BỘ GIÁO DỤC VÀ ĐÀO TẠO

**TRƯỜNG ĐẠI HỌC DUY TÂN**

**KHOA CÔNG NGHỆ THÔNG TIN**

🙡🕮🙣



**BÁO CÁO**

**ĐỀ TÀI NGHIÊN CỨU KHOA HỌC**

**Đề tài:**

**NGHIÊN CỨU TỐI ƯU K TRONG THUẬT TOÁN K\_MEANS BẰNG PHƯƠNG PHÁP ELBOW, SILHOUETTE VÀ GAP STATISTIC**

**Giáo viên hướng dẫn: Th.S Nguyễn Dũng**

**Sinh viên thực hiện : Lê Văn Hiếu**

**Mã số sinh viên : 24211906618**

**Đà Nẵng, 2021**

**MỤC LỤC**

[DANH MỤC CÁC KÍ HIỆU 3](#_Toc67301797)

[DANH MỤC CÁC HÌNH MÌNH HỌA 4](#_Toc67301798)

[DANH MỤC CÁC BẢNG 5](#_Toc67301799)

[Chương 1. TỔNG QUAN ĐỀ TÀI VÀ CƠ SỞ LÝ THUYẾT 6](#_Toc67301800)

[1.1. Lý do chọn đề tài: 6](#_Toc67301801)

[1.2. Mục tiêu và phương pháp nghiên cứu 6](#_Toc67301802)

[1.3. Bố cục báo cáo 7](#_Toc67301803)

[1.4. Giới thiệu lý thuyết về Machine Learning 7](#_Toc67301804)

[1.5. Giới thiệu về ngôn ngữ lập trình. 9](#_Toc67301805)

[1.6. Một số vấn đề về dữ liệu 11](#_Toc67301806)

[Chương 2. XÂY DỰNG CÁC PHƯƠNG PHÁP TÌM K TỐI ƯU CHO THUẬT TOÁN KMEAN. 14](#_Toc67301807)

[2.1. Thuật toán phân cụm K-means 14](#_Toc67301808)

[2.2. Phương pháp chọn tâm khởi tạo: Kmeans++ 17](#_Toc67301809)

[2.3. Vấn đề chọn số lượng tâm với Kmeans. 19](#_Toc67301810)

[2.4. Phương pháp Elbow 21](#_Toc67301811)

[2.5. Phương pháp Silhouette 22](#_Toc67301812)

[2.6. Phương pháp Gap Statistic 26](#_Toc67301813)

[Chương 3. Xây dựng Lớp Kmean và Kiểm Thử 30](#_Toc67301814)

[3.1. Xây dựng Lớp Kmean. 30](#_Toc67301815)

[3.2. Thư viện sklearn và hàm make\_blops. 32](#_Toc67301816)

[3.3. Thực thi phân cụm bằng Kmean. 34](#_Toc67301817)

[3.4. Lựa chọn K trên phương pháp Elbow. 37](#_Toc67301818)

[3.5. Lựa chọn K trên phương pháp Silhouette. 38](#_Toc67301819)

[3.6. Lựa chọn K trên phương pháp Gap Statistic 41](#_Toc67301820)

[KẾT LUẬN 43](#_Toc67301821)

[TÀI LIỆU THAM KHẢO 47](#_Toc67301822)

DANH MỤC CÁC KÍ HIỆU

|  |  |
| --- | --- |
| **Kí hiệu** | **Ý nghĩa** |
| Kmeans | Thuật toán phân cụm dữ liệu Kmeans. |
| K/k | Số lượng tâm/cụm trong thuật toán phân cụm Kmeans |
| C, c | Lần lượt là tập ma trận các tâm của Kmeans và vector tâm thuộc **C** |
| X, x | Lần lượt là tập ma trận các điểm của tập dữ liệu đưa vào và vector điểm thuộc **X** |
| D(a,b) | Khoảng cách giữa ma trận hoặc vector a tới ma trân hoặc vector b. |
| d | Dimension. Số chiều hay cột của mảng/Tập dữ liệu |
| N | Numer. Số lượng phần tử của mảng/Tập dữ liệu |
| L | Labels. Tập ma trận các nhãn sinh từ Kmeans. |
| Min | Giá trị nhỏ nhất thuộc một tập nào đó. |
| Max | Giá trị lớn nhất thuộc một tập nào đó. |
| TWSS | Total Within Sum of Square, Giá trị dùng để đánh xem độ tượng đồng của một cụm như thế nào. |

DANH MỤC CÁC HÌNH MÌNH HỌA

[Hình 1. Bài toán phân cụm với số K=3 14](#_Toc66697631)

[Hình 2. Mô hình thuật toán Kmeans 17](#_Toc66697632)

[Hình 3. Đầu ra của bài toán phân cụm với mỗi sự lựa chọn tâm khác nhau 18](#_Toc66697633)

[Hình 4. Mô hình thuật toán Kmeans++ 20](#_Toc66697634)

[Hình 5. Ví dụ một mô hình Elbow method với tâm thích hợp là 5 23](#_Toc66697635)

[Hình 6. Biêu đồ biểu thị các giá trị Silhouette. 26](#_Toc66697636)

[Hình 7. Biểu đồ phương pháp Kmean 29](#_Toc66697637)

[Hình 8. Lớp Kmean 31](#_Toc66697638)

[Hình 9. Nhóm Kmean 32](#_Toc66697639)

[Hình 10. Nhóm Elbow 32](#_Toc66697640)

[Hình 11. Nhóm Silhouette 33](#_Toc66697641)

[Hình 12. Nhóm Gap Statistic 33](#_Toc66697642)

[Hình 13. Tạo dữ liệu 35](#_Toc66697643)

[Hình 14. Kmean init='rd' – 1 36](#_Toc66697644)

[Hình 15. Kmean init='rd' – 2 36](#_Toc66697645)

[Hình 16. Kmeans init = 'Kmeans++' – 1 37](#_Toc66697646)

[Hình 17. Kmeans init = 'Kmeans++' – 2 38](#_Toc66697647)

[Hình 18. Mô hình Silhouette K=2 40](#_Toc66697648)

[Hình 19. Mô hình Silhouette K=3 40](#_Toc66697649)

[Hình 20. Mô hình Silhouette K=4 41](#_Toc66697650)

[Hình 21. Mô hình Silhouette K=5 41](#_Toc66697651)

[Hình 22. Mô hình Silhouette K=6 42](#_Toc66697652)

[Hình 23. Mô hình Silhouette K=7 42](#_Toc66697653)

[Hình 24. Mô hình Gap Statistic 43](#_Toc66697654)

DANH MỤC CÁC BẢNG

[Bảng 1. Kết quả kiểm thử 45](#_Toc67293713)

1. TỔNG QUAN ĐỀ TÀI VÀ CƠ SỞ LÝ THUYẾT
   1. Lý do chọn đề tài:

Kmeans là thuật toán phân cụm thuộc nhóm học không giám sát (Unsupervised Learning). Cách hình thành các cụm dựa trên việc tính khoảng cách các tâm và các điểm (Thường dùng khoảng cách Euclid). Sau đó điều chỉnh các tâm sao cho các tâm không còn thay đổi hoặc sự khác nhau không quá lớn. Sau khi tìm được các tâm thì ta có thể biết được các cụm.

Thuật toán Kmeans có thể giúp ích dược rất có thể rất hữu ích cho nhiều lĩnh vực khác nhau trong cuộc sống. Các lĩnh vực đó có thể là toán học, kinh tế, văn bản, quy hoạch đô thị, cứu nạn cứu trợ,… Một vài ứng dụng trong số đó có thể kể ra như là phân khúc khách hàng, nhận diện chữ viết, phân nhóm khu vực dựa vào phân bố dân cư, tách vật thể trong ảnh,…

Tuy vậy để Kmeans hoạt động được thì phải cần biết trước số cụm đưa vào(K). Đây là vấn đề thường gặp trong bài toán phân cụm nói chung. Với các bài toán đơn giản thì ta có thể nhận diện được số K và có thể đưa vào thuật toán ngay. Nhưng đối với các bài toàn phức tạp hơn thì số K này là một vấn đề lớn. Việc phân cụm mà số cụm chưa biết rõ thì ta cần phải tính toán để có được số cụm K tối ưu.

Để giải quyết vấn đề này, nhóm sẽ phân tích và xây dựng các phương pháp để tìm được số K tối ưu và áp dùng vào một số bài toán cụ thể. Các phương pháp: elbow method, silthoutte method sẽ giúp tìm được số tâm tối ưu. Tăng hiệu quả bài toán.

* 1. Mục tiêu và phương pháp nghiên cứu
     1. Mục tiêu nghiên cứu

Mục tiêu của đề tài này là nghiên cứu tối ưu thuật toán Kmean và tìm các phương pháp hiệu quả đề tìm được số K tối ưu. Khi đã tìm được thì nghiên cứu, áp dụng phương pháp đó vào thuật toán Kmeans. Chạy thử thuật toán Kmean và tìm số K tối ưu dựa trên các phương pháp đó. Xuất ra mô hình và đưa ra các nhận xét về các mô hình đó.

Hiểu thêm về các thuật toán máy học, cách sử dụng ngôn ngữ python và các công cụ kèm theo.

* + 1. Phương pháp nghiên cứu.

Nghiên cứu về ý nghĩa thuật toán, công thức toán học của thuật toán phân cụm Kmeans và thực hiện trên ngôn ngữ lập trình python.

Nghiên cứu triển khai các thuật toán về tìm K tối ưu cho thuật toán phân cụm Kmeans. Áp dụng các công cụ, thư viện trong ngôn ngữ lập trình python để thực hiện, mô tả các phương pháp nói trên bằng biểu đồ.

Thử nghiệm trên một số mẫu dữ liệu có sẵn. Lấy giá trị K có được nhờ các phương pháp trên. Áp dụng phân cụm Kmeans và đánh giá kết quả.

* 1. Bố cục báo cáo

Nội dụng báo cáo bao gồm các chương sau:

* Chương I. Tổng quan đề tài và Cơ sở lý thuyết: Giới thiệu tổng quan, mục tiêu và phương pháp nghiên cứu đề tài. Giới thiệu về machine learning và ngôn ngữ lập trình cùng với các công cụ hỗ trợ.
* Chương III. Thuật toán phân cụm Kmeans. Giới thiệu tổng quan về thuật toán phân cụm Kmeans. Các công thức toán học đằng sau Kmeans và thực thi thuật toán. Đồng bàn về vấn đề khởi tạo tâm ban đầu K-means++ và về chọn số k tâm ban đầu.
* Chương IV. Xây dựng phương pháp tìm K tối ưu: Giới thiệu và xây dựng thuật toán về hai phương pháp tìm K tối ưu là phương pháp Elbow, Silhouette và Gap Statistic.
* Chương V. Xây dựng lớp Kmean và Kiểm thử: xây dựng lớp Kmean chạy trên python. Sử dụng hàm Kmeans và các phương pháp tìm K tối ưu vào tập dữ liệu thử.
* Kết luận: Đưa ra các đánh giá về các phương pháp. Nêu ra những kết quả đạt được và hạn chế của đề tài.
  1. Giới thiệu lý thuyết về Machine Learning

Cuộc cách mạng công nghiệp 4.0 với bằng chứng là sự nổi lên của AI - Artificial Intelligence (Trí tuệ nhân tạo), cụ thể hơn là Machine learning(Học máy) đã làm thay đổi thế giới(1.0 – Động cơ hơi nước, 2.0 – Điện năng, 3.0 – Công nghệ thông tin). Ta có thể thấy rằng trí tuệ nhân tạo đang dần len lõi vào mọi mặt của đời sống của chúng ta. Một số ví dụ về AI trong cuộc sống mà mọi người có thể dễ dàng nhận thấy như là: Hệ thống đề xuất video trên theo sở thích người dùng trên Youtube, xe tự hành của Tesla hay Camera AI trên các mẫu điện thoại thông minh đời mới. Và đặc biệt là trị tuệ nhận tạo AlphaGo được tạo ra bởi Google DeepMind tại London, trong một trận thi đấu cờ vây, đã có thể chiến thắng trước Lee Sedol – Người từng 18 lần vô địch thế giới trong bộ môn này. Hay gần đây hơn là OpenAI có khả năng chơi game Dota2 chiến thắng trước những tuyển thủ chuyên nghiệp hàng đầu thế giới. Từ những vị dụ trên đã cho thấy được sự phát triển không ngừng của AI trên mọi lĩnh vực trong đời sống.

Machine learning hay học máy(máy học) là một lĩnh vực con thuộc Trị tuệ nhận tạo(AI) nhằm nghiên cứu và sử dụng các thuật toán giúp máy tính hoặc các hệ thống có thể “học” từ dữ liệu nhằm giải quyết những vấn đề cụ thể. Các thuật toán máy học cần được đưa vào một lượng dữ liệu cần thiệt để cho ra một mô hình chính xác. Các tập dữ liệu này sẽ được thu gom, làm sạch và sử dụng một thuật toán máy học để đưa ra một mô hình mô tả bộ dữ liệu này. Do tập dữ liệu đầu vào là khác nhau nên các thuật toán máy học sẽ có những phương pháp và kỹ thuật “học” khác nhau. Các thuật toán máy học có thể được chia làm thành 3 nhóm:

* Học có giám sát(Supervised Learning): là một phương pháp học mà ở đó các tập dữ liệu đầu vào đã được gán nhãn và nhiệm vụ là tạo ra một mô hình để mô tả hoặc một dự đoán dựa trên tập dự liệu tương tự khác. Học có giám sát có thế được phân loại thành hai bài toán nhỏ hơn đó là Phân loại(Classification) và Hồi quy(Regession). Bài toán Phân loài là bài toán mà các nhãn của bộ dữ liệu đầu vào thuộc vào một số nhóm hữu hạn. Bài toán Hồi quy là bài toán mà các nhãn của bộ giữ liệu đầu vào là các giá trị cụ thể nào đó.
* Học không có giám sát(Unsupervised Learning): Là một phương pháp học mà ở đó các tập dữ liệu đầu vào khi đó chưa hoặc không được gán nhãn. Phương pháp này sẽ dựa vào tập dữ liệu đó để có thể thực hiện một công việc nào đó như phân cụm dữ liệu hoặc tìm ra các quy luật hay nguyên tắc cho bộ dữ liệu đó. Do đó sẽ có hai bài toán con trong Học không có giám sát là Phân cụm dữ liệu(Clustering) và tìm Quy tắc kết hợp(association rules). Thuật toán Phân cụm dữ liệu giúp phân nhóm các điễm dữ liệu dữa trên các tính chất đặc điểm tương tự nhau. Quy tắc kết hợp là dựa vào tập dữ liệu đã cho để khai phá, tìm ra những quy luật hoặc quy tắc nào đó có ích trong một công việc cụ thể.
* Học tăng cường(Reinforcement Learning): là một phương pháp học tiên tiến nhằm tự xác định các hành động tiếp theo sao cho đạt được hiệu quả hay lợi ích cao nhất. Bài toán này có thể dễ dàng thấy được trong hệ thống xe tự hành.
  1. Giới thiệu về ngôn ngữ lập trình.
     1. Python và phần mềm Jupyter lab

Python là một ngôn ngữ lập trình bậc cao được sử dụng cho nhiều mục đích khác nhau do Guido Van Rossum tạo ra và ra mắt lần đầu tiên vào năm 1991. Đặc điểm của Python là một ngôn ngữ mạnh mẽ, dễ học, dễ sử dụng. Nó chứa các cấu trúc câu lệnh với cú pháp tối giản, đơn giản và rõ ràng giúp việc học tập và trao đổi trong lập trình trở nên dễ dàng.

Python là cộng cụ tuyệt vời cho việc lập trình các thuật toán máy học. Ngôn ngữ này cho phép việc lập trình trở nên ngắn ngọn và dễ hiểu. Đồng thời, chứa đựng nhiều các hàm và thư viện được xây dựng và tích hợp sẵn trong nó. Trong khi đó, máy học và trí tuệ nhân tạo có nhiều các thuật toán và các công thức toán học phức tạp cùng với lượng dữ liệu lớn để có thể hoạt động. Sự đơn giản của Python giúp lập trình viên tập trung vào các vấn đề của máy học thay vì là các kĩ thuật sử dụng ngôn ngữ lập trình.

Python chứa nhiều các thư viện hỗ trợ việc tính toán và nghiên cứu máy học. Việc thực thi các thuật toán máy học và sử dụng dữ liệu lớn có thể sẽ phức tạm và tốn nhiều thời gian. Trong khi đó, Python cung cấp môi trường làm việc, thư viện hỗ trợ đã được lập trình và có hiệu năng mạnh mẽ. Việc này sẽ giúp tiết kiệm được nhiều vấn đề khác nhau và giúp quá trình lập trình diễn ra được tốt nhất.

Project Jupyter là một dự án nhằm để phát triển phần mềm mã nguồn mở, các tiêu chuẩn mở và các dịch vụ cho tính toán tương tác trên nhiều ngôn ngữ lập trình khác nhau. Jupyter Notebook là một ứng dụng trên web mã nguồn mở cho phép tạo và chia sẽ các tài liệu có chưa mã lập trình trực tiếp, các phương trình, hình ảnh và văn bản tường thuật. Jupyter lab là một môi trường phát triển có tính tương tác cao trên web dành cho Jupyter Notebook, lập trình và làm việc với dữ liệu. JupyterLab có tính linh rất hoạt cao, giao điện người dùng đã được tinh chỉnh và sắp xếp lại để hỗ trợ cho nhiều quy trình làm việc khác nhau trong khoa học dữ liệu, khoa học máy tính và máy học. JupyterLab có thể được mở rộng và tích hợp nhiều các tính năng phụ trợ khác nhau tuy theo nhu cầu của người dùng.

Anaconda là một phần mềm miễn phí, giúp dễ dàng thực hiện các tạc vụ như tải về và quản lý các gói tài nguyên, môi trường làm việc và các phiên bản Python. Anaconda cung cấp JupyterLab miễn phí và người dùng có thể dễ dàng tải về những thư viện mà mình cần.

Đề tài này sẽ sử dụng Anaconda JupyterLab với Python phiên bản 3.8.5 cùng với các thư viện dưới đây để chạy hầu hết các phần lập trình.

* + 1. Thư viện numpy

Numpy là một thư viện mã nguồn mở thuộc Python dùng để làm việc với các mảng và ma trận lớn, nhiều chiều, cùng với số lượng lớn các hàm toán học cấp cao để hoạt động trên các mảng này. Nó hỗ trợ nhiều hàm phục vụ cho việc thực hiện các phép biến đổi, tinh toán trong đại số tuyến tính và ma trận. Numpy được xây dựng bởi Tavis Oliphant vào năm 2005. NumPy là viết tắt của Numerical Python.

Một đặc điểm của Numpy là các mảng Numpy được lưu trữ liên tục lại một nơi trong bộ nhớ, làm cho tốc độ truy cập và xữ lý trở nên hiệu quả. Tốc độ xữ lý này nhanh hơn nhiều so với kiểu dữ liệu List được viết sẵn của python do kiểu dữ liệu List lưu trữ dữ liệu rời rạc trên bộ nhớ. Việc này rất hữu ích cho tiệc tính toán các phép tính có khối lượng dữ liệu lớn và nhiều chiều.

* + 1. Thư viện pandas

Pandas là một thư viện mã nguồn mở thuộc Python. Đây là một cộng cụ mạnh mẽ, nhanh nhẹn, linh hoạt và dễ dàng sử dụng để thực hiện các công việc phân tích và xữ lý dữ liệu. Pandas hỗ trợ nhiều hàm khác nhau giúp phân tích, làm sạch và các thao tác khác với dữ liệu. Pandas được tạo ra bởi Wes McKinney vào năm 2008. Pandas là công cụ tuyệt vời trong ngành khoa học dữ liệu nói chung và máy học nói riêng.

Pandas là một cộng cụ tuyệt vời cho tiền xữ lý dữ liệu. Việc đọc file dữ liệu lớn với python trở nên rất dễ dàng với hàm pandas.read\_csv(). Đây là một ví dụ cho thấy được sự tối ưu của thư viện pandas và Python cho sự tiện lời và mạnh mẽ của nó. Đồng thời nó còn hỗ trợ các hàm xữ lý dữ liệu khác như pandas.drop(): bỏ một cột hay dòng nào đó; pandas.isnull(): Kiểm tra xem trong dữ liệu có phần từ nào bị trống không,…

* + 1. Thư viện matplotlib

Matplotlib là một thư viện mã nguồn mở thuộc Python giúp hỗ trợ vẽ đồ thị. Đây là một cộng cụ rất hữu ích khi làm việc với Python và numpy. Module được sử dụng nhiều nhất của Matplotib là Pyplot cung cấp giao diện như MATLAB. Thư viện này giúp biểu đồ hóa các công thức toán học và các dạng dữ liệu liên tục một cách dễ dàng và thuận tiện.

* 1. Một số vấn đề về dữ liệu
     1. Tập dữ liệu kiểm thử

Vì dữ liệu rất quan trọng trong việc đánh giá xem việc một thuật toán máy học có chính xác hay không. Để đảm bảo hoạt động chính xác nhất thì dữ liệu cần phải được thu gom lại và làm sách, đồng thời tập dữ liệu đó phải thích hợp với thuật toán mà lập trình viên đang triển khai.

Trong bài báo cáo này sẽ sử dụng dữ liệu sinh ra từ thư viện. Do các điểm dữ liệu trong tập dữ liệu được sinh ra bằng thư viện có sự nhất quán rất cao, dễ điều chình và chính xác, cho nên kết qua cho ra thường đạt kì vọng. Đồng thời số liệu của mỗi điễm dữ liệu là hợp lý để cho tốc độ tính toán cao. Thông qua các thông số cụ thể thì ta có thể điều chỉnh đườc các thông số của bộ dữ liệu thử như: số cụm, phương sai, số lượng dữ liệu,…. Cho nên việc đánh giá sẽ trở nên đúng đắn và dễ dàng hơn.

Tuy nhiên, số lượng điễm dữ liệu cho mỗi cụm tạo ra bằng cách trên là tương tự nhau. Cho nên khi xuất lên biều đồ thì sẽ cân bằng hơn và dễ hiệu hơn. Cho nên, khi sử dụng các tập dữ liệu khác với số lượng điểm mỗi cụm không cân bằng thì khi xuất lên biều đồ có thể sẽ khác những nhìn chung thì kết qua vẫn sẽ có thể chính xác.

* + 1. Tối ưu hóa dữ liệu

Đối với thuật toán Kmeans thì bộ dữ liệu đầu vào thì kiểu dữ liệu phải là các số thuộc tập số thực hoặc dữ liệu ở dạng liên tục. Các số này không được quá lớn vì các số này sẽ khiên máy tính toán nặng hơn khi áp dụng các công thức toán học, gây lãng phí tài nguyên máy tính. Trong bộ dữ liệu không có các điểm dữ liệu trống sai định dạng.

Khi áp dụng những bộ dữ liệu thức tế thì dữ liệu cần phải được làm sạch. Bộ dữ liệu đầu vào có thể sẽ có những cột mà ở đó các dữ liệu có thể là các giá liên tục hoặc rời rác, dữ liệu mang số quá lớn hay sai định dạng. Khi đó để thuật toán có thể sử dụng được thì dữ liệu cần phải được làm sạch. Sau đây là một số phương pháp để làm sạch dữ liệu.

Đối với dữ liệu bị sai định dạng: Ta cần chuyên định sang sai đó thành định dạng đúng hoặc xóa điểm dữ liệu đó.

* Ví dụ: Môt điểm dữ liệu có dạng x=(2,’415’) thì ‘415’ không phải là một số nguyên dương mà là kiểu dữ liệu dạng chuỗi(string). Khi đó ta phải chuyển về thành x =(2,415).

Đối với dữ liệu dạng rời rạc: Ta cần chuyển định dạng dữ liệu đó về dạng liên tục.

* Vị dụ: Nhãn của cột giới tính có hai giá trị là ‘Nam’ và ‘Nữ’. Với kiểu dữ liệu này thì thuật toán khó có thể hiểu được. Khi đó ta có thể sử dụng phương pháp LabelEncoder để chuyên các nhãn này thành các kí tự số. Theo đó ta có thể chuyển ‘Nam’🡪1 và ‘Nữ’ 🡪 0.

Đối với dữ liệu là các số thực những các giá trị của nó là quá lớn hay một cột dữ liệu có khoảng giá trị quá lớn 1-10000 và cột khác có khoảng giá trị quá nhỏ 0-10. Lúc này ta cần phải chuẩn hóa dữ liệu. Có 2 phương pháp chính để tối chuẩn hóa dữ liệu là:

* Chuyển khoảng giá trị: là phương pháp rất đơn giản để đưa các giá trị của các cột về cùng một khoảng. Để có thể đưa điểm dữ liệu thứ thuộc cột dữ liệu **.** Ta có thể sử dụng công thức sau:

Với và lần lượt là giá trị thứ của cột dữ liệu trước và sau khi được chuẩn hóa. và lần lượt là giá trị lớn nhất và giá trị thấp nhất của cột dữ liệu **.**

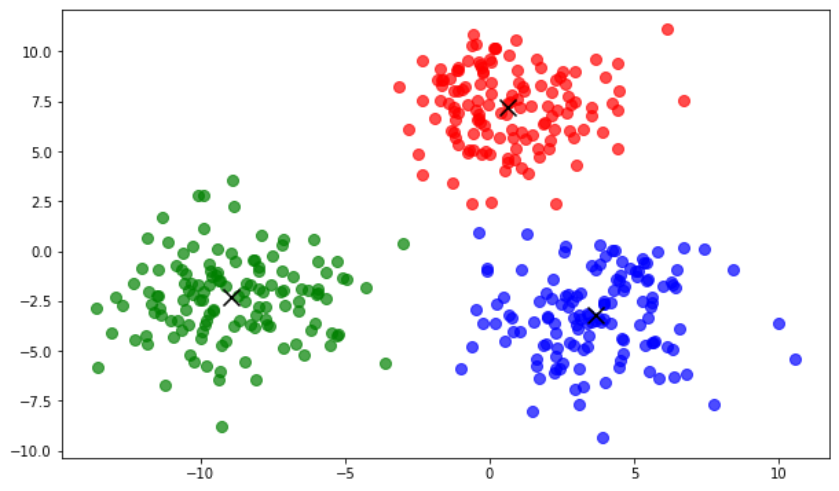
Code thực thi:

1. def normalize(X):
2. n,d=X.shape
3. temp=np.zeros((n,d))
4. for i in range(d):
5. temp[:,i]=(X[:,i]-np.min(X[:,i]))/(np.max(X[:,i])-np.min(X[:,i]))
6. return temp

* Chuẩn hóa theo phân phối chuẩn:
* Code thực thi:

1. def normalize\_std(X):
2. n,d = X.shape
3. temp = np.zeros(shape=(n,d))
4. for i in range(d):
5. mean = np.mean(X[:,i])
6. std = np.std(X[:,i])
7. temp[:,i] = (X[:,i]-mean)/std
8. return temp
9. XÂY DỰNG CÁC PHƯƠNG PHÁP TÌM K TỐI ƯU CHO THUẬT TOÁN KMEAN.
   1. Thuật toán phân cụm K-means
      1. Giới thiệu thuật toán phân cụm Kmeans

Phân cụm là nhiệm vụ phân chia các điểm dữ liệu thành một số nhóm sao cho các điểm dữ liệu trong cùng một nhóm giống với các điểm dữ liệu khác trong cùng một nhóm hơn các điểm dữ liệu trong các nhóm khác. Nói cách đơn giản, mục đích là để tách các nhóm có những đặc và thuộc tính tương tự nhau và gán chúng thành các cụm. Mục tiêu của thuật toán Kmean là phân cụm và gãn nhãn các điểm thuộc tập dữ liệu, với số lượng nhóm được đại diện bởi biến K. Thuật toán hoạt động lặp đi lặp lại để gán mỗi điểm dữ liệu cho một trong k nhóm dựa trên các thuộc tính được cung cấp. Trong hình ảnh tham khảo bên dưới, k = 3, và có ba cụm được xác định từ tập dữ liệu nguồn.



Hình 1. Bài toán phân cụm với số K=3

* + 1. Phân tích toán học:

Cho tập dữ liệu với d là số chiều của vector , Kmean mục tiêu là chia N số lượng phần từ trong thành KN cụm và các tâm cụm sao cho giá trị hàm mất mát là thấp nhất. Hàm mất mát của Kmean được tính như sau:

Xét mỗi điểm dữ liệu được phân vào cụm thì sẽ có sai số là:

|  |  |  |
| --- | --- | --- |
|  |  | (2.1) |

Khi đó ta sẽ tính sai số cho tất cả các điểm trong mỗi cụm. Tương tự với số cụm K= 1 có thì tổng sai số của cụm này đươc tính như sau:

|  |  |  |
| --- | --- | --- |
|  |  | (2.2) |

Tương tự khi có K cụm và dữ liệu được đã được gán nhãn và phân vào mỗi cụm ta có:

|  |  |  |
| --- | --- | --- |
|  |  | (2.3) |

Vậy khi đó ta sẽ có được ma trận nhãn và tâm cụm

|  |  |  |
| --- | --- | --- |
|  |  | (2.4) |

* + 1. Tính khoảng cách Euclid:

Cho a, b là 2 điểm dữ liệu, khoảng cách Euclid từ điểm a đến điểm b:

|  |  |  |
| --- | --- | --- |
|  |  | (2.5) |

Với a, b có số chiều lớn hơn 1 ta có:

Với số chiều là 2:

|  |  |  |
| --- | --- | --- |
|  |  | (2.6) |

Với số chiều là n:

|  |  |  |
| --- | --- | --- |
|  |  | (2.7) |

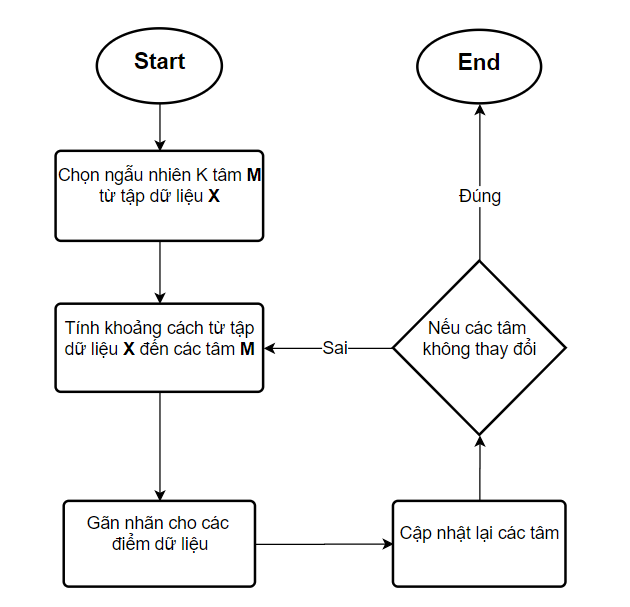
* + 1. Giải thích thuật toán

Thuật toán Kmeans có để được thực hiện như sau.

* Đầu vào: Ma trận dữ liệu và số lượng tâm KN cần tìm
* Đầu ra: Ma trận tâm cụm , và ma trận nhãn .

1. Khởi tạo K điểm bất kì làm các tâm khởi tạo ban đầu của **M**.
2. Tính khoảng cách Euclid từ các điểm đến tâm **M**.
3. Gán nhãn cho mỗi điểm dữ liệu vào cụm có tâm gần nó nhất.
4. Cập nhật các tâm cụm bằng cách lấy trung bình cộng các điểm dữ liệu đã được gán nhãn.
5. Nếu các tâm không thay đổi so với lần cập nhập trước thì kết thúc thuật toán.
6. Quay lại bước 2.

Mô hình đơn giản hóa thuật toán:



Hình 2. Mô hình thuật toán Kmeans

* + 1. Nhược điểm của Kmean:

Kmean không phải là một thuật toán máy học hoàn hảo để phân cụm dữ liệu. Dễ thấy rằng Kmean tính khoảng cách Euclid để tìm ra nhãn của tập dữ liệu và các tâm cụm. Đồng thời dựa vào thuật toán trên thì ta có thấy được một số nhược điểm đối với Kmean như sau:

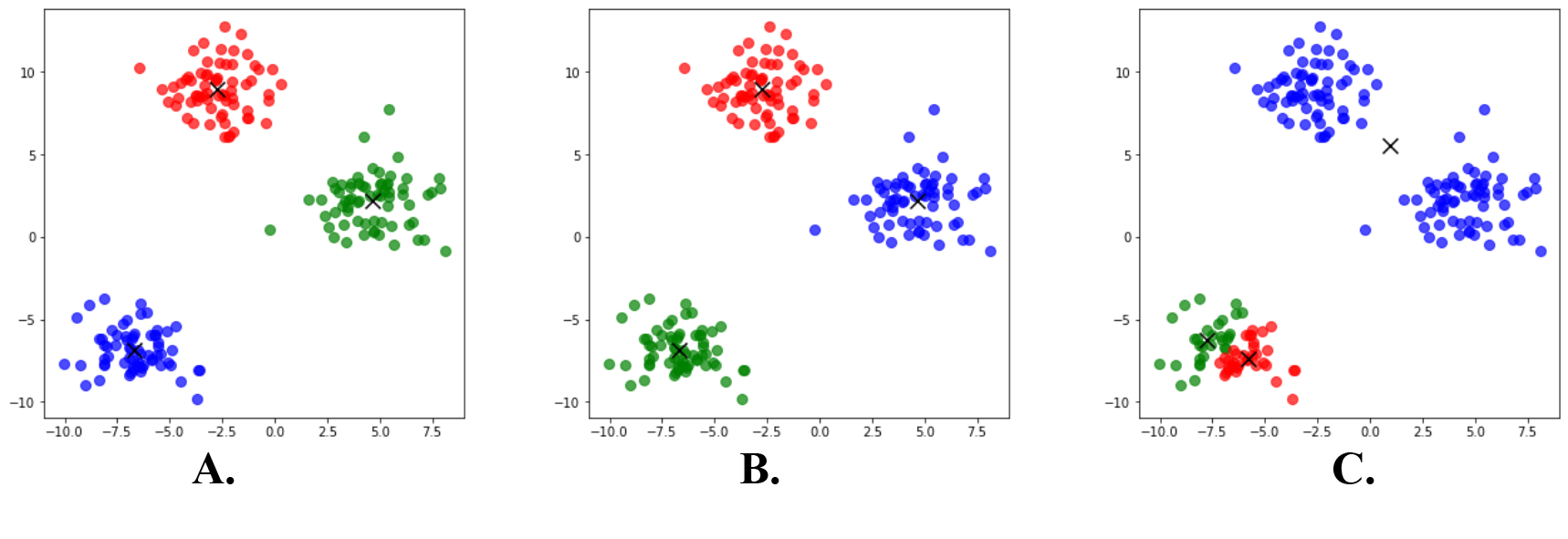
Thứ nhất là vấn đề chọn số lượng tâm cụm(K). Phân này sẽ được giải thích cụ thể hơn ở phần 3.3 và đây chính là vấn đề chính trong bài báo cáo này.

Thứ hai là thuật toán Kmean không thế xữ lý được được những dữ liệu bị nhiễu và các ngoại lệ. Đây chính là vấn đề chính khiến độ chính xác của thuật toán này thấp. Do kết quả tính khoảng cách là chính xác tuyết đối và Kmean cần tính cho toàn bộ các điểm nên rất khó để có thể xử lý các điểm này. Việc dữ liệu có nhiễu và các ngoại lệ là không thể tránh khỏi trọng việc thu thập dữ liệu trong thực tế. Cho nên muốn cho thuật toán này mang lại độ chính xác cao thì dữ liệu, một là được thu thập cẩn thận sao cho không hoặc ít xuất hiện các điểm nhiễu hay ngoại lệ. Hai là sử dụng các phương pháp đặc biệt để có thể loại bỏ nó.

Cuối cùng là Kmean chỉ có thể phân cụm được những cụm dữ liệu dạng lồi. Vì Kmeans sử dụng phương pháp tính khoảng các giữa các điễm và lấy những điễm gần nhất với các tâm. Cho nên, hình dạng của càng cụm thường có dạng lồi hay các cụm có dạng tròn gần nhau. Cho nên nếu hình dạng của cụm khác đi hay tuân theo một hình dạng khác thì Kmean khó có thể xữ lý được. Khi đó ta có thể sử dụng các phương pháp phân cụm khác như DBCAN,…

* 1. Phương pháp chọn tâm khởi tạo: Kmeans++

Trong bài toán Kmeans, ta phải lựa chọn K tâm ban đâu để thuật toán có thể cập nhật tâm và phân cụm. Lựa chọn tâm theo cách chọn ngẫu nhiên K điểm dữ liệu đã cho ban đầu là một cách đơn giản và nhanh chóng. Tuy nhiên phương pháp này nảy sinh ra một số vấn đề có thể gây ảnh hưởng đến kết quả cuối cùng cho bài toán phân cụm. Các tâm được sinh ra khác nhau khiến cho kết quả của bài toán là khác nhau, có thể là vì trị các tâm khác nhau hay việc phân nhãn khác nhau. Trường hợp tệ nhất là các tâm khởi tạo ban đầu quá gần nhau dẫn đến việc phân cụm bị sai.



Hình 3. Đầu ra của bài toán phân cụm với mỗi sự lựa chọn tâm khác nhau

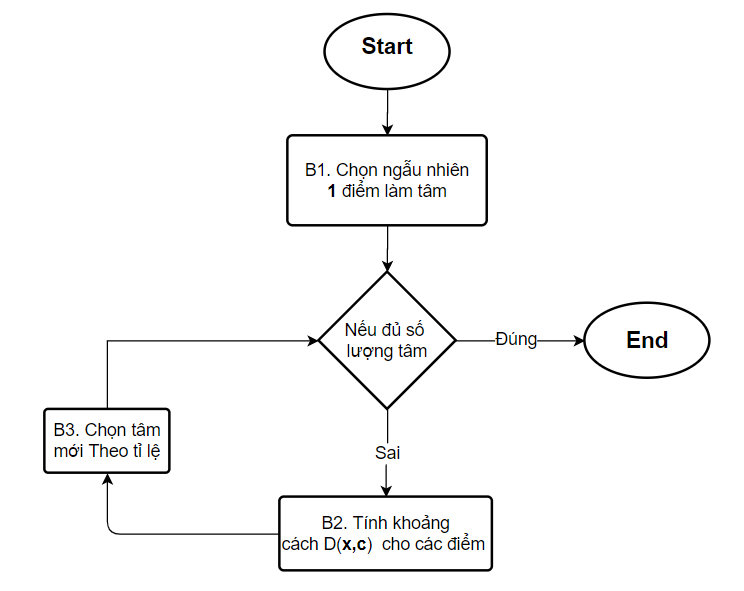
Trong hình trên ta có thể thấy: Hình A. B. là có sự khác nhau về màu trong 2 cụm. Riêng hình C. thì các cụm được phân bố khác đi so với hình A. và B.. Dễ thấy rằng, sự phân cụm như hình C. là có sự sai lệch. Sự lựa chọn tâm ngẫu nhiên quá gần nhau đã dẫn đến điều này.

Đề làm giảm đi sự sai lệch do khởi tao tâm ban đầu, ta có thể khởi tạo các tâm ban đầu sử dụng phương pháp Kmeans++. Đây là phương pháp được David Arthur và Sergei Vassilvitskii viết trong một bài báo nói về ưu điểm của việc gieo hạt cẩn thận(“*k-means++: The Advantages of Careful Seeding*”). Đây là phương pháp giúp việc lựa chọn tâm ban đầu được giàn trải hơn so với phương pháp lựa chọn tâm ngẫu nhiên cho toàn bộ các tâm.

Thuật toán được mô tả như sau:

1. Lựa chọn một tâm ngẫu nhiên là một điểm thuộc tập dữ liệu đầu vào
2. Với mỗi điểm x, xác định D(x,c) sao cho D(x,c) là khoảng cách ngắn nhất từ x đến tâm c gần nhất mà ta đã chọn.
3. Chọn tâm c tiếp theo thuộc X sao cho tỉ lệ với khoảng cách d(x,ci)2 trên tổng khoảng cách đối với tâm gần nhất . Tức là chọn điểm xa nhất từ x đến tâm c gần nhất đã chọn.
4. Lặp lại bước 2, 3 cho đến khi đủ số lượng tâm cần tìm.

Mô hình đơn giản hóa thuật toán Kmeans++:



Hình 4. Mô hình thuật toán Kmeans++

Áp dụng thuật toàn này vào nghiên cứu sẽ giúp hạn chế dược sai số trong quá trình thử nghiệm các thuật toán. Tuy vậy Kmeans++ sẽ có tốc độ chậm hơn so với việc lựa chọn ngẫu nhiên cho toàn bộ các tâm do phải thực hiện tính toán lại khoảng cách các điểm với các tâm. Điều này thậm chí còn tệ hơn nếu số K lớn.

Trong suốt bài báo cáo này, các tâm ban đầu sẽ được khởi tạo ngẫu nhiên bằng thuật toán Kmeans++ để tăng độ chính xác và hiệu suật bài toán mặc dù tốc độ thực thi có thể sẽ chậm hơn.

* 1. Vấn đề chọn số lượng tâm với Kmeans.

Muốn cho thuật toán phân cụm Kmeans phân chia được thì ta cần phải biết được số cụm cần thiết để đưa vào. Câu hỏi đặt ra là: với một tập dữ liệu đã có sẵn thì phân bao nhiêu cụm là hợp lý, tối ưu? Đây là câu hỏi phổ biến nhất trong thuật toán phân cụm Kmeans.

K tối ưu là số lượng cụm K sao cho sự khác biết giữa cụm này và cụm khác là thấp nhất so với tổng thể bài toán.

Đối với một bộ dữ liệu đẹp, tức là tập dữ liệu có số chiều nhỏ và số lượng dữ liệu ít. Đồng thời các cụm đã được phân tách rõ ràng. Thì ta có thể dễ dàng biết được số cụm K cần phân chia nhờ vào cảm quan. Tuy nhiên số k này có thể được chia khác nhau cho một tập dữ liệu. Vì các cụm trong Kmeans được chia sao cho các điểm trong mỗi cụm là giống nhau nhất có thể. Sự giống nhau này càng tăng thì số cụm càng tăng. Cho nên, số lượng cụm tối đa trọng một bài toán có thể bằng số lượng điểm dữ liệu trong tập dữ liệu. Cho nên ta cần phải đi tìm K tối ưu cho tập dữ liệu đó.

Trong khi đó với bộ dữ liệu có số chiều nhiều và lượng dữ liệu lớn thì việc tìm K bằng cảm quan sẽ không còn hiệu quả nữa. Khi đó ta chỉ có thế biết được K bằng hai cách. Thứ nhất là số K này đã được biết trước cho tập dữ liệu đã cho. Khi đó ta chỉ cần áp dụng thuật toán phân cụm để lấy nhãn của từng điểm dữ liệu. Thứ hai là khi ta chưa biết được số tâm K cho sẵn, lúc đó ta phải đi tìm kiếm số K tối ưu. Chương sau sẽ trình bày hai phương pháp để tìm K tối ưu cho thuật toán Kmeans.

Trong thực tế việc chọn số tâm K sẽ còn dựa vào nhiều yếu tố khác. Đối với dữ liệu dạng hai chiều hay ba chiều thì ta có thể dễ hàng hình dung được hình dạng của tâm. Tuy nhiên khi số chiều dữ liệu d từ 4 trở lên thì ta chỉ có thể tự hình dung. Khi đó thì cần thiệt phải có sự phân tích dữ liệu trong tập dữ liệu đã có được. Phân tích ở đầy có thể là phân tích dưa trên mỗi tương quan của các thuộc tính với nhau. Một thuộc tình này có thể tương tác mạnh với thuộc tính kia. Đối với một số trường hợp ta có thể gán thêm trọng số đối với một số thuộc tính quan trọng dẫn đến phân cụm. Ngoài ra ta có thể sử dụng phương pháp như PCA để có thể giảm được chiều dữ liệu, nếu có thể giảm số chiều về 2,3 thì có thể biểu thị ra và phân tích.

Việc chọn K còn dựa trên mục đích của phân cụm. Với mỗi bộ dữ liệu mà số chiều trong nó là lớn thì việc phân tích K có thể gặp nhiều khó khăn hơn. Khi đó ta có thể dựa vào các phương pháp và phân tích dữ liệu để có thể chọn ra những đặc tính nổi bật trong dữ liệu, theo đó có cái hình tông quát hơn về dữ liệu. Theo đó ta có thể chọn được số K phù hợp để phân cụm.

* 1. Phương pháp Elbow

Phương pháp Elbow(Củi chỏ) là một trong những phương pháp nổi tiếng nhất và đơn giản mà ta có thể dùng để chọn giá trị K phù hợp và tăng độ chính xác của mô hình.

Ý tưởng chính của thuật toán Kmean như sau. Với một tập dữ liệu ban đầu, phân tập dữ liệu đó thành K cụm khác nhau. Với mối cụm sẽ có một tâm gọi là Centroid, ta tính tổng khoảng cách từ tâm này đến toàn bộ các điểm dữ liệu thuộc cụm của tâm đó. Tham số này gọi là WSS (Within-cluster Sum of Square). Sau đó ta tính tiếp tục tìm WSS cho các tâm còn lại là lấy tổng của nó. Tổng này gọi là TWSS(Total Within-cluster Sum of Square). Giá trị này có thể được tính như sau:

|  |  |  |
| --- | --- | --- |
|  |  | (2.8) |

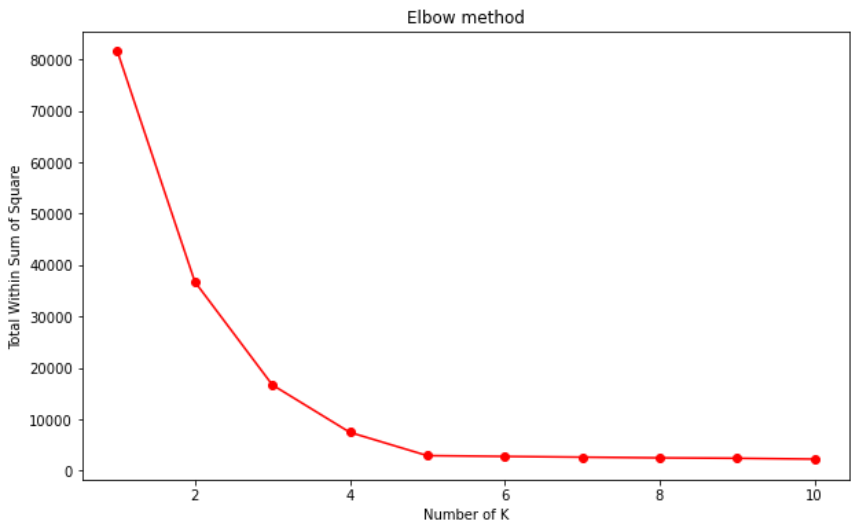
Vì mục tiêu của thuật toán phân cụm là phân chia toàn bộ các điểm của tập liệu ban đầu thành K cụm khác nhau sao cho các điểm dữ liệu thuộc cùng một cụm là giống với nhau nhất có thể, hay TWSS nhỏ nhất có thể, nên việc lựa chọn phương thức tính khoảng cách là rất quan trọng vì khoảng cách được sử dụng như là một thước đo mức độ tương đồng của các quan sát. Tuy vậy với số lượng K càng lớn thì các giá trị TWSS càng nhỏ dần và không có sự thay đổi lớn và dẫn vô nghĩa với số K rất lớn. Do đó ta nên chọn giá trị K cho TWSS tối ưu nhất.

Theo đó ta có quy trình triển khai Elbow method như sau:

* Đầu vào: Ma trận dữ liệu và khoảng số lượng NumK<N cần tìm.
* Đầu ra: TWSS của từng giá trị K trong khoảng NumK

1. Triển khai thuật toán phân cụm với số cụm K trong khoảng NumK bắt đầu từ 1.
2. Với mỗi giá trị K, tính giá trị TWSS tương ứng.
3. Biểu thị biểu đồ Elbow Method theo các giá trị TWSS của K đã tính ở bước 2.
4. Dựa vào biểu đồ trên, lấy điểm K sao cho giá trị TWSS của K đó không thay đổi quá nhiều so với trước.

Ta lựa cọn điểm K trong biểu đồ sao cho giá trị của nó có sự giảm không đáng kể với các điểm tiếp theo nó.



Hình 5. Ví dụ một mô hình Elbow method với tâm thích hợp là 5

Tuy vậy như đã nói ở các chương trước thì phân bố các điểm dữ liệu có thể là không được chuẩn hoàn theo như dữ liệu mô hình ở trên. Cho nên biểu đồ sẽ cho ra kết quả không rõ ràng. Do đó việc chọn tâm sẽ trở nên khó khăn hơn. Khi đó ta sẽ dựa vào cảm quan để có thể lựa chọn k tốt nhất trong mô hình.

* 1. Phương pháp Silhouette

Phương pháp Silhouette(Hình bóng) là một phương pháp dùng để tính toán, đánh giá mức độ thích hợp của kĩ thuật phân cụm. Giá trị của có được gọi là hệ số Silhouette(Silhouette Coefficient) hay điểm Silhouette(Silhouette score). Giá trị này giao động từ trong khoảng [-1;1].

* Các giá trị càng gần 1 chứng tỏ điểm đó đã tách xa và phân biệt với các điểm thuộc cụm khác.
* Giá trị 0 mang ý nghĩa là điểm đó không có sự khác biệt giữa cụm này và cụm kia. Hoặc khoảng cách giữa các điểm trong cụm là không khác biệt
* Giá trị càng về -1 thì có nghĩa là điểm đó đã bị phân cụm sai.

Chung quy lại nếu giá trị Silhouette càng lớn thì cụm đó đã được phận cụm chính xác

Giá trị Silhouette cho mỗi điểm có thể được tính như sau:

Với mỗi điểm dữ liệu (điểm dữ liệu i thuộc cụm ), đặt:

|  |  |  |
| --- | --- | --- |
|  |  | (2.9) |

Với là tổng khoảng cách từ điểm dữ liệu i đến tất cả các điểm dữ liệu j thuộc sau cho i khác j chia cho số điểm dữ liệu thuộc trừ 1(không tính i vào số lượng). Ta có thể hiểu như là cách để đánh giá điểm dữ liệu i có thích hợp với cụm hay không. Giá trị càng nhỏ thì điểm dữ liệu i càng hợp với cụm và ngược lại.

Tiếp theo ta xác định sự khác biệt trung bình của điểm dữ liệu i đến các tâm tính trung bình khoảng cách giữa điểm dữ liệu i đến tất cả các điểm dữ liệu của các cụm không phải là cụm :

|  |  |  |
| --- | --- | --- |
|  |  | (2.10) |

Giá trị nhỏ nhất ở đây tức là ta lấy trung bình tổng khoảng cách đến gần cụm nhất. Muốn làm vậy ta buộc phải tính trung bình khoảng cách của tất cả các cụm và chọn giá trị nhỏ nhất.

Sau đó ta có thể tính giá trị Silhouette bằng công thức sau:

|  |  |  |
| --- | --- | --- |
|  |  | (2.11) |

Và gắn cho nếu số lượng của cụm, có rằng buộc này để ngắn việc số lượng cụm tăng lên một cách đáng kể và tránh gặp lỗi khi số lượng nhỏ bằng 0 khiến máy tính chia cho 0:

|  |  |  |
| --- | --- | --- |
|  |  | (2.12) |

Ta có thể viết lại rằng

|  |  |  |
| --- | --- | --- |
|  |  | (2.12) |

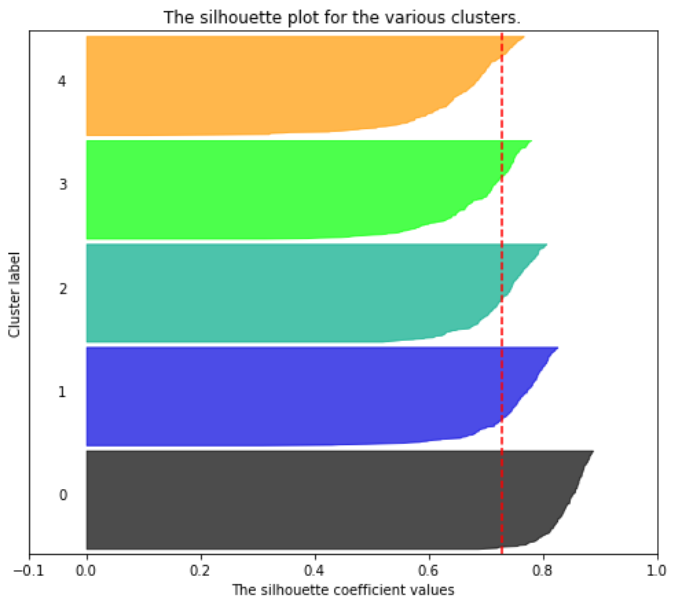
Do đó cho thấy:

|  |  |  |
| --- | --- | --- |
|  |  | (2.13) |

Muốn cho giá trị về gần 1 thì cần phải có . Do là giá trị dùng để đánh giá sự không giống nhau của một phần tử trong cụm, a(i) càng nhỏ thì càng thích hợp. Và thêm nữa, là giá trị chỉ độ không phù hợp với cụm lân cận của nó, b(i) càng lớn thì càng không thích hợp. Nếu s(i) càng về -1 thì ngược lại. Tức giá trị s(i) càng lớn thì độ thích hợp của điểm đó với cụm là càng cao.

Để có thể biểu diễn ra biểu đồ thì ta cần tính giá trị Silhouette cho toàn bộ các điểm và phân cụm mỗi điểm ra. Đồng thời tính giá trị Silhouette trung bình cho cụm đó và biểu thị ra biểu đồ. Thông thường thì ta sẽ chọn K nào có biểu đồ Silhouette có giá trị trung bình cao và phân bố các giá trị s(i) đồng đều, ít hoặc tối thiểu giá trị s(i) nào bé hơn 0 ở mỗi cụm.

Tương tự như Phương Pháp Elbow, việc mô hình sẽ không được đẹp và rõ ràng như dữ liệu kiểm thử. Do đó ta phải dựa vào cảm quan và các yêu tố liên quan khác để chọn được số K phù hợp.



Hình 6. Biêu đồ biểu thị các giá trị Silhouette.

* 1. Phương pháp Gap Statistic

Gap Statistic là một phương pháp khá mới dùng để ước lượng số K tối ưu cho bài toán Kmeans. Phương pháp này được một nhóm nghiên cứu thuộc trường Đại học Stanford phát triển. Phương pháp này ít phổ biến hơn hai phương pháp trên, nhưng kết quả mang lại có thế vượt xa hai phương pháp trên.

Cho tập dữ liệu , với là số lượng điểm dữ liệu độc lập và là số chiều của dữ liệu. Cho là khoáng cách giữa điểm và được tính bằng khoáng cách Euclid. . Cho rằng là ta có tập dữ liệu đã được phân thành cụm và là tập các điểm dữ liệu thuộc cụm và . Cho

|  |  |  |
| --- | --- | --- |
|  |  | (2.14) |

Là Tổng khoảng các giữa các 2 điểm cho tất cả các điểm thuộc cụm , và

|  |  |  |
| --- | --- | --- |
|  |  | (2.15) |

Là Tổng trung bình khoảng cách giữa cặp các điểm của mỗi cụm(ta chia với 2 để thuật toán hoạt động chuẩn nhất).

Ý tưởng của phương pháp này là tìm cách chuẩn hóa việc so sánh với một tập điễm thuộc một phân phối tham chiếu rỗng hay là một phân phối không có sự phân cụm rõ ràng(kí hiệu là \*). Ước lượng cụm tối ưu là khi cụm mà ở đó giảm xa nhất so với đường tham chiếu. Do đó ta có thể định nghĩa phương trình Gap statistic sau:

|  |  |  |
| --- | --- | --- |
|  |  | (2.16) |

Với là một kì vọng của phân phối tham chiếu có phần tử. Để có thể có được là sẽ đi tính giá trị trung bình của bản sao của với . Mỗi bản sao sinh ra ngẫu nhiên bằng cách sử dụng phương pháp Monte Carlo. Với các giá trị đã được tính ở trên, có có thể tìm được đố lệch chuẩn mà sau khi đã được tính toán thêm sự sai lệch do quá trình sinh ta có đại lượng

|  |  |  |
| --- | --- | --- |
|  |  | (2.17) |

Khi đó ta có thể chọn được số K tối ưu là nhỏ nhất sau cho:

|  |  |  |
| --- | --- | --- |
|  |  | (2.18) |

Theo đó ta có các bước để triển khai mô hình như sau:

1. Với tập dữ liệu đã có, xác định số lượng tổng số lượng K từ , và tính với mỗi k tương ứng.
2. Sinh ra tập dữ liệu tham chiếu ngẫu nhiên bằng phương pháp Monte Carlo và tính , tương ứng. Sau đó tính giá trị Gap statistic.

|  |  |  |
| --- | --- | --- |
|  |  | (2.19) |

1. Đặt (2.20), rồi tính độ lệch chuẩn:

|  |  |  |
| --- | --- | --- |
|  |  | (2.21) |

Và xác định .

1. Lừa chọn K tối ưu bằng cách chọn k nhỏ nhất sao cho

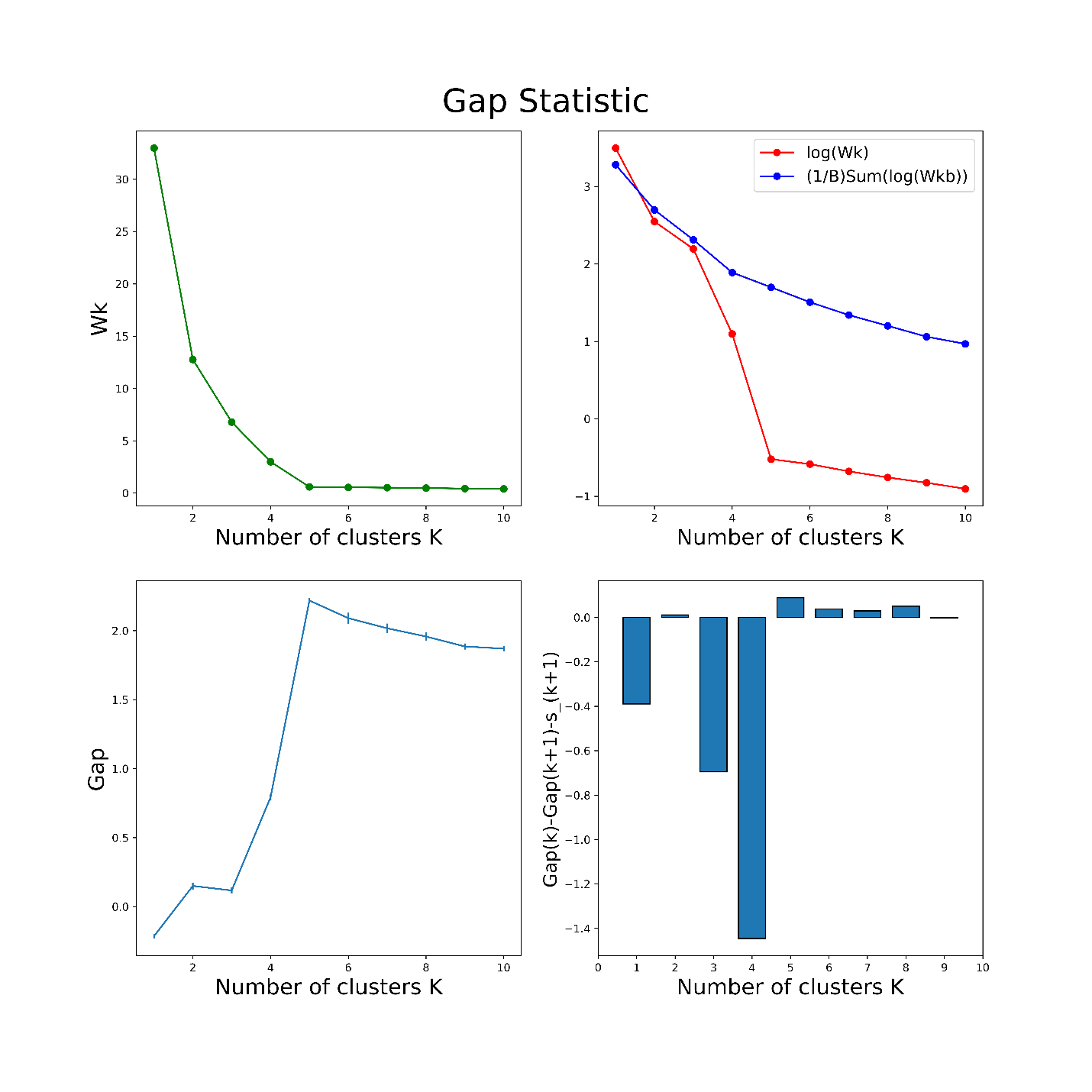
|  |  |  |
| --- | --- | --- |
|  |  | (2.22) |

Kết quả của thuật toán sẽ cho ta rất nhiều biến để có thể biểu diễn lên biểu đồ. Các biến này là giá trị giá trị này có ý nghĩa gần tường tự với phương pháp Elbow, , Gap(k), và các giá trị .

Tương tự như phương pháp Silhouette, phương pháp này đòi hỏi việc tính toán khoảng các của các cặp điểm. Mà như ở bước 2 ta cần sinh ra các tập điễm phẫn phỗi rỗng, do đó thuật toán sẽ mất thêm thời gian để có thể sinh ra các tập này và đồng thời tính toán khoảng cách. Thế nên là tốc độ chạy của thuật toán này sẽ lớn hơn nhiều với các thuật toán trước vì số lượng điểm cần được tính khoảng cách với nhau là rất lớn.

Tuy vậy kết quả bài toán này lại cho ta một cái nhìn tông thể hơn về số K phù hợp là bao nhiêu. Bằng cách sử dụng nhiều biến khác nhau để có thể tham khảo k, một số biến có thể bị lệch nhưng biến khác lại mang lại kết quả chính xác hơn.

Sau khi hoàn thành các bước để tiến hành thuật toán, ta có thể sử dụng biểu đồ sau để đánh giá kết quả bài toán:

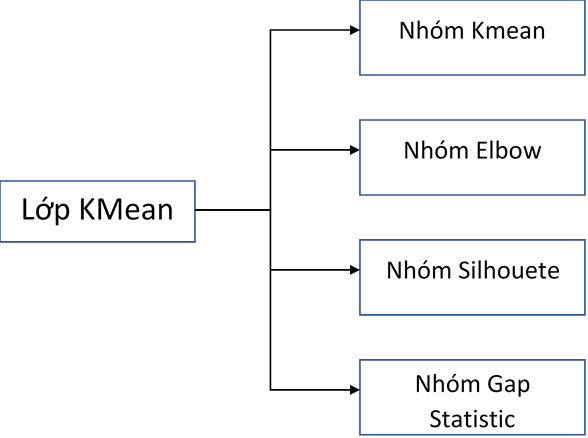


Hình 7. Biểu đồ phương pháp Kmean

Biều đồ này sinh bằng phương pháp Gap statistic bởi cùng một bộ dữ liểu đã sự dụng ở 2 phương pháp trước để cho kết quả đánh giá tối ưu nhất. Như ta đã thấy ở trên biểu đồ, có tổng cộng 4 hình khác nhau. Hình thứ nhất chính là kết quả của phương pháp Elbow, nhưng Wk có cách tính khác hơn đôi chút so với phương pháp Elbow gốc. Cách xác định K ở hình này tương tự như ở phương pháp Elbow. Ở hình thứ hai, ta có hai đường màu đỏ và xanh dương. Đường màu đỏ biểu thị giá trị log(Wk) và đường màu xanh ở phía trên biểu thị giá trị trung bình của các log(Wkb). Ta có thể so sanh 2 hai giá trị này với nhau để xác định được K chính xác. Ở hình này K có thể được xác định bằng cách xác định khoàng cánh ở giữa xa nhất đầu tiên ở hai đường này hay con gọi là Gap, và hình trên chỉ ra K=5 chính là giá trị đúng. Với hình thứ 3 biểu thị giá trị của Gap tương ứng như là khoảng cách ở giữa của hình thứ 2, Giá trị này càng lớn thì khả năng K chính xác càng lớn. Hình thứ 4 là mức thay đổi giá trị Gap so với Gap ở k tiếp theo sau khi đã cộng thêm độ lệch chuẩn. Tương tự thì độ lệch này càng cao thì khả năng giá trị K tại điểm đó chính xác càng cao.

1. Xây dựng Lớp Kmean và Kiểm Thử
   1. Xây dựng Lớp Kmean.

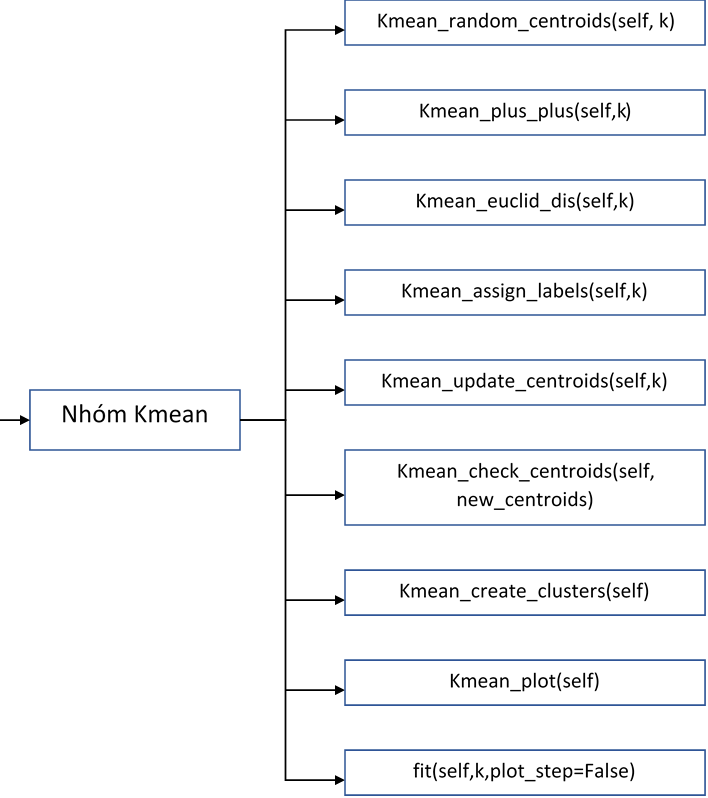
Việc xây dựng lớp Kmean sẽ được xây dựng theo mô hình sau.



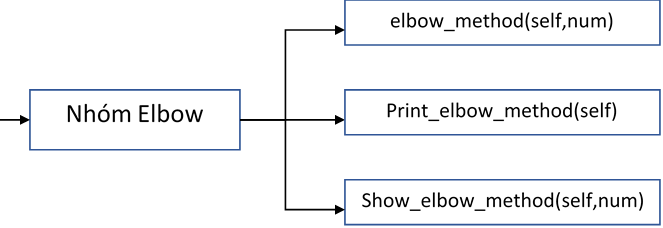
Hình 8. Lớp Kmean

Trong lớp Kmean sẽ có 3 nhóm chức năng chính là:

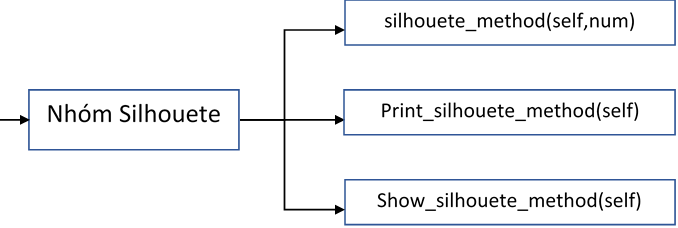
* Nhóm Kmean: Gồm các hàm và lệnh để hỗ trợ và thực thi thuật toán Kmean.
* Nhóm Elbow: Các hàm thực thi phương pháp Elbow
* Nhóm Silhouette: Các hàm thực thi phương pháp Silhouette.
* Nhóm Gap Statistic:Các hàm thực thi phương pháp Gap Statistic



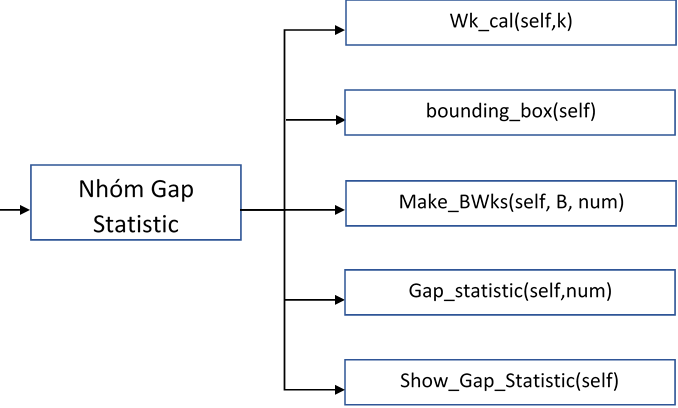
Hình 9. Nhóm Kmean



Hình 10. Nhóm Elbow



Hình 11. Nhóm Silhouette



Hình 12. Nhóm Gap Statistic

* 1. Thư viện sklearn và hàm make\_blops.

Để có thể chọn kiểm tra được tính chính xác và hiệu quả của bài toán, ta cần có một tập dữ liệu kiểm thử. Điều kiện để có một bộ dữ liệu kiểm thử đẹp là các cụm được phân tích rõ ràng, sự phân tán không quá lớn, độ lớn của giá trị không quá cao và có sự nhất quán nhật định. Tuy nhiên các tập dữ liệu thực tế hiếm khi có được những tính chất như vậy dể đánh giá. Do đó ta cần một phương án khác để thay thế.

Phương án thay thế đó là hàm make\_blops trong thư viện sklearn. Thư viện sklearn là một thư viện nổi tiếng trong Machine Learning và lập trình python. Nó chưa nhiều các hàm khác nhau để hỗ trợ việc tính toán, phân tích dữ liệu, các thuật toán máy học, tạo và xữ lý dữ liệu,...Trong đó có hàm make\_blops.

Hàm make\_blops là một hàm giúp tạo ra số lượng tùy ý các cụm theo hàm phân phối chuẩn Gaussian giúp hỗ trợ kiểm tra các bài toán phân lớp. Trong hàm này có các đại lượng có thể tùy chỉnh như:

* n\_sample: số lượng điểm dữ liệu đưa vào
* n\_features: Số lượng thuộc tính hay số nhiều đưa vào.
* centers: số lượng tâm hay cụm tạo ra
* cluster\_std: độ lệch chuẩn theo phân phối Gaussian, kiểm soát được phân bộ hay độ rộng của mỗi cụm.
* random\_state: Xác định việc tạo ngẫu nhiên của tập dữ liệu, nếu đưa vào cùng một số và các tham số khác dữ nguyên thì đâu ra của hàm luôn giống nhau.

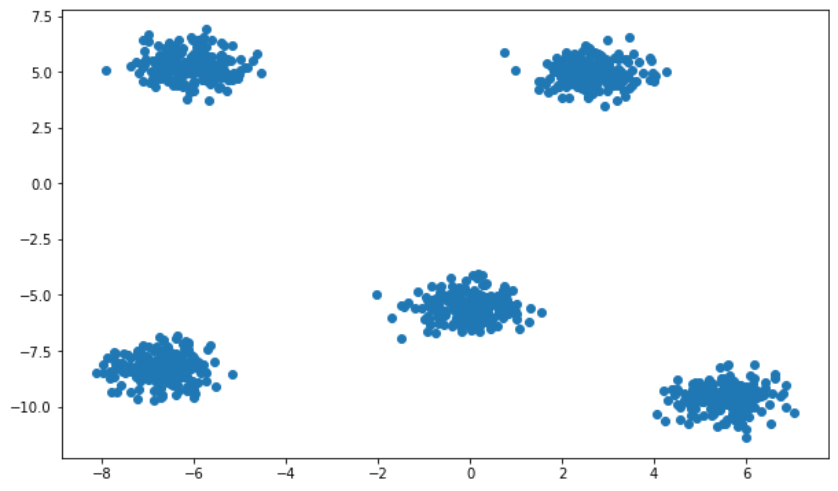
Ta thêm thư viện và hàm.

1. from sklearn.datasets import make\_blobs

Sau đó ta có thể tạo ra một tập dữ liệu để tính toán và kiểm thử như sau, các tham số đưa vào lần lượt là: số điểm = 1000, số chiều = 2, số tâm = 5, std = 0.5, tham số ngẫu nhiên = 10.

1. features, true\_labels = make\_blobs(n\_samples=1000,n\_features=2,centers=5,cluster\_std=0.6,random\_state=10)
2. fig = plt.figure(figsize=(10,6))
3. ax = fig.add\_subplot()
4. ax.scatter(features[:,0],features[:,1])
5. plt.show()

Theo đó ta có một tập dữ liệu và tập dữ liệu đó có dạng:



Hình 13. Tạo dữ liệu

Theo như dữ liệu tạo ra ta có số tâm đúng K=5.

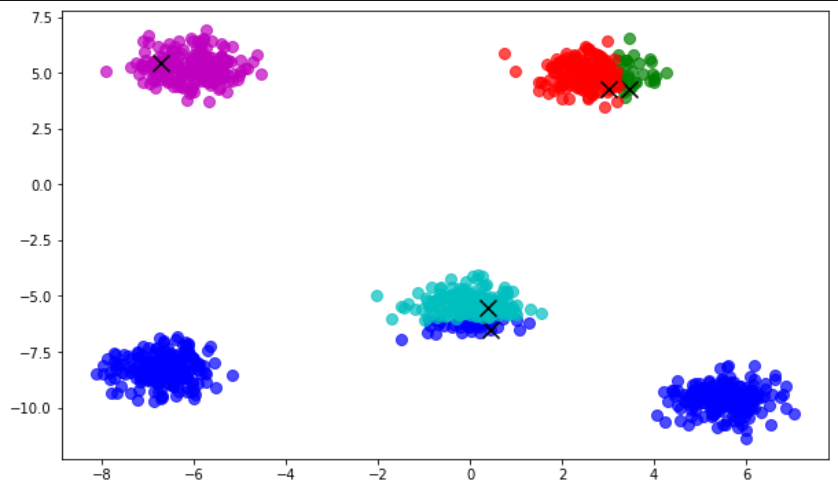
* 1. Thực thi phân cụm bằng Kmean.

Để kiểm tra xem hàm Kmean đã được xây dựng chính xác chưa. Ta sẽ chạy thử trên bộ dữ liệu vừa được tạo ở trên. Ta sẽ xuất ra các bước điều chỉnh tâm. Tuy nhiên trong bài báo cáo sẽ không đưa ra hết toán bộ các bước mà chỉ lấy bước khởi tạo đầu tiên và bước cuối.

Ta sẽ chạy thử với các tâm đầu vào được khởi tạo ngẫu nhiên. Với cách khới tạo ngẫu nhiên thì khả năng rất cao là các tâm khởi tạo ban đầu sẽ sai dẫn đến kết quả cuối cùng không chính xác.

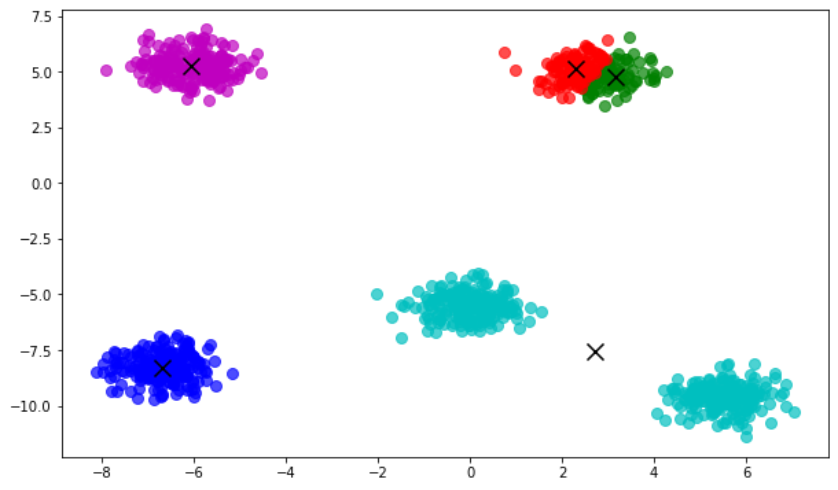
Ta sẽ gán tham số init='rd' để có thể khởi tạo tâm ngẫu nhiên và plot\_steps = True để có tể biểu thị quá trình cập nhật tâm. Hàm .fit(5) để thực kiện thuật toán Kmean với số tâm là 5

1. Km1=KMean(features,init='rd',plot\_steps=True)
2. values1=Km1.fit(5)



Hình 14. Kmean init='rd' – 1

Ta có thể thấy ngay ở bước đầu tiên thì có 2 cặp tâm được khởi tạo quá gần nhau. Điều này có thể dẫn đến số lần chạy lớn để có thể điều chỉnh được tâm và khả năng lớn dẫn đến kết quả cuối cùng sai.



Hình 15. Kmean init='rd' – 2

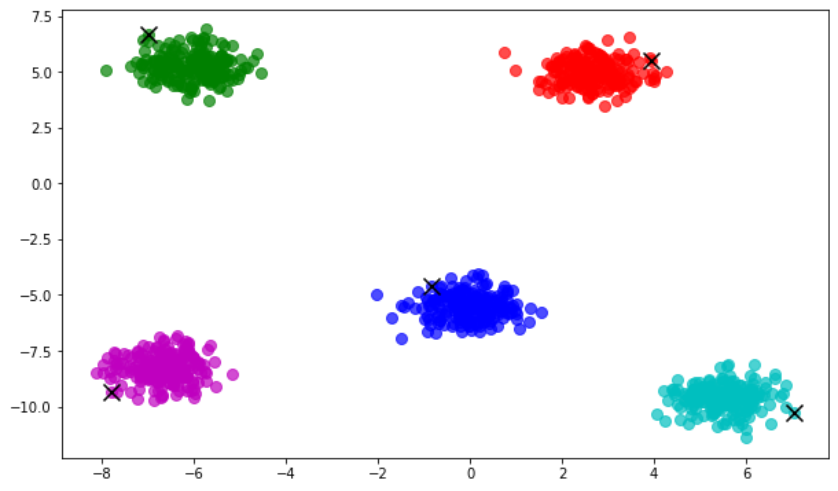
Như đã nói ở trên thì sau 11 lần cập nhật thì cho ra kết quả. Dễ thấy rằng một số tâm đã bị điều chỉnh sai.

Lưu ý: Với việc khởi tạo tâm ngẫu nhiên thì kết qua cho cuối cụng cho ra vẫn sẽ chính xác tuy nhiên với tỉ lệ thấp. Khi sử dụng cho các bài toán lớn hơn thì có thể dẫ đến kết quả sai hoặc tốc độ tính toán chậm do số lần cập nhật tâm lớn. Và mô hình trên chỉ chứng tỏ việc tâm nhẫu nhiên dẫn đến kết quả cuồi cùng có thể bị sai.

Ta sẽ tiếp tục chạy thuật toán Kmeans với các tâm ban đầu được khởi tạo ban đầu sử dùng thuật toán Kmeans++. Với điều này, ta kĩ vọng sẽ cho kết quả chính xác và sẽ dựa trên kết quả này để đánh giá xem thuật toán đã chạy đụng chưa.

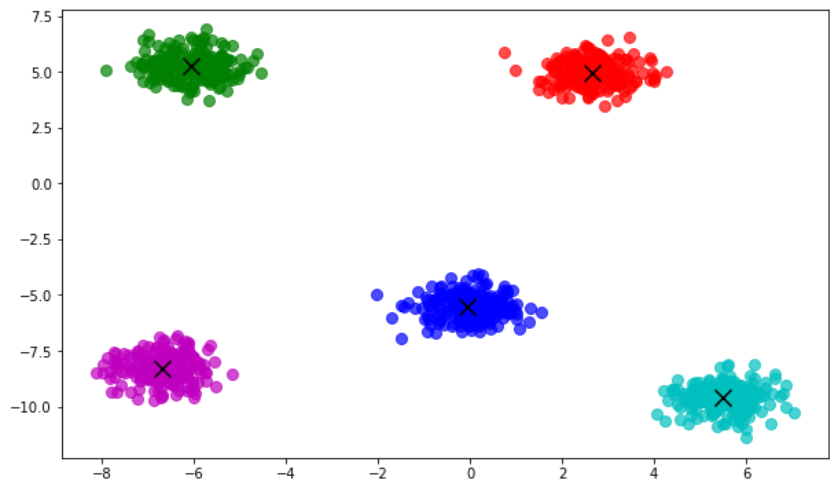
Vì thuật toán Kmean đã được xây dựng với thuật toán khởi tạo tâm ban đầu là Kmean++ nên ta không cần gán giá trị init. Plot\_steps=True để biểu thị quá trình khởi tạo tâm và .fit(5) để thực thi thuật toán Kmean với số tâm là 5.

1. Km2=KMean(features,plot\_steps=True)
2. Values2=Km2.fit(5)



Hình 16. Kmeans init = 'Kmeans++' – 1

Dễ thấy rằng, các tâm lần này đã được điều chỉnh theo thuật toán Kmeans++ đã được phân bố rộng hơn. Điều này cho phép việc cập nhật tâm trở nên nhanh hơn và chính xác hơn.



Hình 17. Kmeans init = 'Kmeans++' – 2

Chỉ cần 1 lần cập nhật tâm thì các tâm và các cụm đã được cập nhật chính xác.

**Kết luận:** Từ các kết quả trên, cho thấy các hàm và thuật toán đã được xây dựng chính xác.

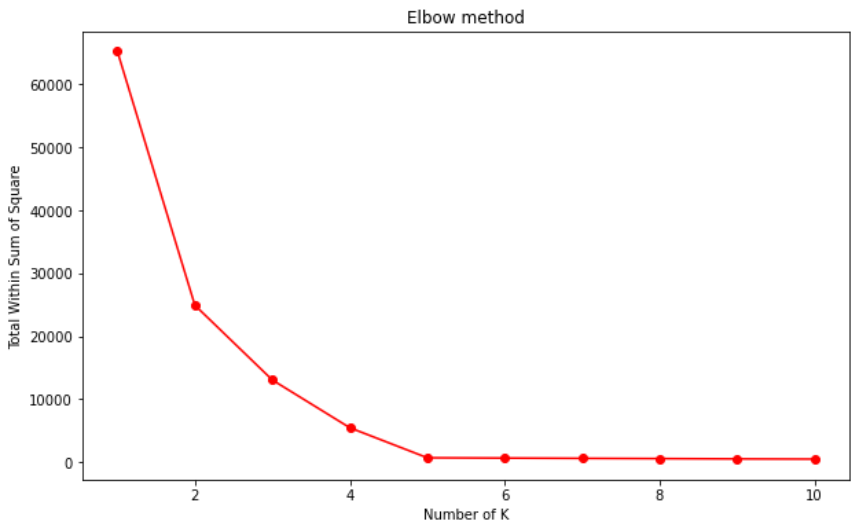
* 1. Lựa chọn K trên phương pháp Elbow.

Ta sẽ kiểm tra và tìm số k tối ưu dưa trên phương pháp Elbow với hàm đã được xây dụng. Ta sẽ chọn số lượng dữ đoán tối đa là 10 khi đó ta có thể bằng lệnh như sau.

1. Km3.elbow\_method(10)
2. Km3.Show\_elbow\_method()

Kết quả:

1. k=1:65250.6871957869
2. k=2:24908.134768466385
3. k=3:13030.531842396324
4. k=4:5831.7188469759885
5. k=5:669.4928101529986
6. k=6:627.512762141783
7. k=7:576.1830116287447
8. k=8:529.647378440143
9. k=9:486.1602391797609
10. k=10:440.3949439862806



Như ta thấy ở trên thì K=1 có giá trị lớn nhất và luôn sẽ là lớn nhất. Sau đó các giá trị bắt đầu giảm giần và dừng lại ở K=5. Sau K=5 thì giá trị không có sự biến động cho thấy K=5 là tối ưu nhất.

**Kết luận:** Phương pháp này đã có thể tìm được số K chính xác.

* 1. Lựa chọn K trên phương pháp Silhouette.

Tương tự với phương pháp Elbow. Ta chạy hàm Silhouette như sau với K tối đa là 7.

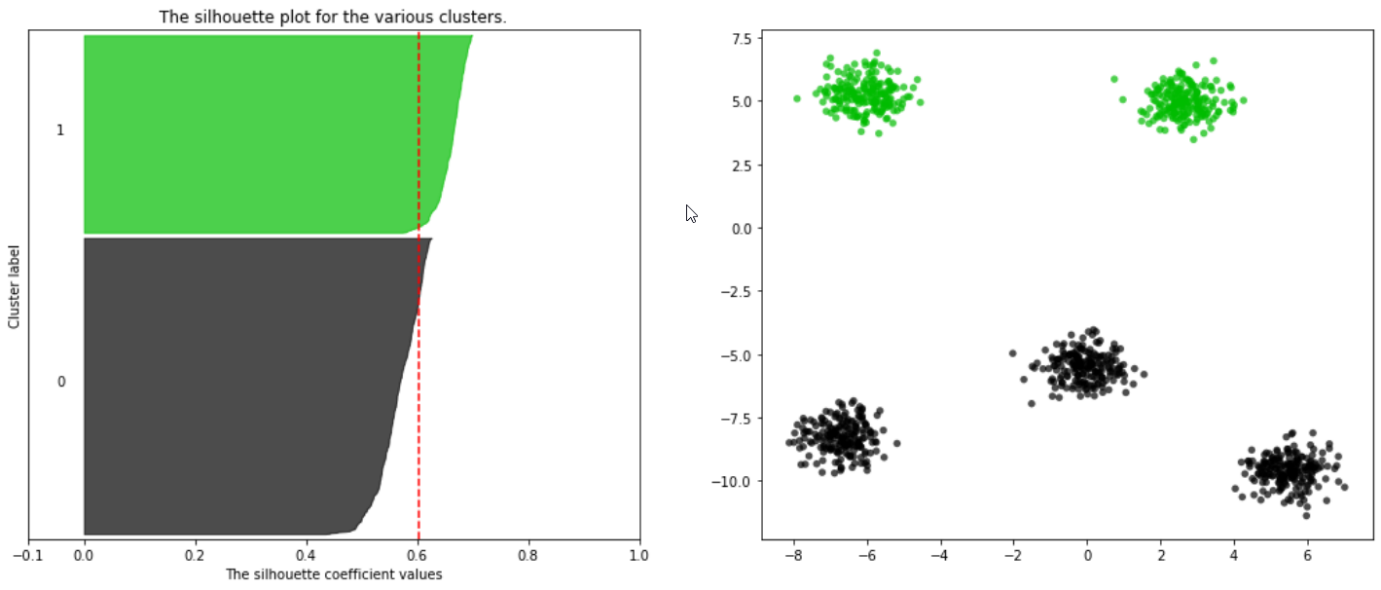
1. Silhouette\_values1=Km3.silhouette\_method(7)
2. Km3.Print\_silhoutte\_method()
3. Km3.Show\_silhoutte\_method()

Kết quả:

1. k=2:0.6036157364018033
2. k=3:0.6530334786547721
3. k=4:0.7543039824604906
4. k=5:0.8629392768705967
5. k=6:0.7545617804247856
6. k=7:0.6362134294159024

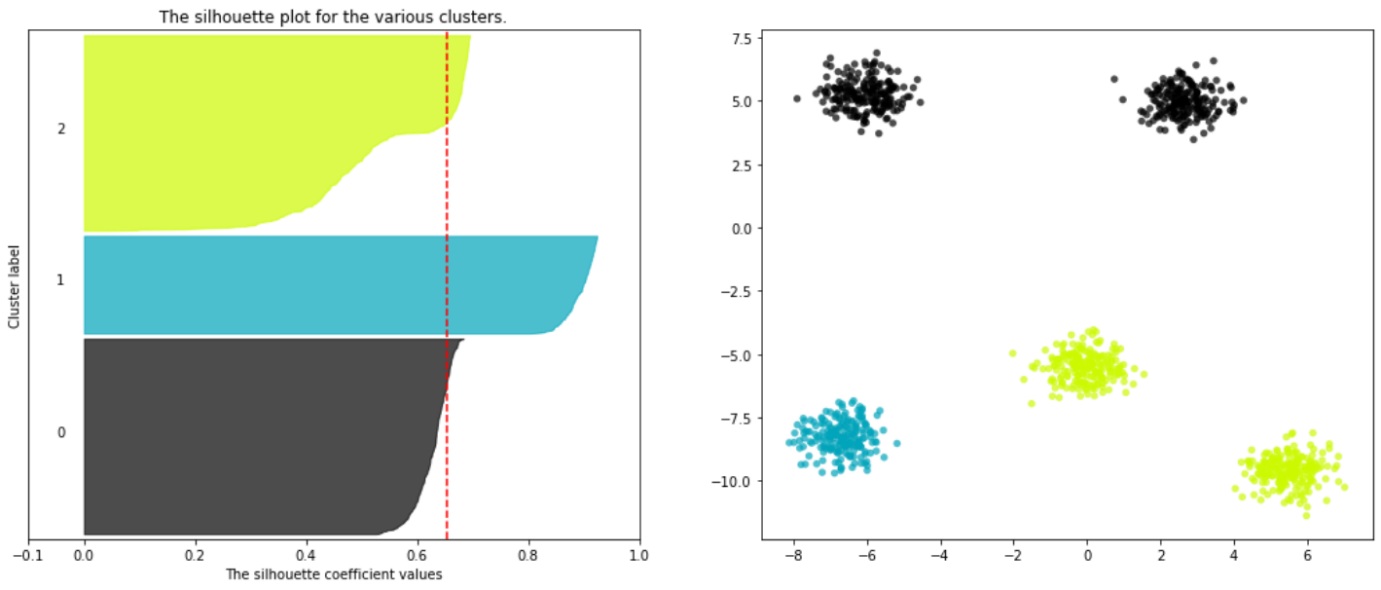
Như ta có thể thấy kết quả ở trên. Với K = 5 cho ra kết quả cao nhất so với các K còn lại. Điều có thế cho ta biết rằng, với K = 5 thì độ tách biệt giữa cụm này và các cụm khác của giữ liệu là cao nhất. Cụ thể hơn ta có thể xem mô hình phân tích điểm Silhouette cho từng điểm và mô hình chia cụm.

Chú ý: Mô hình phía bên phải của mỗi hình chỉ có thể biểu diễn được dưới dạng 2 chiều và không thể biễn diễn cho số chiều khác lơn hơn. Mỗi cụm được gãn nhãu và có màu riêng biệt. Mỗi điễm dữ liệu được gán nhãn và phân mỗi cụm, được tính điểm silhouette và sắp xếp lại và được biểu diễn ở mô hình bên trái. Đồng thời đã được căn chỉnh tỷ lệ sao cho thích hợp với mô hình. Đường gạch đỏ là điểm Silhouette trung bình cho mỗi cụm được có được ở kết quả.



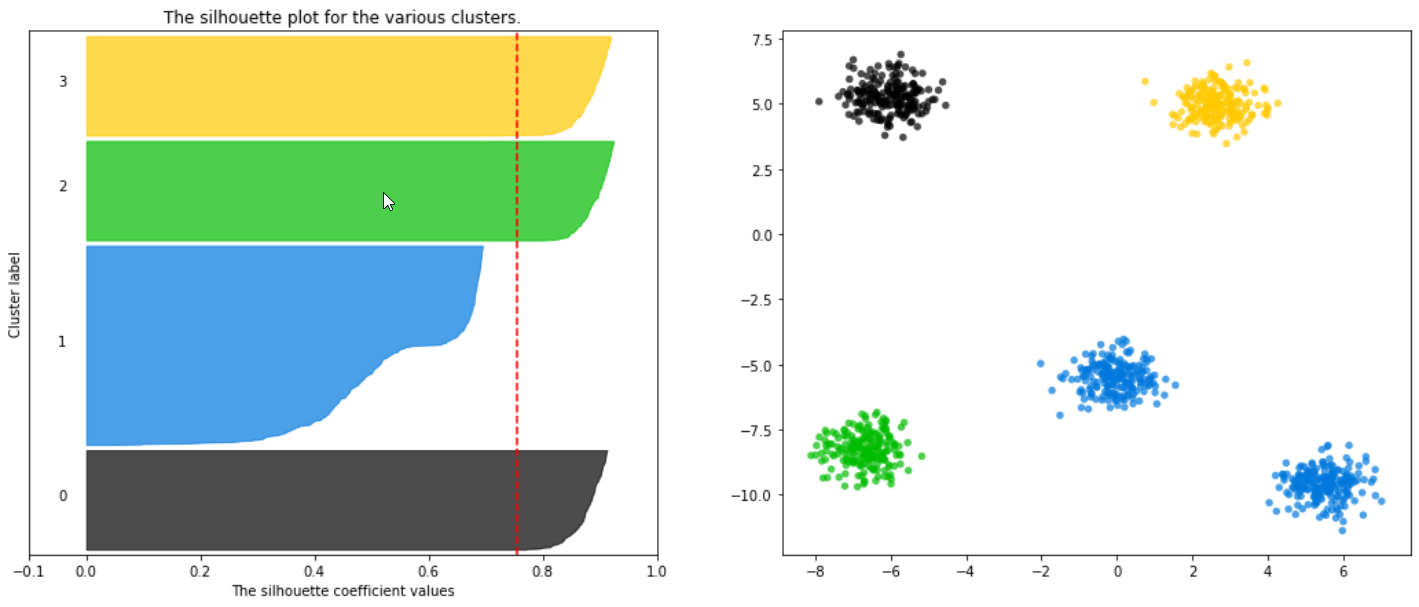
Hình 18. Mô hình Silhouette K=2

Với K = 2 thì ta có thể coi là các cụm đã được phân tách với nhau và số điễm mỗi cụm là khá đồng đều. Tuy vậy hơn 2/3 số điểm ở cụm 0 là dưới trung bình và mô hình bên phải cho thấy k=2 vẫn chưa thích hợp.



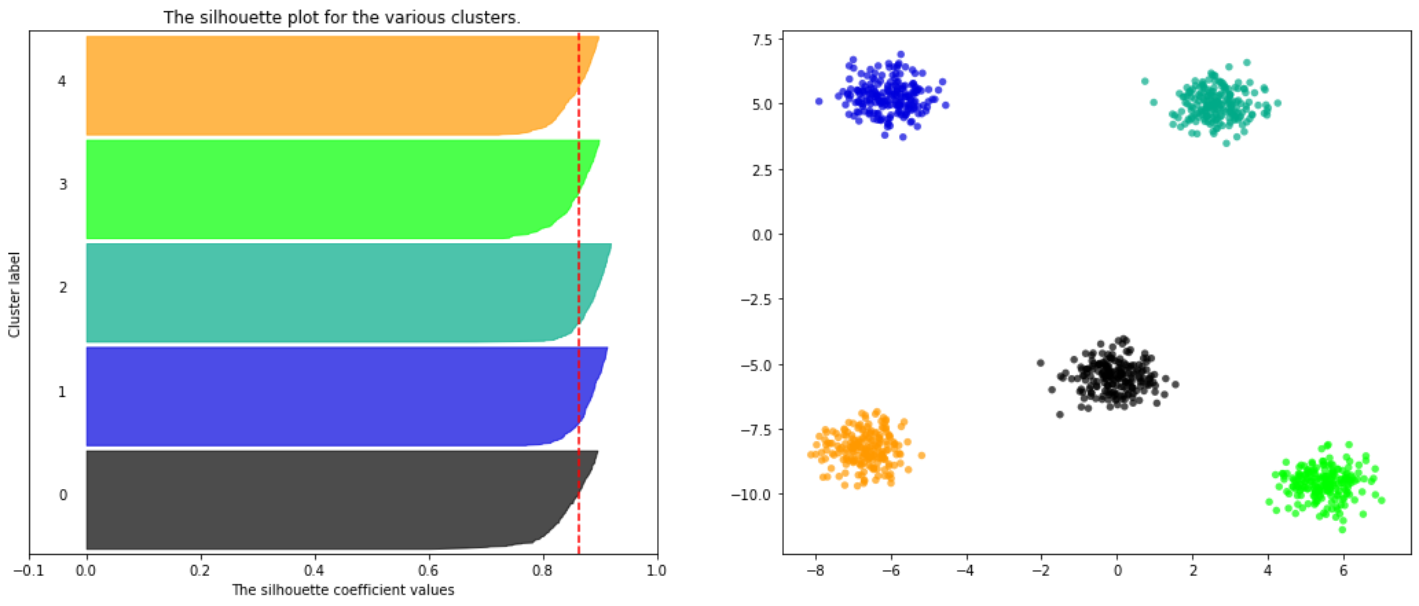
Hình 19. Mô hình Silhouette K=3

Với K=3 ta có thể nhận thấy một số điểm bất thường trong phân bổ điểm silhouette ở các cụm. Cụm nhãn 1 có điểm rất cao chúng cho cụm này đã được phân chia rõ rệt với các cụm khác trong khi các cụm còn lại có số điểm thấp hơn trung bình và cụm nhãn 2 phần nữa có giá trị thấp.



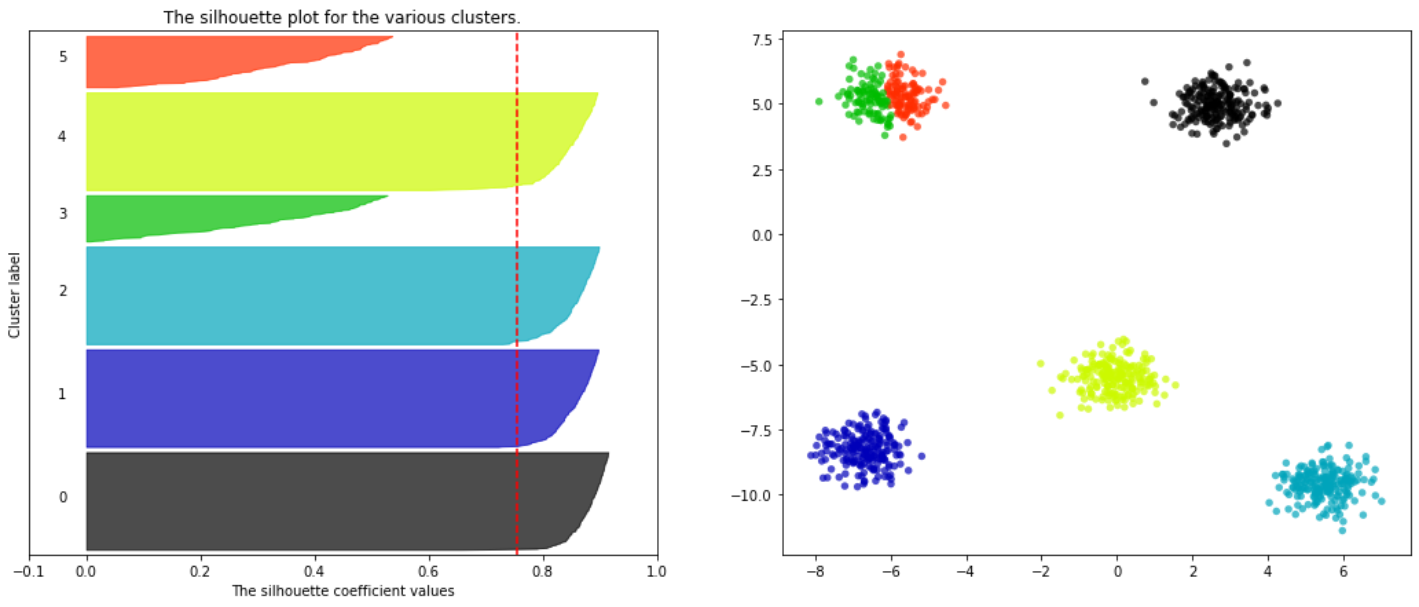
Hình 20. Mô hình Silhouette K=4

Với K=4 ta dễ dàng nhận thấy được các cụm có nhãn 0,2,3 có số điểm cao và đồng đều. Chúng cho các cụm này đã được phân chia tốt. Còn Cụm nhãn 1 không đạt giá trị trung bình.

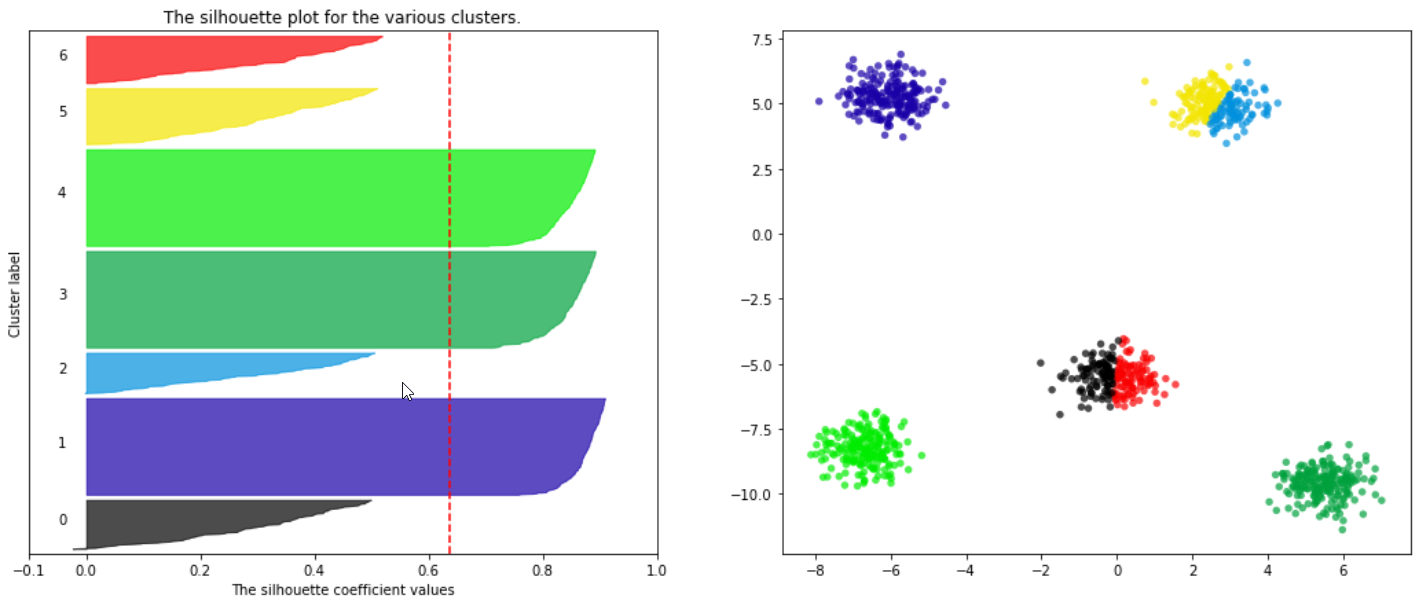


Hình 21. Mô hình Silhouette K=5

Với số K=5 ta có thấy số điểm của mỗi cụm là rất cao và không có cụm nào là khác biệt. Cho thấy đây là mô hình đẹp nhất và ta nên chọn số K=5. Qua mô hình này ta có thể có cái hình tổng quan nhất về thế nào khi các cụm được chia tách rõ ràng.



Hình 22. Mô hình Silhouette K=6



Hình 23. Mô hình Silhouette K=7

Đối với số cụm = 6 và 7 ta dễ dàng nhận thấy có một số cụm không đạt ngưỡng trung mình và điểm rất thấp so với các cụm còn lại nên ta không chọn số k tượng tự

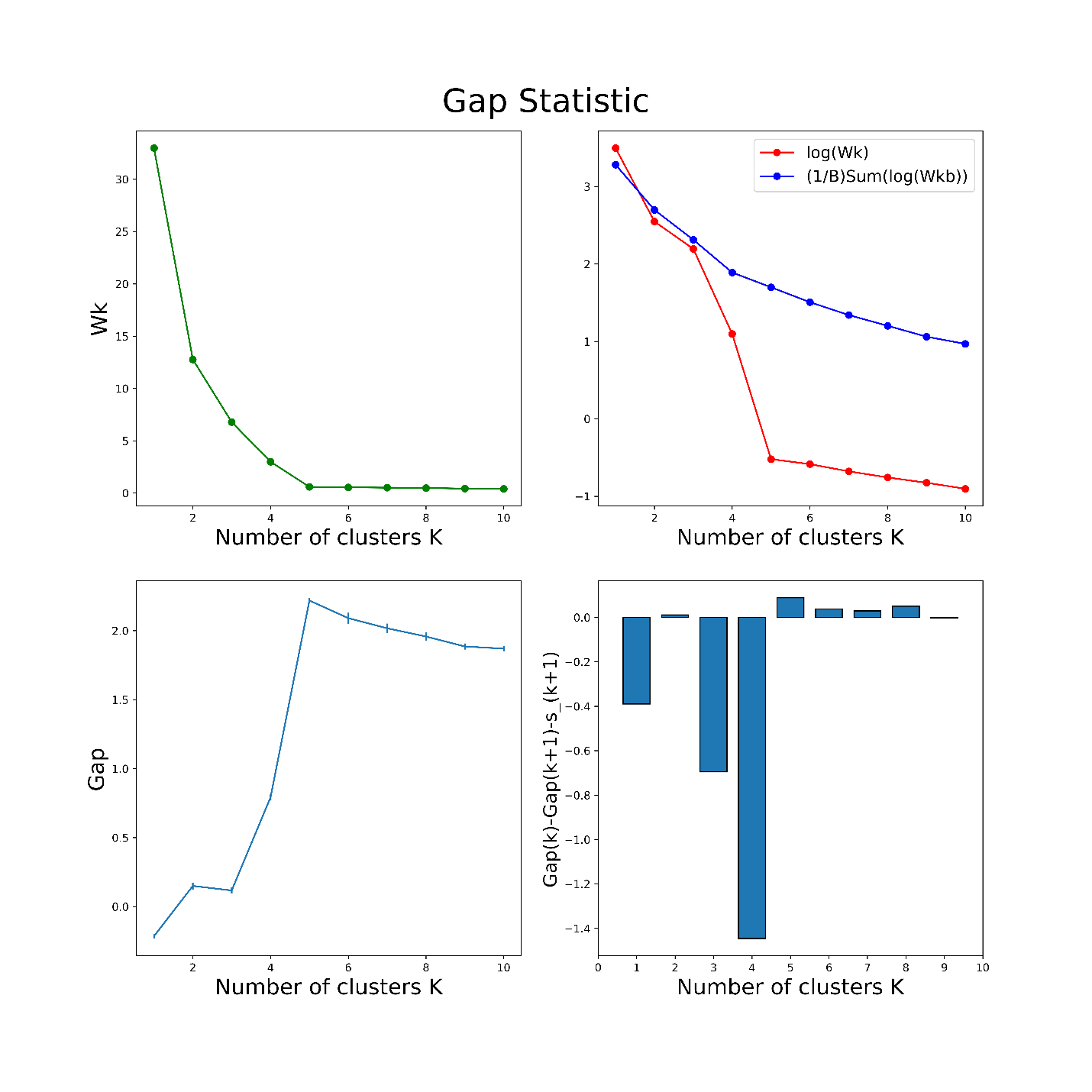
**Kết luận:** Với mô hình K=5 như trên thì đã chứng tỏ rằng được rằng hàm Silhouette đã được xây dựng chính xác và cho ra được số k đúng.

* 1. Lựa chọn K trên phương pháp Gap Statistic

Tương tự như hai phương pháp trên, ta sẽ thực thi phương pháp Gap Statistic trên lương num=10.

1. Km3.Gap\_statistic(10)
2. Km3.Show\_Gap\_Statistic()

Tương tự như đã giới thiệu, ta sẽ có 4 hình biệu thị cho các giá trị khác nhau của phương pháp Gap Statistic.



Hình 24. Mô hình Gap Statistic

Với mô hình trên thì ta có thể dễ thấy ở K=5 thì tất cả các hình đều chỉ ra răng là K = 5 là K phù hợp nhất với tập dữ liệu. Với hình một tương tự phương pháp Elbow, giá trị từ K=5 về sau không thay đổi nữa. Với hình 2 thì khoảng cách(Gap) ở K=5 là lớn nhất. Với hình 3 là phản ảnh của hình 2 thì K=5 cũng đạt giá trị là cao nhất. Tương tự với hình thứ 4 thì cột thự 5 có giá trị lớn nhất.

**Kêt luận:** Qua 4 hình ở phương pháp Gap Statistic đều đồng ý là K=5 là phù hợp nhất, cho nên K = 5 là kết quả dự đoán của phương pháp này. Theo đó thì phương pháp này đã chứng minh được tín chính xác của mình.

KẾT LUẬN

So sánh các thuật toán.

Ưu điểm chung: các thuật toán đã có thể chạy thành công và cho ra được số K đúng đối với điều kiện dữ liệu lý tưởng.

Nhược điểm chung: Các thuật toán sẽ chạy chậm hơn đáng kể với số K đầu vào lớn.

Phương pháp Elbow: Nhìn chung thì đơn giản hơn so với phương pháp Silhouette. Có tốc độ thực thi cao hơn. Tuy vậy mức chi tiết thì sẽ kém hơn.

Phương pháp Silhouette: Phức tạp và có tốc độ thực thi chậm hơn phương pháp Elbow. Tuy nhiên phương pháp này cho cái nhìn cụ thể hơn đối với mỗi cụm trong mỗi K ta lựa chọn, cho phép ta có thể lựa chọn chính xác hơn. Tuy vậy phương pháp này tỏ ra nhược điểm là có quá nhiều mô hình để quan sát. Điều này sẽ gây khó khăn cho người dùng về sau.

Phương pháp Gap Statistic: Với sự phức tạp trong tính toán, phương pháp này tỏ ra nhiều quả hơn so với 2 phương pháp trên khi chỉ ra được nhiều đặc điểm hơn để có thể nhận diện được K chính xác. Nhược điểm lớn nhất của thuật toán này là có thời gian chạy lớn hơn rất nhiều. Và có thể cần được xem xét sử dụng khi áp dụng cho lượng dữ liệu lớn.

Đánh giá các thuật toán

Để có một cái nhìn tổng quan nhất về các phương pháp và đảm bảo rằng, các phương pháp có thể chạy chính xác trên các tập dữ liệu khác nhau. Các phương pháp sẽ được chạy thử trên các tập dữ liệu kiểm thử khác nhau. Các tập dữ liệu đó bao gồm

* Tập dữ liệu trống, tức chỉ cho 1 cụm, và số chiều lớn.
* Tập dữ liệu tiêu chuẩn, tập dữ liệu với số cụm và số chiều tối thiểu để có thể thực hiện kiểm kiểm thử chính xác.
* Tập dữ liệu có số chiều và cụm nhiều hơn một chút và gấp đôi số lượng dữ liệu. Việc tăng gấp đôi này nhằm để đánh giá sự phụ thuộc khối lương vào tộc độ thực thi thuật toán.
* Tập dữ liệu với số chiều lớn. Để kiểm chứng các phương pháp có thể thực thi trên các tập dữ liệu có số chiều lớn.

20 tập dữ liệu kiểm thử sẽ được sinh bằng thư viện để có kết quả chính xác nhất với mỗi trường hợp đã liệt kê ở trên. Các phương pháp sẽ được thực hiện và xét xem các phương pháp có kết quả chính xác hay không và thời gian chạy trung bình là bao lâu.

Bảng dưới đây là kết quả sau quá trình kiểm thử. Số lượng K tối đa là 9, số lượng dữ liệu là n=500. Số lượng các kết quả đúng dự kiến sẽ được đánh dấu là: *k* với k là các số tự nhiên từ 0.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ***Phương pháp*** | ***Thời gian trung bình(s)*** | ***Số lượng ước tính của mỗi cụm k dự đoán*** | | | | | | | | |
|  |  | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** |
| K = 1, d = 10, n=500 | |  |  |  |  |  |  |  |  |  |
| Elbow | *0.79295* | *12* | 0 | 1 | 0 | 1 | 2 | 1 | 1 | 2 |
| Silhouette\* | *2.8515* | *0* | *18* | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| Gap statistic | *9.275* | *13* | 1 | 4 | 0 | 0 | 0 | 0 | 1 | 1 |
| \*Vì Silhouette cần phải có ít nhất 2 cụm để thực thi nên kết quả sẽ không thể rơi vào k = 1 | | | | | | | | | | |
| K=3, d = 2, n=500 | |  |  |  |  |  |  |  |  |  |
| Elbow | *0.67335* | 0 | 0 | *20* | 0 | 0 | 0 | 0 | 0 | 0 |
| Silhouette | *1.88565* | 0 | 0 | *20* | 0 | 0 | 0 | 0 | 0 | 0 |
| Gap statistic | *8.51415* | 0 | 0 | *20* | 0 | 0 | 0 | 0 | 0 | 0 |
|  | | | | | | | | | | |
| K = 6, d = 3, n = 1000 | |  |  |  |  |  |  |  |  |  |
| Elbow | *1.4345* | 0 | 0 | 0 | 0 | 6 | *14* | 0 | 0 | 0 |
| Silhouette | *4.4105* | 0 | 0 | 0 | 0 | 2 | *18* | 0 | 0 | 0 |
| Gap statistic | *17.07* | 0 | 0 | 0 | 0 | 0 | *20* | 0 | 0 | 0 |
|  | | | | | | | | | | |
| K=7, d = 10, n = 500 | |  |  |  |  |  |  |  |  |  |
| Elbow | *0.7107* | 0 | 0 | 0 | 0 | 0 | 1 | *19* | 0 | 0 |
| Silhouette | *2.77* | 0 | 0 | 0 | 0 | 0 | 1 | *19* | 0 | 0 |
| Gap statistic | *9.379* | 0 | 0 | 0 | 0 | 0 | 0 | *20* | 0 | 0 |

Bảng 1. Kết quả kiểm thử

Đánh giá kết quả: từ bảng trên cho ta thấy được các phương pháp đều hiện việc ước tính K tối ưu theo như dự đoán. Đối với trường hợp số cùm và số chiều là tối ưu thì tất cả các phương pháp đều dự đoán chính xác. Với phường pháp Gap Statistic thì tỉ lệ dự đoán là 100% với các tập dữ kiểu không phải rỗng.

Từ số liệu cho thấy thì yếu tố ấy ảnh hưởng đến tốc độ thực thi không phải là số chiều hay cụm của dữ liệu mà là số lượng dữ liệu. Khi số lượng dữ liệu tăng gấp đôi thì thời gian đồng thời tăng tương ứng. Từ thời gian chạy cho thấy Gap Satistic cho thời gian chạy lớn nhất so với 2 phương pháp còn lại.

Các kết quả đạt được

Qúa quá trình nghiên cứu thì bài báo cáo này đã đạt được môt số kết quả nhất định.

Thứ nhất, đã làm rõ được các vấn đề liên quan tới thuật toán Kmeans. Bài báo cáo đã trình bày những nét nổi bật, công dụng của thuật toán phân cụm Kmeans. Đồng thời nêu lên được thuật toán và cách thực thi trên ngôn ngữ lập trình Python. Kết hợp với một số cải thiện trọng việc khởi tạo tâm ban đầu của thuật toán. Làm cho thuật toán này được đảm bảo sự tin cậy.

Thứ hai, bài báo cáo này đã trình bày được 3 phương pháp phổ biến và thông dụng để tìm K tối ưu đó là Elbow method, Silhouette method và Gap Statistic. Đã giải thích được thuật toán cũng như là cách sử dụng. 3 phương pháp này đã được thực thi thành công trên ngôn ngữ lập trình Python. Đồng thời đã cho ra được một số kết quả như ý muốn.

Thứ ba, với các phương pháp tìm K có sẳn. Đã có thể áp dụng được để tìm K tối ưu cho một số tập dữ liệu mẫu. Tuy rằng kết quả phân cụm có thể khác với thực tế. Những đã cho thấy rằng việc thực thi các phương pháp trên là thành công.

Các hạn chế

Kmeans là một thuật toán đơn giản và phổ biển dùng để phân cụm dữ liệu. Tuy vậy Kmeans gặp phải nhiều hạn chế khi chỉ có thể phân cụm dữ liệu mạng một số thuộc tính nhất định. Khi gặp một số cụm giữ liệu phức tạp, có kiểu hình khác nhau khí khó có thể áp dụng đúng được. Đồng thời với tập dữ liệu có số chiều D lớn thì tý lệ chính xác phần nào cũng giảm. Điều này là cho trong bào báo cáo này Kmeans vẫn chưa được điều chính về phương pháp tính khoảng cách. Chú yếu vẫn dùng phương pháp tính khoảng cách cố điển là Euclid.

Qua trình vẫn dụng dữ liệu thực tế cho thấy k có thể bị lệch do dữ liệu. Trong thực tế thì các tập dữ liệu thì các cụm có thể được phân bố ngẫu nhiên và không đồng đều. Đẫn đến số K tối ưu có thể phân bố trong một khoảng nào đó mà không phải là một số K cụ thể. Điều này dẫn đến việc phân tích kĩ càng hơn để chọn số K cụ thể hơn mà bài báo cáo này vẫn chưa đề cập tới được.

Hướng phát triển:

Bài toán này có thể được phát triển trên bằng cách tiếp tục tối ưu một số khía cạnh khác của Kmean như cách tính khoảng cách.

Dữ liệu là một vấn đề lớn đối với mọi bài toán Machine Learning. Nếu có thể tối ưu dữ liệu bằng nhiều cách khác như giảm chiều, giảm độ nhiễu,… thì bài toán sẽ cho kết quả chính xác hơn và nhanh hơn.

Vận dụng vào một số lĩnh vực mà Kmeans có thể áp dụng tốt như phân loại hình ảnh, nhận diện chữ biết, phân bố dân cư…

TÀI LIỆU THAM KHẢO

J.Rousseeuw, P. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, Pages 53-65.

Tiệp, V. H. (2020). *Machine Learning cơ bản.* VietNam.

Vassilvitskiz; David Arthur; Sergei. (n.d.). *k-means++: The Advantages of Careful Seeding.*