

BÁO CÁO CUỐI KỲ

Môn học

CS2205.CH1501 - RM.FinalReport


PHƯƠNG PHÁP LUẬN NCKH

Giảng viên

PGS.TS. LÊ ĐÌNH DUY

Thời gian

03/2021 - 06/2021

| | |
|-----------------------|--|
| Họ và tên (IN HOA) | LÊ QUANG KỲ |
| Ảnh |  |
| Số buổi vắng | 0 |
| Bonus | 28 |
| Tên đề tài (VN) | PHƯƠNG PHÁP PHÂN CỤM DỮ LIỆU DỰA TRÊN TẬP THÔ, HỖ TRỢ XÁC ĐỊNH SỐ LƯỢNG PHÂN CỤM TỐI ƯU |
| Tên đề tài (EN) | <i>OPTIONAL - KHÔNG BẮT BUỘC</i> |
| Giới thiệu | <p>Phân cụm dữ liệu là một trong những nghiên cứu quan trọng trong khai thác dữ liệu và được áp dụng cho đa lĩnh vực.</p> <p>Mục tiêu chính trong phân cụm dữ liệu là để phân loại các đối tượng không có nhãn thành nhiều cụm mà các đối tượng thuộc cùng một cụm thì tương tự nhau và khác nhau đối với các cụm khác nhau. Phân cụm dữ liệu được chia làm hai loại là phân cụm cứng/rõ và phân cụm mềm/mờ.</p> <p>Một kỹ thuật được sử dụng phổ biến trong phân cụm dữ liệu là thuật toán K-Means, thuộc phân cụm rõ, với sự hội tụ nhanh chóng và khả năng tìm kiếm địa phương mạnh mẽ. Trong quá trình phân cụm K-Means truyền thống, các đối tượng dữ liệu thu được trong cụm là nhất định. Tuy nhiên, trong thực tế giữa những đối tượng thường không có ranh giới rõ ràng. Để tăng hiệu quả và kết quả chính xác cho phân cụm việc sử dụng lý thuyết tập thô tiếp cận hỗ trợ phân cụm K-Means được đề xuất. Mặc dù thuật toán K-Means thô có khả năng tìm kiếm địa phương mạnh mẽ nhưng lại dễ rơi vào cực trị địa phương. Do đó nghiên cứu này đề xuất hỗ trợ xác định số lượng phân cụm tối ưu theo hướng cải tiến, kết hợp <i>Phương pháp Elbow - Silhouette</i> (ELSI), các cụm sẽ được gom lại nếu các xấp xỉ trên các phân cụm giao nhau khác rỗng. Sử dụng lý thuyết tập thô để giải quyết sự thiếu chính xác và không đầy đủ tri thức, tính</p> |

chính xác cụm được cải tiến. Các kết quả thực nghiệm cho thấy rằng thuật toán phân cụm được cải tiến tốt hơn cho việc phân cụm dữ liệu thông thường.

Khả năng ứng dụng của nghiên cứu này vào thực tế rất cao, cụ thể:

1. Tổ chức hoặc cá nhân muốn chạy sự kiện khuyến mãi cho các nhóm khách hàng khác nhau dựa trên một vài thông tin mà tổ chức hoặc cá nhân đã có (năm sử dụng dịch vụ, số tiền đã chi trả, độ tuổi, giới tính, ...). Để hoạch định việc làm cách nào phân nhóm khách hàng khác nhau để thực hiện việc chạy khuyến mãi hiệu quả nhất, đảm bảo công việc kinh doanh trực tuyến phù hợp trong tình hình mới do bị tác động của dịch bệnh Covid-19.

2. Phân cụm tài liệu web:

- Tìm kiếm và rút trích tài liệu
- Tiền xử lý tài liệu: tách từ và vector hóa tài liệu, biểu diễn dưới dạng vector

3. Phân vùng ảnh.

- Phân nhóm chữ số viết tay
- Tách vật thể trong ảnh

4. Nén ảnh và nén dữ liệu.



K = 5



K = 10



K = 15



K = 20

Thuật toán K-Means:

❖ **Input:**

- k: số cụm
- X: tập dữ liệu chứa n đối tượng

❖ **Output:** tập hợp k các cụm

Bước 1: Xác định số lượng cụm k và điều kiện dừng

Bước 2: Gom các đối tượng vào cụm mà nó gần tâm nhất

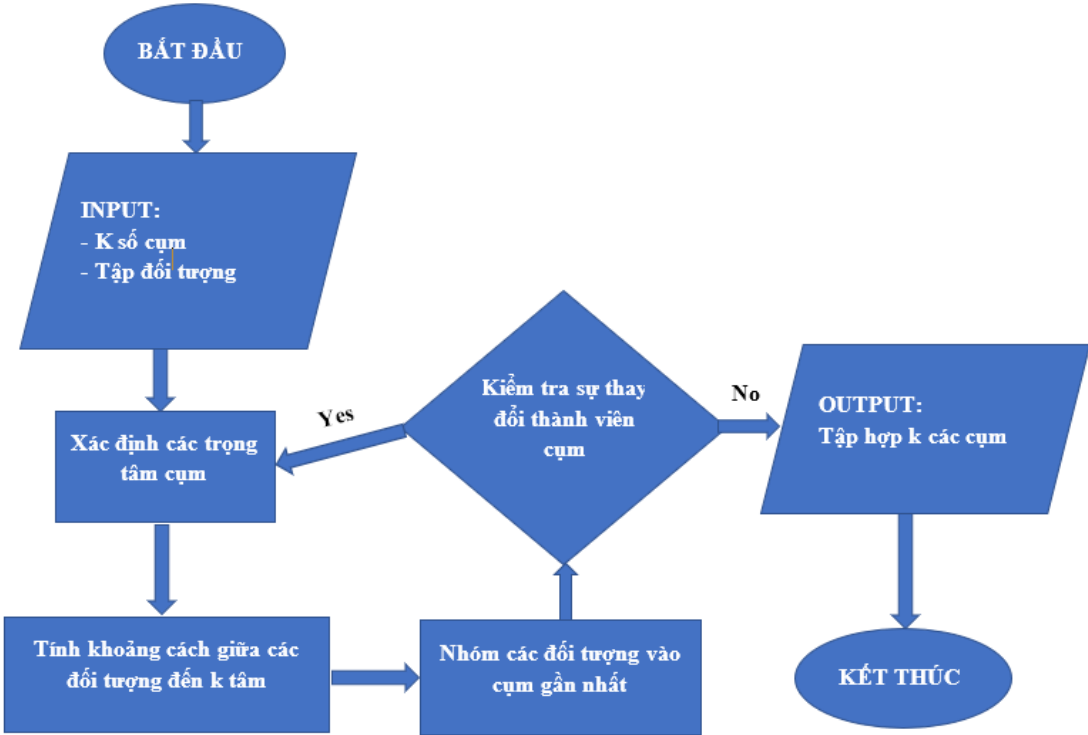
Bước 3: Tính lại các tâm theo đối tượng đã được phân hoạch ở bước 2

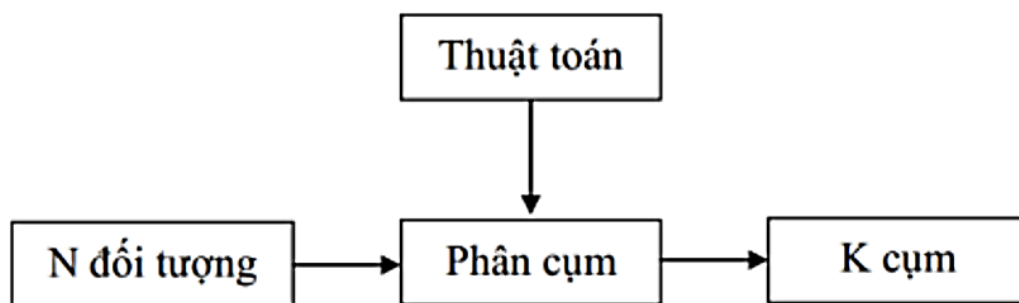
Lặp cho đến khi điều kiện dừng thỏa mãn

Điều kiện dừng thường chọn các điều kiện sau:

1. Số lần lặp $t = T_{max}$ trong đó T_{max} là số cho trước
2. $|E^t - E^{t-1}| < \Delta$ trong đó Δ là hằng số bé cho trước
3. Tới khi các cụm không đổi

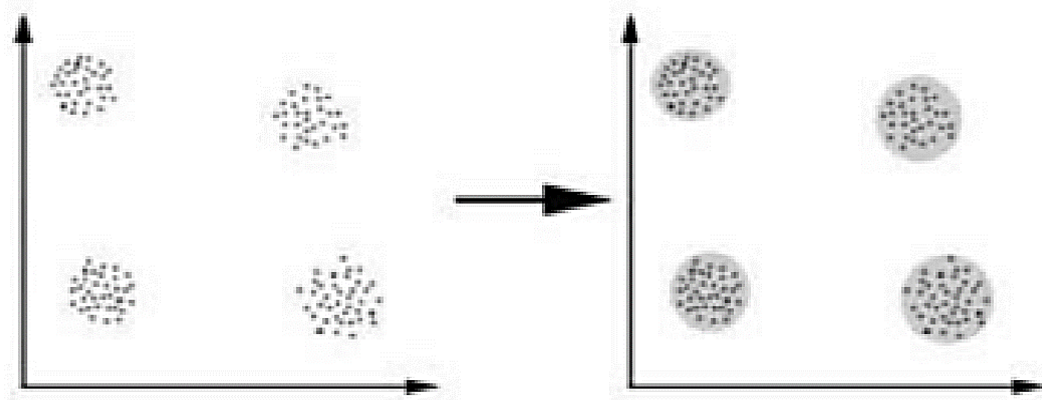
Lưu ý: khi tập dữ liệu không quá lớn thì chúng ta dùng điều kiện dừng thứ 3

| | |
|--|---|
| |  <pre> graph TD Start([BẮT ĐẦU]) --> Input[/INPUT: - K số cụm - Tập đối tượng/] Input --> Init[Xác định các trọng tâm cụm] Init --> CalcDist[Tính khoảng cách giữa các đối tượng đến k tâm] CalcDist --> Group[Nhóm các đối tượng vào cụm gần nhất] Group --> Check{Kiểm tra sự thay đổi thành viên cụm} Check -- Yes --> Init Check -- No --> Output[/OUTPUT: Tập hợp k các cụm/] Output --> End([KẾT THÚC]) </pre> |
| Mục tiêu | <p>Mục tiêu của nghiên cứu:</p> <ol style="list-style-type: none"> 1. Trình bày quá trình phân cụm theo thuật toán K-Means, sử dụng lý thuyết tập thô tiếp cận hỗ trợ phân cụm K-Means, xác định rõ số lượng phân cụm tối ưu, theo phương pháp Elbow. 2. Đề xuất thuật toán GMM thay cho thuật toán K-Means, sử dụng lý thuyết tập thô tiếp cận hỗ trợ phân cụm GMM, xác định rõ số lượng phân cụm tối ưu, theo phương pháp Elbow, đồng thời đề xuất cải tiến, kết hợp phương pháp Elbow - Silhouette (ELSI). 3. Chạy thử nghiệm, đánh giá, so sánh quá trình phân cụm theo thuật toán K-Means và GMM đối với dữ liệu tuyến tính và phi tuyến tính. |
| Nội dung và phương pháp thực hiện | <p>I. Nội dung</p> <p>Bài toán phân cụm dữ liệu là một nhánh ứng dụng chính của lĩnh vực học không giám sát, mà dữ liệu mô tả trong bài toán là không được dán nhãn. Trong trường hợp này, thuật toán sẽ tìm cách phân cụm dữ liệu thành từng nhóm có đặc điểm tương tự nhau, nhưng đồng thời đặc tính giữa các nhóm đó lại phải càng khác biệt càng tốt. Số các cụm dữ liệu có thể được xác định trước theo kinh nghiệm hoặc có thể được tự động xác định theo thuật toán.</p> |



Hình 1. Quy trình phân cụm

Độ tương tự được xác định dựa trên giá trị các thuộc tính mô tả đối tượng. Thông thường, phép đo khoảng cách thường được sử dụng để đánh giá độ tương tự hay phi tương tự. Vấn đề phân cụm có thể minh họa như hình 2:



Hình 2. Mô phỏng sự phân cụm dữ liệu

II. Các phương pháp thực hiện

1. Thuật toán K-Means

Thuật toán K-Means (MacQueen, 1967)[1, 2, 5, 7, 8] là một trong những thuật toán học không giám sát đơn giản nhất để giải quyết vấn đề phân cụm dữ liệu nổi tiếng, với số cụm được xác định trước là k cụm. Thuộc nhóm phân cụm dữ liệu cứng/rõ, ý tưởng chính là để xác định k trọng tâm cho k cụm, một trọng tâm cho mỗi cụm. Những trọng tâm nên được đặt ở vị trí thích hợp nhất vì vị trí khác nhau gây ra kết quả khác nhau. Vì vậy, sự lựa chọn tốt hơn là đặt chúng càng nhiều càng tốt và cách xa nhau. Bước tiếp theo là với mỗi điểm thuộc tập dữ liệu cho trước và liên kết nó với trọng tâm gần nhất.

2. Phương pháp phân cụm tập thô

Khái niệm về phân cụm thô tương tự như lý thuyết tập thô - với bộ xấp xỉ dưới và trên cho phép các đối tượng thuộc nhiều cụm trong tập hợp dữ liệu. Theo định nghĩa, xấp xỉ dưới của một cụm thô chứa các đối tượng mà nó chắc chắn thuộc về cụm đó, và các đối tượng

thuộc về xấp xỉ trên có thể thuộc về nhiều hơn một cụm. Đối với kỹ thuật phân cụm, lý thuyết tập thô được tiếp cận hỗ trợ phân cụm dựa vào hai hướng:

a) Cải tiến các thuật toán phân cụm cổ điển như K-Means, K-Medoid thành Rough K-Means, Rough K-Medoid... bằng cách kết hợp các khoảng cách hay độ tương đồng với các phép xấp xỉ.

b) Hỗ trợ xác định số lượng phân cụm tối ưu: Dựa trên số lượng phân cụm ban đầu theo các phương pháp Elbow, Silhouette, các cụm sẽ được gom lại nếu các xấp xỉ trên các phân cụm giao nhau khác rỗng.

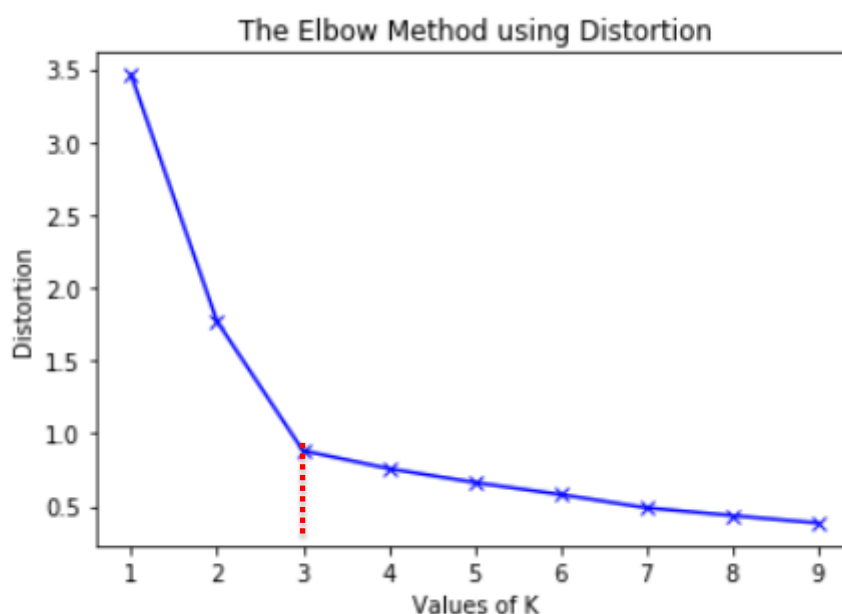
Trong phân cụm thô ta không xét tất cả các thuộc tính của tập thô. Tuy nhiên, bộ xấp xỉ trên và dưới được yêu cầu phải làm theo một số các thuộc tính tập thô cơ bản như sau:

- Một đối tượng v thuộc nhiều nhất một xấp xỉ dưới
- Một đối tượng v thuộc xấp xỉ dưới của một tập thì cũng thuộc xấp xỉ trên của nó.

Nếu một đối tượng v không thuộc bất kỳ xấp xỉ dưới thì nó thuộc hai hoặc nhiều hơn xấp xỉ trên.

3. Phương pháp Elbow

Bước cơ bản đối với bất kỳ thuật toán không được giám sát nào là xác định số lượng các cụm tối ưu mà dữ liệu có thể được gom vào. Phương pháp Elbow là một trong những phương pháp phổ biến nhất để xác định giá trị tối ưu của k cụm. Để xác định số lượng cụm tối ưu, chúng ta phải chọn giá trị của k tại “khủy tay” tức là điểm mà sau đó biến dạng hoặc quán tính bắt đầu giảm theo kiểu tuyến tính (Hình vẽ)

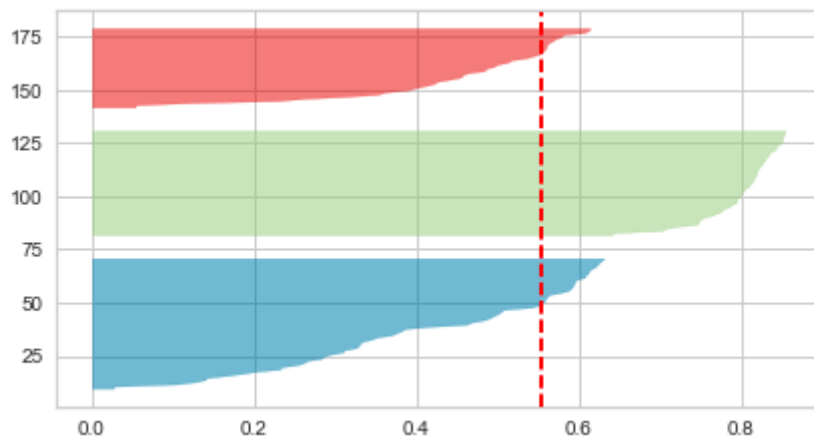


4. Phương pháp Silhouette

Phương pháp Silhouette cũng là một phương pháp để tìm số tối ưu của các cụm và giải thích và xác nhận tính nhất quán trong các cụm dữ liệu. Phương pháp Silhouette tính toán các hệ số Silhouette của mỗi điểm đo bao nhiêu điểm tương tự với cụm của chính nó so với các cụm khác, bằng cách cung cấp một biểu diễn đồ họa ngắn gọn về mức độ phân loại của từng đối tượng.

Silhouette là một thước đo về cách một thuật toán phân cụm đã hoạt động. Sau khi tính toán hệ số Silhouette của mỗi điểm trong tập dữ liệu, hãy vẽ biểu đồ đó để có được một hình ảnh trực quan về mức độ tốt của tập dữ liệu được nhóm thành k cụm. Biểu đồ Silhouette hiển thị thước đo mức độ gần của mỗi điểm trong một cụm với các điểm trong các cụm lân cận và do đó cung cấp một cách để đánh giá các thông số như số lượng cụm một cách trực quan. Số đo này có phạm vi là $[-1, 1]$.

Như đã nêu ở trên, biểu đồ Silhouette cho $k = 3$ có vẻ phù hợp nhất so với các biểu đồ khác vì nó phù hợp với tất cả ba tiêu chí đo lường (điểm dưới điểm số Silhouette trung bình, Sự dao động lớn về kích thước của đồ thị và độ dày không đồng đều).



5. Mô hình hỗn hợp Gaussian (Gaussian Mixture Model – GMM)

GMM là một hàm số được tổng hợp từ rất nhiều bộ Gaussians, được sử dụng để giải quyết các bài toán liên quan đến dữ liệu ở cùng một tập chứa các phân phối khác nhau [11, 3], mỗi phân phối được định nghĩa bởi $k \in \{1..K\}$, trong đó K là số cụm của bộ dữ liệu. Mỗi Gaussian k trong hỗn hợp này được tổng hợp từ các tham số sau:

- Giá trị trung bình μ định nghĩa trung tâm của cụm.
- Hiệp phương sai Σ định nghĩa biên của cụm.
- Giá trị xác suất α định nghĩa mức độ lớn hay nhỏ của hàm Gaussian.

GMM giả định một hỗn hợp các phân phối gaussian đã tạo ra dữ liệu. Nó sử dụng với phép gán mềm các điểm dữ liệu cho các cụm (tức là theo xác suất và do đó tốt hơn) tương

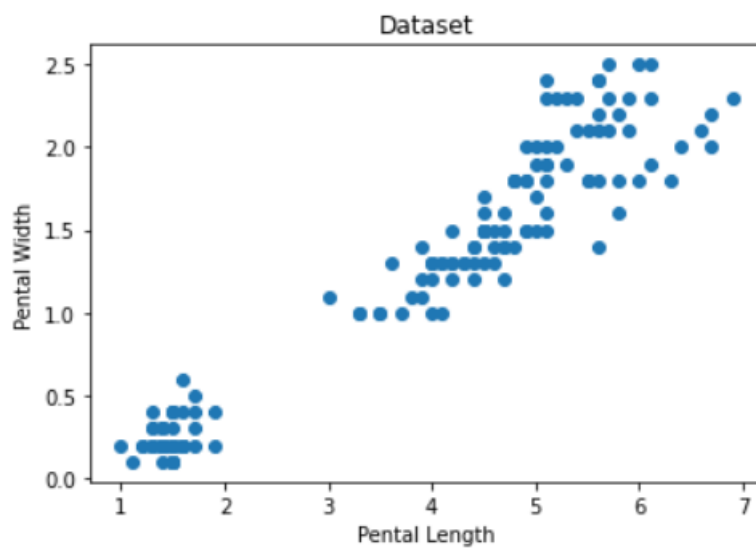
phản với phương pháp K-mean về gán cứng các điểm dữ liệu cho các cụm với giả định phân bố dữ liệu theo vòng tròn xung quanh các trung tâm.

Kết quả dự kiến

Trong nghiên cứu này phần mềm sử dụng là Jupyter Notebook 6.1.4. Sử dụng ngôn ngữ python để minh họa và demo các thuật toán: K-Means và GMM; phương pháp Elbow, Silhouette với Bộ dữ liệu: Iris.xls [12].

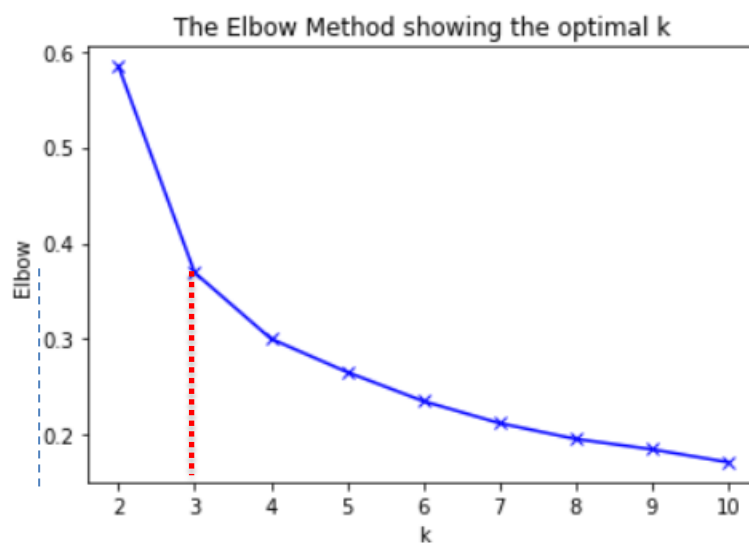
1. Về quá trình phân cụm theo thuật toán K-Means, sử dụng lý thuyết tập thô tiếp cận hỗ trợ phân cụm K-Means, xác định rõ số lượng phân cụm tối ưu, theo phương pháp Elbow.

Bước 1: Chuẩn hóa dữ liệu: X chỉ chứa 2 dữ liệu là petallength và petalwidth

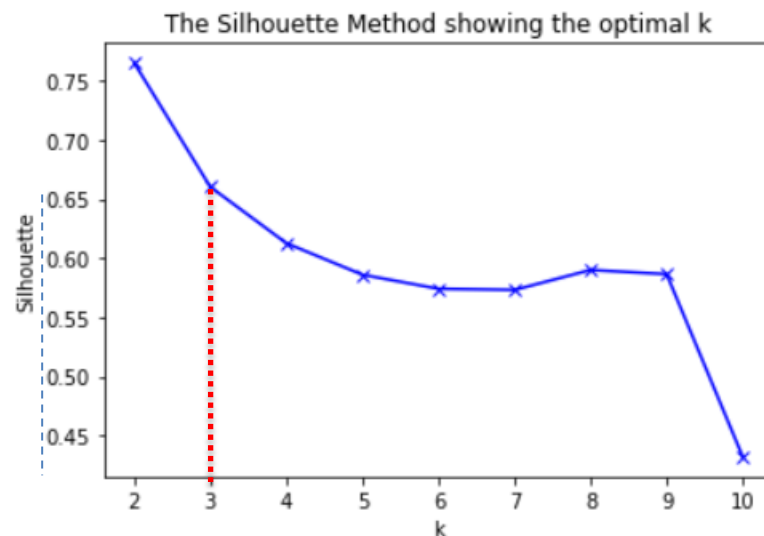


Bước 2: Áp dụng một trong hai phương pháp sau:

a) Áp dụng Elbow tìm k (phân cụm với $K > 1$)

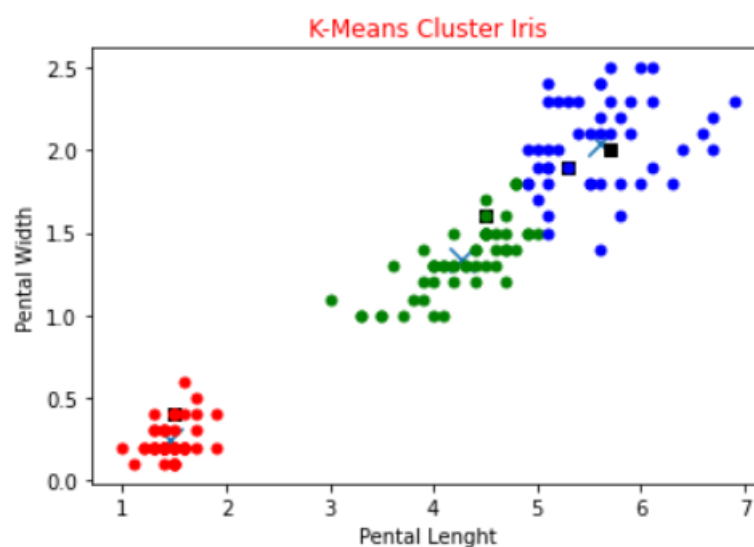


b) Áp dụng Silhouette tìm k (phân cụm với $K > 1$)



Bước 3: Áp dụng thuật toán K-Means để giải bài toán phân cụm theo K (chọn với $K=3$)

Bước 4: Vẽ hình, xem kết quả

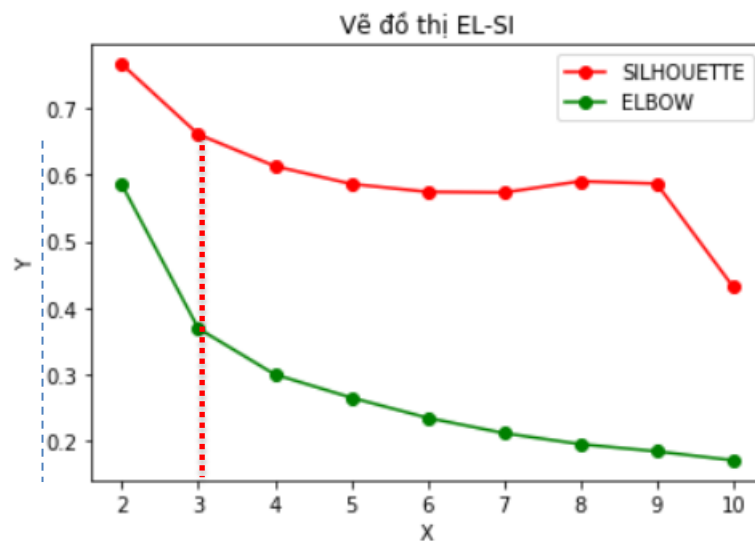


2. Về đề xuất thuật toán GMM thay cho thuật toán K-Means, sử dụng lý thuyết tập thô tiếp cận hỗ trợ phân cụm GMM, xác định rõ số lượng phân cụm tối ưu, theo phương pháp Elbow, đồng thời đề xuất cải tiến, kết hợp phương pháp Elbow - Silhouette (ELSI).

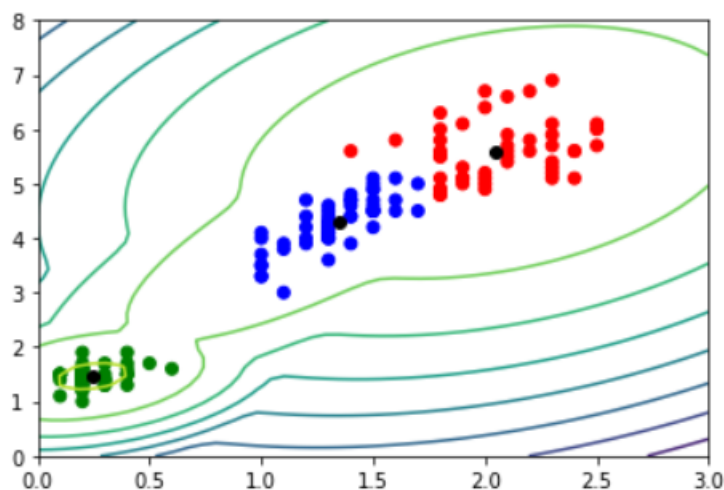
Bước 1, 2: tương tự phân trình bày quá trình phân cụm theo thuật toán K-Means và được thay bằng thuật toán GMM, xác định rõ số lượng phân cụm tối ưu, theo phương pháp Elbow hoặc Silhouette.

Bước 3: Đề xuất cải tiến, kết hợp phương pháp Elbow - Silhouette (ELSI) tìm k cụm dữ liệu tuyến tính và phi tuyến tính một cách chính xác và tối ưu nhất. Đối với cải tiến này sẽ phát

huy hiện quả tốt nhất đối với dữ liệu phi tuyến tính, nếu chỉ áp dụng phương pháp Elbow sẽ khó tìm giá trị k một cách chính xác, tối ưu.



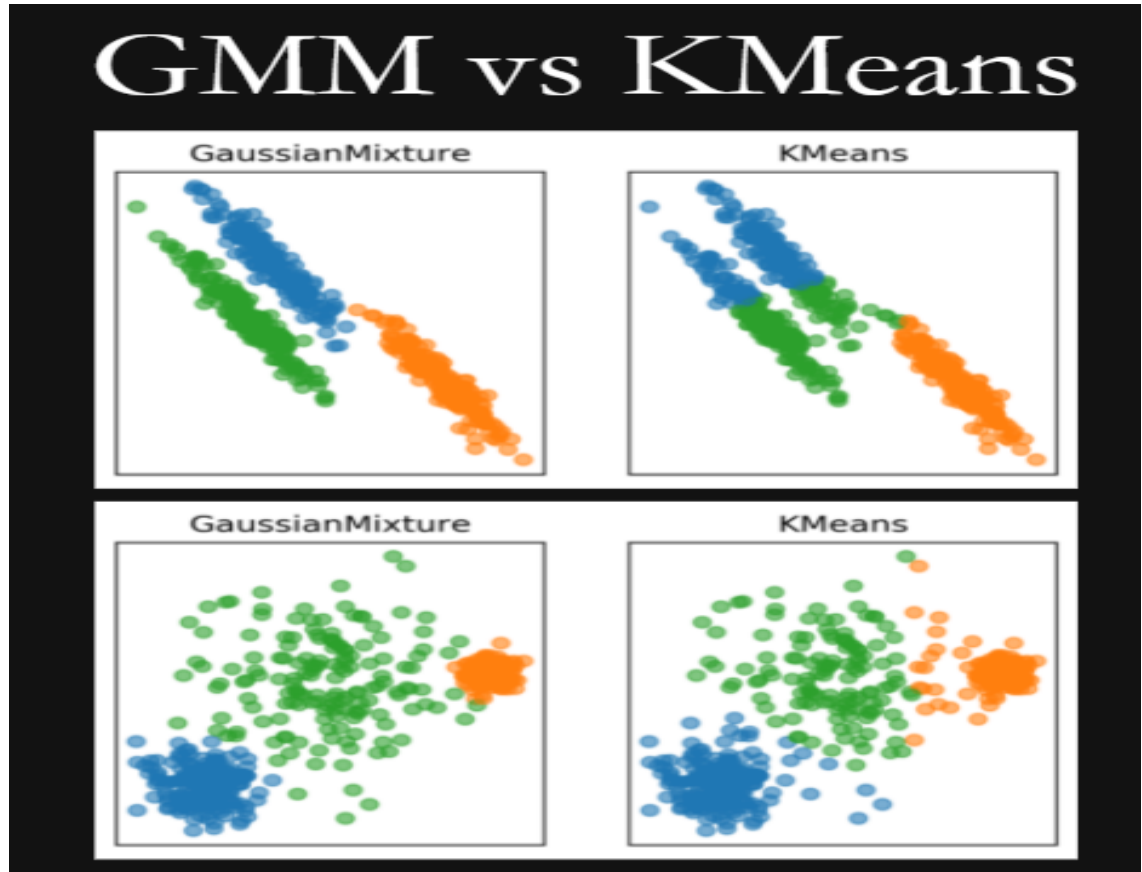
Bước 4: Áp dụng thuật toán GMM để giải bài toán phân cụm theo K (chọn với $K=3$)



3. Chạy thử nghiệm, đánh giá, so sánh quá trình phân cụm theo thuật toán K-Means và GMM đối với dữ liệu tuyến tính và phi tuyến tính.

a) Phương pháp Elbow và Silhouette được sử dụng để tìm số lượng cụm tối ưu. Sự mơ hồ nảy sinh đối với phương pháp Elbow để chọn giá trị của k khi dữ liệu phi tuyến tính. Phân tích Silhouette có thể được sử dụng để nghiên cứu khoảng cách tách biệt giữa các cụm kết quả và có thể được coi là một phương pháp tốt hơn so với phương pháp Elbow. Nhưng khi kết hợp cả hai phương pháp Elbow - Silhouette (ELSI) việc chọn giá trị của k sẽ là tối ưu nhất, cho dù dữ liệu tuyến tính hay phi tuyến tính.

b) Thuật toán K-means sử dụng hàm khoảng cách để khám phá các cụm trong dữ liệu. Cách tiếp cận này hoạt động tốt miễn là dữ liệu tuân theo phân bố vòng tròn đối với các điểm trung tâm. Nhưng nếu dữ liệu là phi tuyến tính, hình elip,... thì thuật toán K-means phân cụm chưa tối ưu.



Thuật toán GMM giả định một hỗn hợp các phân phối gaussian đã tạo ra dữ liệu. Nó sử dụng với phép gán mềm các điểm dữ liệu cho các cụm (tức là theo xác suất và do đó tốt hơn) tương phản với phương pháp K-mean về gán cứng các điểm dữ liệu cho các cụm với giả định phân bố dữ liệu theo vòng tròn xung quanh các trung tâm [9, 10]. Nói tóm lại, phương pháp phân cụm theo thuật toán GMM hoạt động tốt hơn vì nắm bắt sự không chắc chắn của các điểm dữ liệu thuộc các cụm khác nhau bằng cách sử dụng phép gán mềm và không có độ lệch cho các cụm tròn. Vì vậy, áp dụng thuật toán GMM kết hợp cải tiến, kết hợp phương pháp Elbow - Silhouette (ELSI) tìm số k cụm tối ưu hoạt động tốt ngay cả với các phân phối dữ liệu tuyến tính và phi tuyến tính.

Tài liệu tham khảo

- [1] K-Means clustering, Wikipedia, https://en.wikipedia.org/wiki/K-means_clustering.
- [2] K-Mean Clustering Algorithm Approach for Data Mining of Heterogeneous Data, Monika Kalra – Niranjan Lal, 2018,
https://www.researchgate.net/publication/320932435_KMean_Clustering_Algorithm_Approach_for_Data_Mining_of_Heterogeneous_Data.

- [3] Mixture model, Wikipedia, https://en.wikipedia.org/wiki/Mixture_model
- [4] Advances in K-means Clustering A Data Mining Thinking, Junjie Wu, 2012, Springer Theses.
- [5] K-Means clustering, 2017, <https://machinelearningcoban.com/>
- [6] Hoàng Xuân Huân (2012), “Giáo trình Nhận dạng mẫu”, Trường Đại học công nghệ – Đại Học Quốc Gia Hà Nội.
- [7] Xây Dựng Clustering Model Bằng Giải Thuật K-Means Với Thư Viện Scikit-Learn Skills AI, <https://insights.magestore.com/posts/xay-dung-clustering-model-bang-giai-thuat-k-means-voi-thu-vien-scikit-learn-skills-ai>
- [8] K-means Clustering, <https://machinelearningcoban.com/2017/01/01/kmeans>.
- [9] Sounak Bhattacharya and Ankit Lundia. “Movie Recommendation System Using Bag Of Words and Scikit-learn”. In: International Journal of Engineering Applied Sciences and Technology 04 (Oct. 2019), pp. 526–528. DOI: 10.33564/IJEAST.2019.v04i05.076.
- [10] Dilan G˘or˘ur and Carl Rasmussen. “Dirich-let Process Gaussian Mixture Models: Choice of the Base Distribution”. In: J. Comput. Sci. Technol. 25 (July 2010), pp. 653–664. DOI: 10.1007/s11390-010- 9355-8.
- [11] Carl Rasmussen. “The Infinite Gaussian Mixture Model”. In: vol. 12. Apr. 2000, pp. 554–560.
- [12] Iris flower data set, Wikipedia, https://en.wikipedia.org/wiki/Iris_flower_data_set