Phân cụm dữ liệu

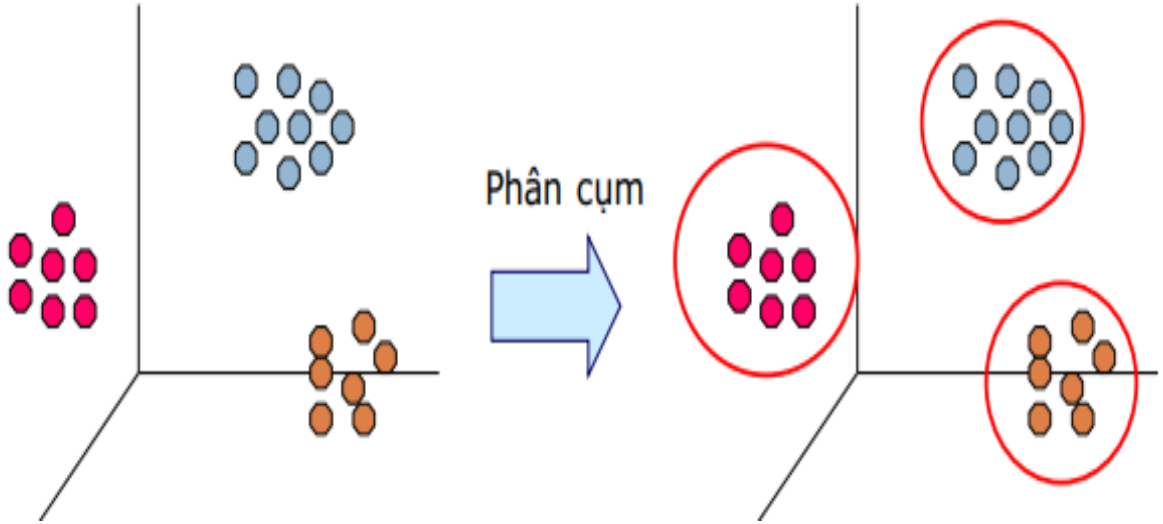
TS. Nguyễn Quốc Tuấn

Nội dung

- Tổng quan về phân cụm dữ liệu
- Thuật toán K-means

Tổng quan về phân cụm dữ liệu

- Phân cụm (Clustering) : là quá trình phân chia tập dữ liệu ban đầu thành các cụm dữ liệu sao cho các phần tử trong một cụm thì "tương tự" (Similar) với nhau các phần tử trong các cụm khác nhau sẽ "không tương tự" (dissimilar) với nhau.
- Mục tiêu phân cụm: nhóm các đối tượng tương tự, nhờ đó phát hiện cấu trúc ẩn của dữ liệu



Tổng quan về phân cụm

- Lĩnh vực ứng dụng
  - Nghiên cứu thị trường (Marketing): Xác định các nhóm khách hàng (khách hàng tiềm năng, khách hàng lớn, phân loại và dự đoán hành vi khách hàng, …)
  - Sinh học (Biology): Phân nhóm động vật, thực vật, ...
  - Tài chính, Bảo hiểm (Finance and Insurance): Phân nhóm các đối tượng sử dụng bảo hiểm và các dịch vụ tài chính, dự đoán xu hướng (trend) của khách hàng,
  - Xử lý ảnh
  - …

Tổng quan về phân cụm

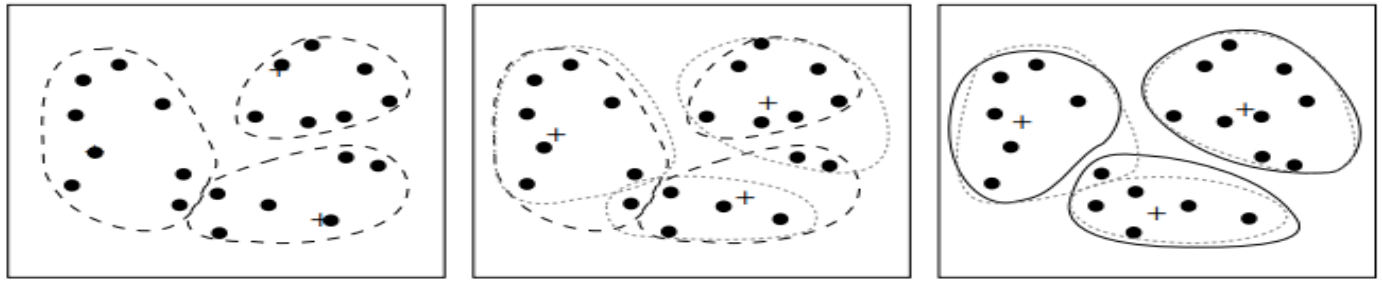
- Một số phương pháp phân cụm
  - **Phân hoạch** (partitioning): phân hoạch một tập hợp dữ liệu có n phần tử thành k nhóm (k <= n) và sau đó đánh giá chúng dựa trên các tiêu chí xác định.
    Thuật toán điển hình: K-means, K-medoids, CLARANS,…
  - **Phân cấp** (hierarchical): xây dựng một phân cấp trên cơ sở các đối tượng dữ liệu đang xem xét.
    Thuật toán điển hình: BIRCH, Chameleon,…
  - **Phân cụm dựa trên mật độ** (Density-Based): nhóm các đối tượng dữ liệu dựa trên hàm mật độ xác định.
    Thuật toán điển hình: DBSCAN, OPTICS,…
  - **Phân cụm dựa trên lưới** (Grid-Based): thích hợp với dữ liệu nhiều chiều, dựa trên cấu trúc dữ liệu lưới để phân cụm.
    Thuật toán điển hình: STING, CLIQUE,…

Nội dung

- Tổng quan về phân cụm dữ liệu
- Thuật toán K-means

Thuật toán K-Means

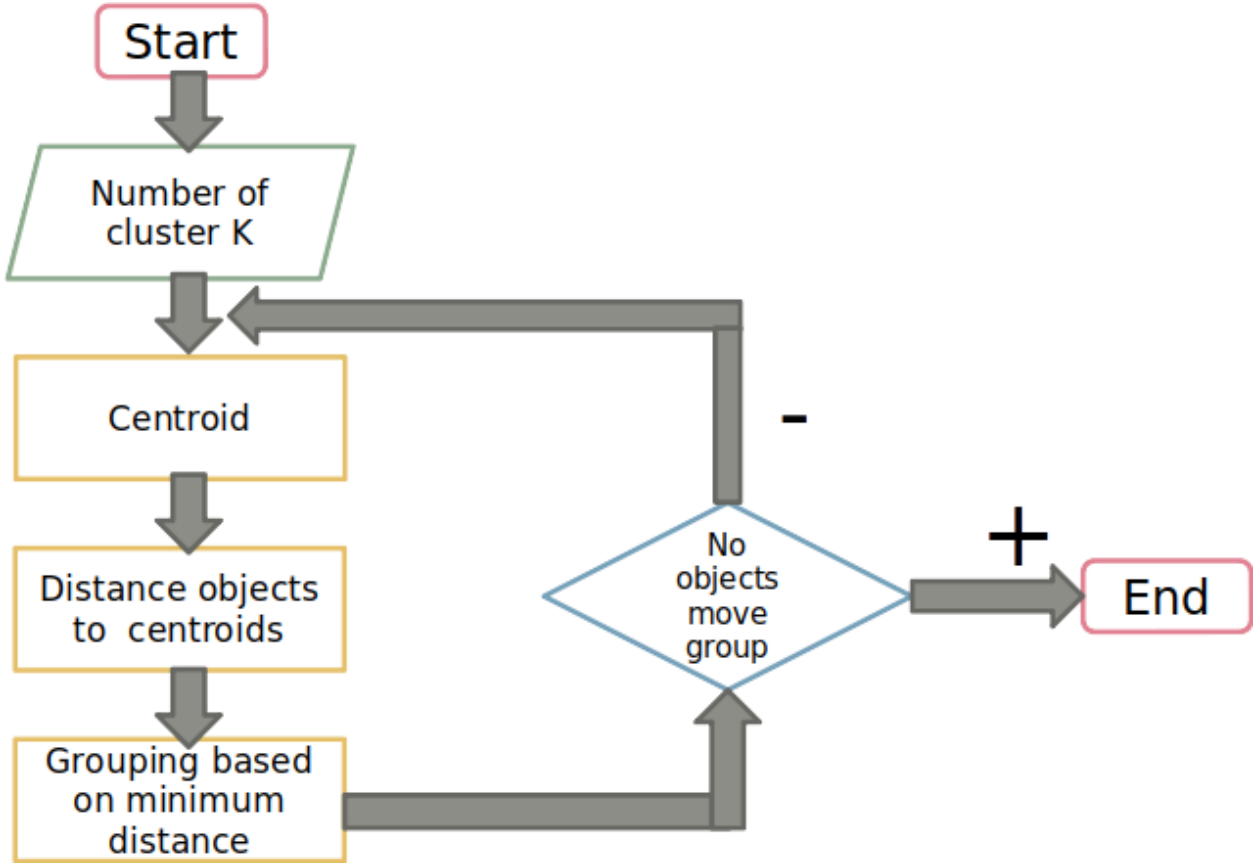
- Ý tưởng:
  - Mỗi cụm được đại diện bởi trọng tâm.
  - Một đối tượng được phân vào một cụm nếu khoảng cách từ đối tượng đó đến trọng tâm của cụm đang xét là nhỏ nhất.
  - Sau đó
  - Quá trì



(a) Initial clustering **(b)** Iterate **(c)** Final clustering

Thuật toán K-Means

- Sơ đồ thuật toán



Thuật toán K-Means

- Thuật toán

Input:

- k: the number of clusters,
- D: a data set containing n objects.

Output: A set of k clusters.

Method:

(1) arbitrarily choose k objects from D as the initial cluster centers;
(2) **repeat**
(3) (re)assign each object to the cluster to which the object is the most similar, based on the mean value of the objects in the cluster;
(4) update the cluster means, that is, calculate the mean value of the objects for each cluster;
(5) **until** no change;

Thuật toán K-Means

- Phương pháp tính khoảng cách

  $i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ and $j = (x_{j1}, x_{j2}, \ldots, x_{jp})$

  - Euclidean: $d(i,j) = \sqrt{(|x_{i_1} - x_{j_1}|^2 + |x_{i_2} - x_{j_2}|^2 + \ldots + |x_{i_p} - x_{j_p}|^2)}$

  - Manhattan: $d(i,j) = |x_{i_1} - x_{j_1}| + |x_{i_2} - x_{j_2}| + \ldots + |x_{i_p} - x_{j_p}|$

  - Minkowski: $d(i,j) = \sqrt[h]{|x_{i1} - x_{j1}|^h + |x_{i2} - x_{j2}|^h + \cdots + |x_{ip} - x_{jp}|^h}$

Thuật toán K-Means

- Ưu điểm:
  - Đơn giản
    - Rất dễ cài đặt
    - Rất dễ hiểu
  - Hiệu quả
    - Độ phức tạp về thời gian ~ O(r.k.t)
      - r: Tổng số các ví dụ (kích thước của tập dữ liệu)
      - k: Tổng số cụm thu được
      - t: Tổng số bước lặp (của quá trình phân cụm)
    - Nếu cả 2 giá trị k và t đều nhỏ, thì giải thuật k-means được xem như là có độ phức tạp ở mức tuyến tính.
  - k means là giải thuật phân cụm được dùng phổ biến nhất

Thuật toán K-Means

- Nhược điểm:
  - Giá trị k (số cụm thu được) phải được xác định trước
  - Giải thuật k-means cần xác định cách tính điểm trung bình (centroid) của một cụm
    - Đối với các thuộc tính định danh (nominal attributes), giá trị trung bình có thể được xác định là giá trị phổ biến nhất
  - Giải thuật k-means nhạy cảm (gặp lỗi) với các ví dụ ngoại lai (outliers)
    - Các ví dụ ngoại lai là các ví dụ (rất) khác biệt với tất các ví dụ khác
    - Các ví dụ ngoại lai có thể do lỗi trong quá trình thu thập/lưu dữ liệu
    - Các ví dụ ngoại lai có các giá trị thuộc tính (rất) khác biệt với các giá trị thuộc tính của các ví dụ khác.

Thuật toán K-Means

- Ví dụ
  - Phân cụm dữ liệu sau với k=2

Đối tượng	Thuộc tính 1(x)	Thuộc tính 2(y)
A	1	2
B	2	2
C	4	4
D	6	5

Thuật toán K-Means

- Bước 1: Khởi tạo
  - Chọn 2 trọng tâm ban đầu:
    $C_1(1,2) \equiv A$ và $C_2(2,2) \equiv B$, thuộc 2 cụm 1 và 2
- Lặp 1:
  - Tính toán khoảng cách

	$C_1(1, 2)$	$C_2(2, 2)$	Cụm
A(1, 2)	**0**	1	1
B(2, 2)	1	**0**	2
C(4, 4)	3.606	**2.828**	2
D(6, 5)	5.381	**5**	2

- Cụm 1 : A(1,2); Cụm 2: B(2,2), C(4,4), D(6,5)
- Cập nhật lại trọng tâm: C_1 = A(1,2); $C_2(x,y)$= ((2+4+6)/3, (2+4+5/3)=(4,11/3)

Thuật toán K-Means

- Lặp 2:
  - Tính toán khoảng cách

	c1(1, 2)	c2(4, 11/3)	Cụm
A(1, 2)	**0**	3.442	1
B(2, 2)	**1**	2.603	**1**
C(4, 4)	3.606	**0.333**	2
D(6, 5)	5.381	**2.404**	2

- Cụm 1 : A(1,2), B(2,2); Cụm 2: C(4,4), D(6,5)
- Cập nhật lại trọng tâm: C_1 (x,y) = ((1+2)/2,(2+2)/2) = (3/2,2)

  C_2(x,y)= ((4+6)/2, (4+5/2)=(5,9/2)

Thuật toán K-Means

- Lặp 3:
  - Tính lại khoảng cách

	C_1(3/2, 2)	C_2(5, 9/2)	Cụm
A(1, 2)	**0.5**	4.717	1
B(2, 2)	**0.5**	3.905	1
C(4, 4)	3.202	**1.118**	2
D(6, 5)	5.408	**1.118**	2

- Cụm 1 : A(1,2), B(2,2); Cụm 2: C(4,4), D(6,5)
- Cập nhật lại trọng tâm: C_1 (x,y) = ((1+2)/2,(2+2)/2) = (3/2,2)
  C_2(x,y)= ((4+6)/2, (4+5/2)=(5,9/2)
- Không có sự thay đổi tâm cụm => Dừng

Bài tập k-Means

Bài 1. Cho tập dữ liệu sau

	X1	X_2
A	1	2
B	2	2
C	2	3
D	3	3
E	3	4
F	2	4

Tiến hành phân cụm tập dữ liệu trên với k=2 với tâm khởi tạo cho 2 cụm là A và F

Bài tập k-Means

Bài 1. Cho tập dữ liệu sau

	X1	X_2
A	1	2
B	2	2
C	2	3
D	3	3
E	3	4
F	2	4

Bài tập k-Means

Bài 2. Cho tập dữ liệu các điểm có tọa độ (x,y) như sau

$A_1(2,10), A_2(2,5), A_3(8,4), B_1(5,8), B_2(7,5), B_3(6,4), C_1(1,2), C_2(4,9)$

Sử dụng hàm tính khoảng cách Mahattan, phân nhóm các điểm trên vào 3 cụm. Giả sử A_1, B_1, và C_1 là tâm khởi tạo ban đầu của 3 cụm tương ứng. Sử dụng thuật toán k-means để đưa ra kết quả sau:

(a) Tâm của 3 cụm sau vòng lặp đầu tiên

(b) Tâm cuối cùng của 3 cụm.

Một số lưu ý

- Kết quả phụ thuộc vào sự lựa chọn ngẫu nhiên ban đầu của các tâm cụm.
- Để có được kết quả tốt trong thực tế, người ta thường chạy thuật toán k-means nhiều lần với các tâm cụm ban đầu khác nhau.
- Phương pháp k-means chỉ có thể được áp dụng khi giá trị trung bình của một tập đối tượng được xác định.
- Cần chỉ định số lượng cụm trước

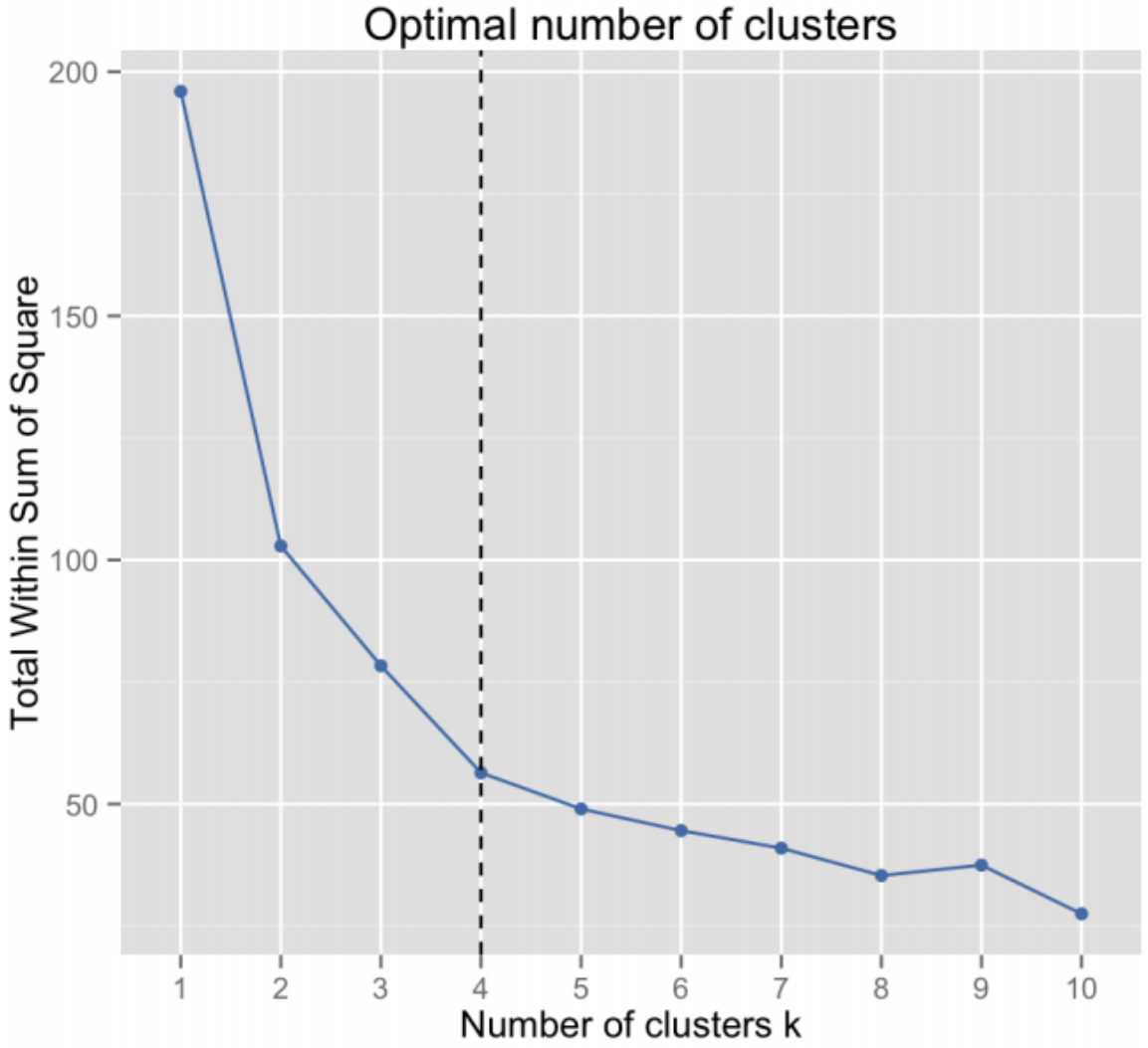
Một số cách chọn số lượng cụm tối ưu

Silhouette Analysis

Elbow Method

Elbow

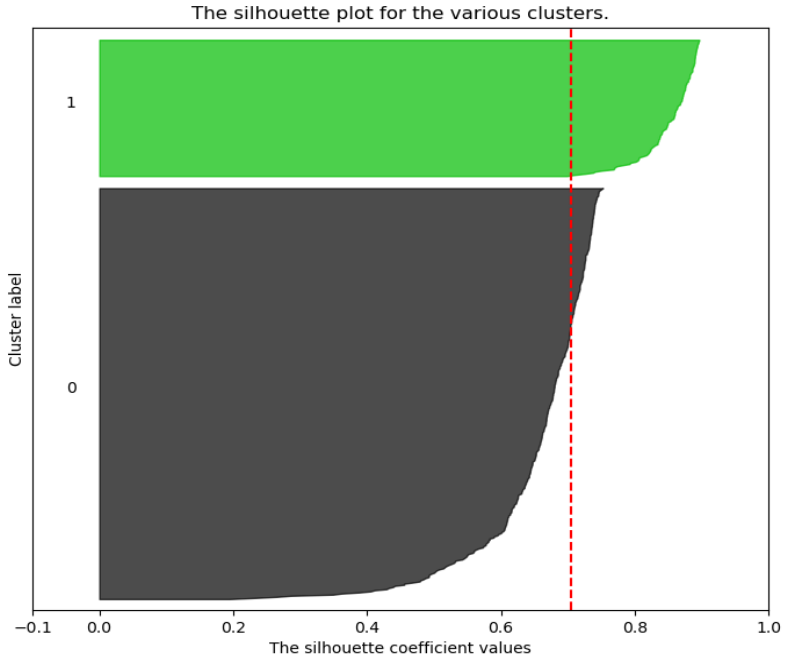
- Distortion: Trung bình cộng bình phương khoảng cách giữa tâm cụm đến các điểm còn lại (thường là Euclidean distance)

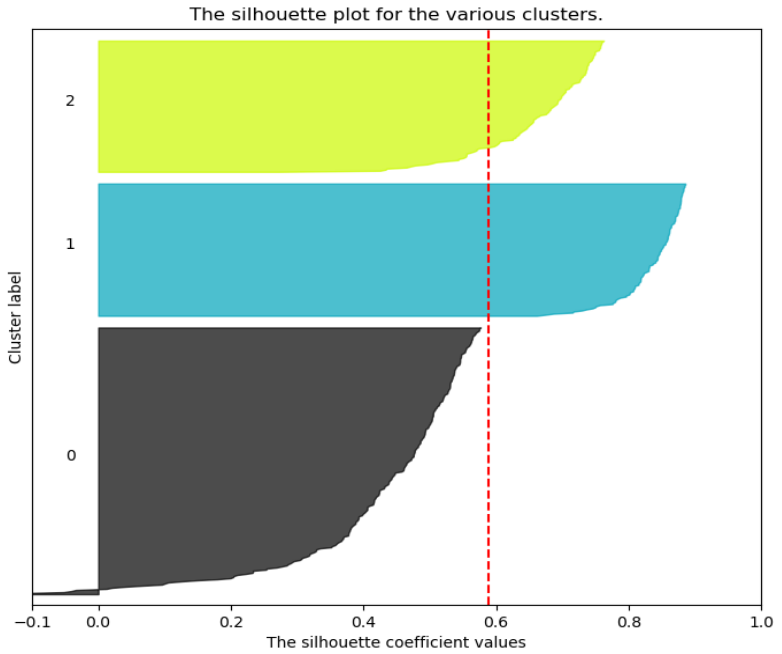


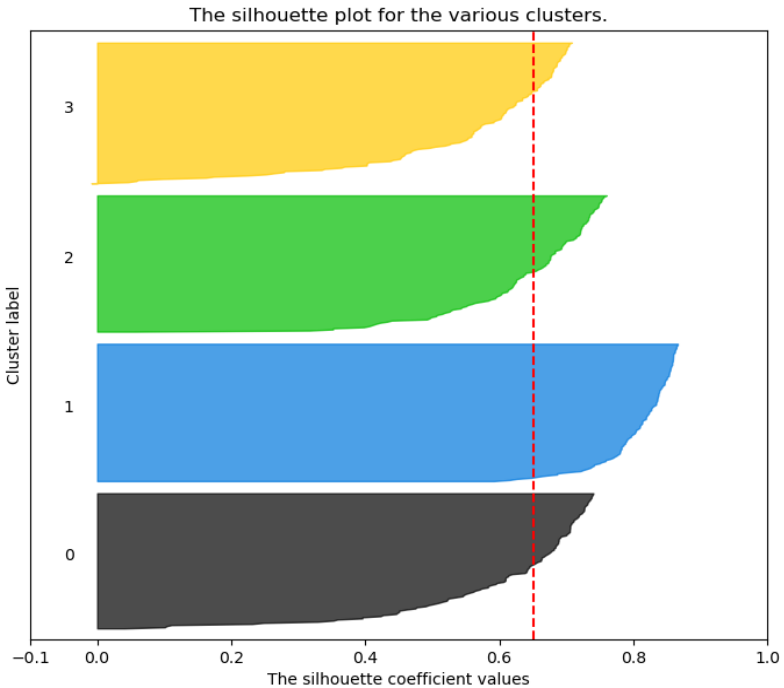
Sihouette Analysis

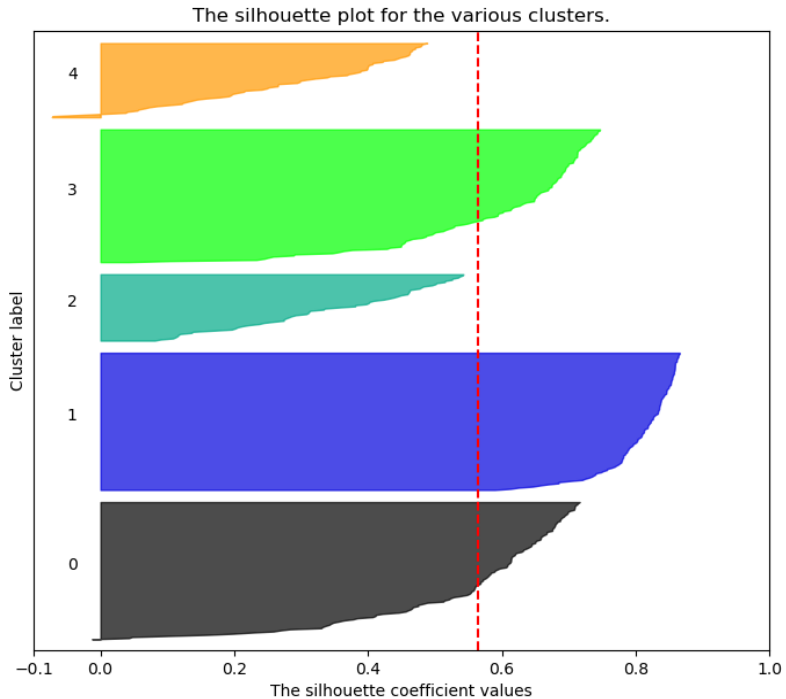
- The silhouette score: là thước đo mức độ giống nhau trung bình của các đối tượng trong một cụm và khoảng cách của chúng với các đối tượng khác trong các cụm khác

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$


The silhouette plot for the various clusters.
Cluster label
1
0
−0.1
0.0
0.2
0.4
0.6
0.8
1.0
The silhouette coefficient values


The silhouette plot for the various clusters.
Cluster label
2
1
0
−0.1
0.0
0.2
0.4
0.6
0.8
1.0
The silhouette coefficient values


The silhouette plot for the various clusters.
Cluster label
3
2
1
0
−0.1
0.0
0.2
0.4
0.6
0.8
1.0
The silhouette coefficient values


The silhouette plot for the various clusters.
Cluster label
4
3
2
1
0
−0.1
0.0
0.2
0.4
0.6
0.8
1.0
The silhouette coefficient values