Elbow cho biết độ biến dạng theo k (ảnh 2)



Hình 2.3.4: Biểu đồ phương pháp Elbow cho biết độ biến dạng theo k (ảnh 3)

Từ biểu đồ trên ta có thể thấy được $k = 4$ là điểm khuỷu tay của ảnh 1, $k = 3$ là điểm khuỷu tay của ảnh 2, và $k = 4$ là điểm khuỷu tay của ảnh 3.

Sau khi tiến hành kiểm tra các giá trị PSNR và MSE tương ứng với từng k chúng em thu được bảng giá trị như sau:

k	MSE	PSNR
2	97.45	28.24 db
3	81.21	29.03 db
4	75.01	29.38 db
5	73.89	29.44 db
6	65.31	29.98 db
7	62.09	30.20 db
8	60.94	30.28 db
9	58.39	30.47 db
10	53.64	30.84 db

Bảng 2.1: Bảng giá trị MSE và PSNR với các k tương ứng (ảnh 1)

k	MSE	PSNR
2	91.71	28.51 db
3	86.12	28.78 db
4	83.62	28.91 db
5	76.45	29.30 db
6	69.68	29.70 db
7	70.40	29.66 db
8	61.53	30.24 db
9	58.27	30.48 db
10	57.28	30.55 db

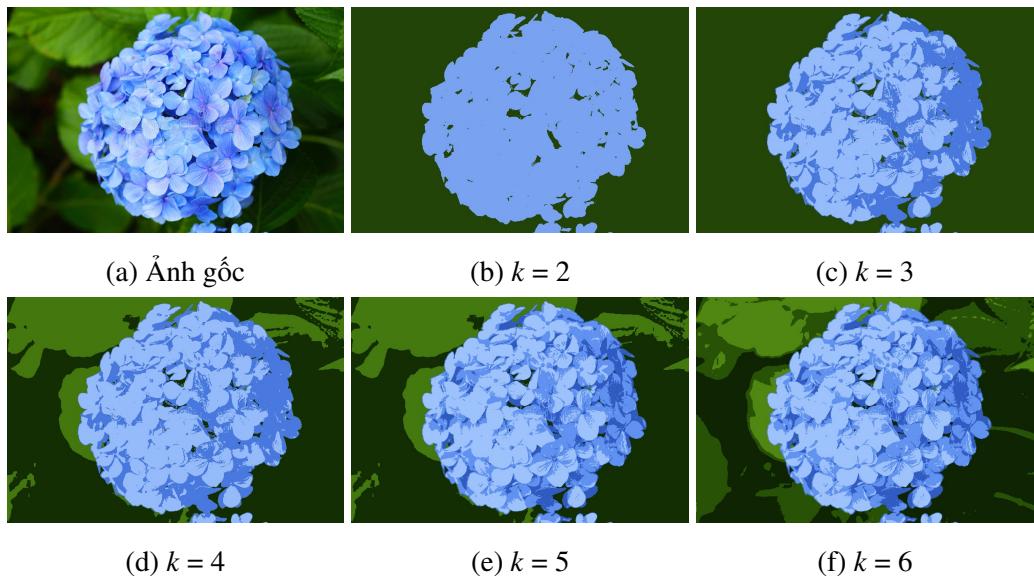
Bảng 2.2: Bảng giá trị MSE và PSNR với các k tương ứng (ảnh 2)

k	MSE	PSNR
2	92.18	28.48 db
3	84.82	28.85 db
4	80.39	29.08 db
5	74.41	29.41 db
6	71.74	29.57 db
7	68.68	29.76 db
8	64.47	30.04 db
9	61.99	30.21 db
10	59.53	30.38 db

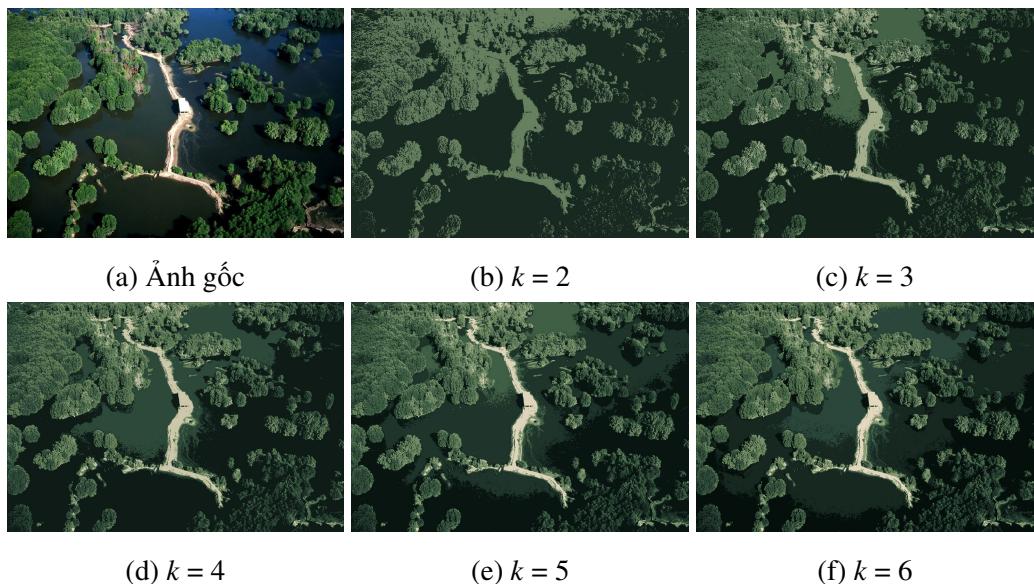
Bảng 2.3: Bảng giá trị MSE và PSNR với các k tương ứng (ảnh 3)

Từ các bảng trên, ta có thể thấy khi k tăng thì MSE giảm và PSNR tăng.

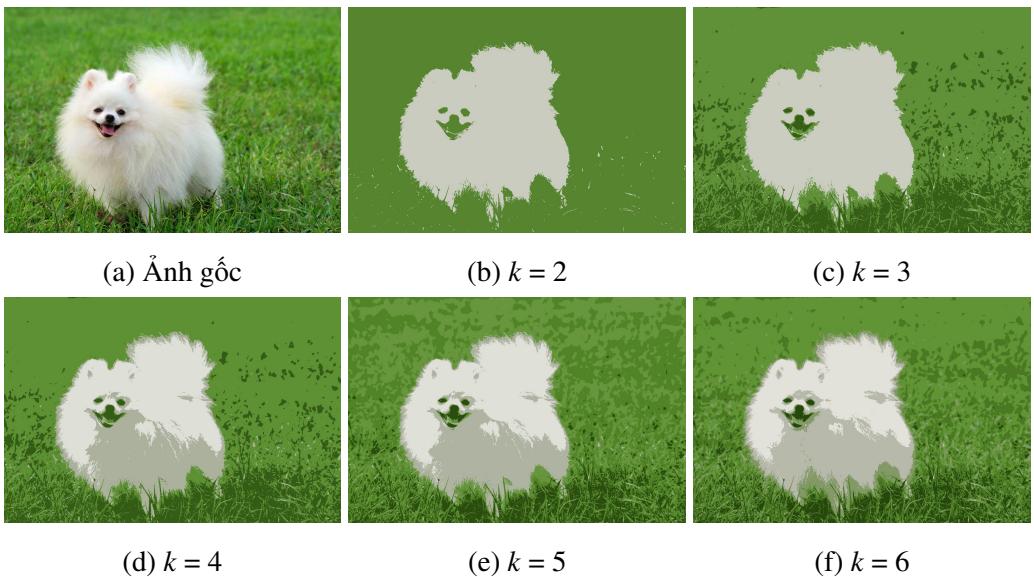
Cuối cùng chúng em biểu diễn một vài ảnh sau khi được phân vùng bằng thuật toán K-Means với các giá trị k lân cận để theo dõi sự thay đổi cũng như kiểm tra tính hiệu quả khi ta tăng giá trị k .



Hình 2.3.5: Ảnh 1 gốc và ảnh 1 sau khi được phân vùng bằng thuật toán K-Means



Hình 2.3.6: Ảnh 2 gốc và ảnh 2 sau khi được phân vùng bằng thuật toán K-Means



Hình 2.3.7: Ảnh 3 gốc và ảnh 3 sau khi được phân vùng bằng thuật toán K-Means

Qua đây ta có thể thấy được là khi tăng k qua điểm khuỷu tay thì hiệu quả tăng lên không rõ rệt.

Kết luận

Qua đề tài nghiên cứu này, chúng em đã nghiên cứu và hiểu được về bài toán phân cụm, các độ đo khoảng cách Euclidean và khoảng cách Manhattan, hiểu được phương pháp phân hoạch và phương pháp phân cấp, ưu điểm và nhược điểm của từng phương pháp. Biết được ứng dụng của phân cụm vào thực tế về các lĩnh vực khác nhau thương mại, y tế, giáo dục.

Chúng em đã tìm hiểu, nghiên cứu thuật toán K-Means và hiểu rõ được thuật toán cũng như ứng dụng được nó. Hiểu được rõ ưu điểm và nhược điểm của phương pháp Elbow từ đó áp dụng nó vào tìm được số tâm cụm. Chúng em đã sử dụng ngôn ngữ lập trình Python để áp dụng thuật toán K-Means có kết hợp với phương pháp elbow để ứng dụng trong bài toán phân vùng ảnh. Chất lượng ảnh thu được sau khi áp dụng K-Means đã được đánh giá thông qua sai số toàn phương trung bình(MSE) và tỉ số tín hiệu cực đại trên nhiễu (PSNR).

Tài liệu tham khảo

1. Jasmine Irani, Nitin Pise, Madhura Phatak(2016): Clustering Techniques and the Similarity Measures used in Clustering: A Survey.
2. Nameirakpam Dhanachandra, Khumanthem Manglem and Yambem Jina Chanu(2015): Image Segmentation using K-means Clustering Algorithm and Subtractive Clustering Algorithm.
3. Mahamed G.H. Omran, Andries P Engelbrecht and Ayed Salman(2007): An Overview of Clustering Methods
4. Peak signal-to-noise ratio: https://en.wikipedia.org/wiki/Peak_signal-to-noise_ratio
5. Mean squared error: https://en.wikipedia.org/wiki/Mean_squared_error
6. Metric space: https://en.wikipedia.org/wiki/Metric_space
7. Phương pháp Elbow: https://phamdinhkhanh.github.io/deepai-book/ch_ml/KMeans.html#phuong-phap-elbow-trong-lua-chon-so-cum