1. Định nghĩa

Ảnh có chứa biểu đồ

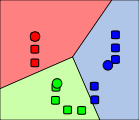
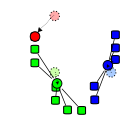
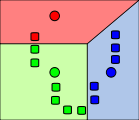
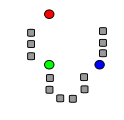
Mô tả được tạo tự động

K-means là một kỹ thuật học máy không giám sát được sử dụng để phân vùng tập dữ liệu thành K nhóm hoặc cụm riêng biệt dựa trên sự giống nhau của các điểm dữ liệu. Mục tiêu chính của phân cụm K-means là giảm thiểu phương sai bên trong cụm đồng thời tối đa hóa phương sai giữa các cụm.

Trong phân cụm K-means, độ tương tự giữa các điểm dữ liệu thường được xác định bằng thước đo khoảng cách, chẳng hạn như khoảng cách Euclide. Thuật toán xác định K trọng tâm, là tâm của cụm và tinh chỉnh lặp đi lặp lại vị trí của chúng để giảm thiểu tổng bình phương khoảng cách giữa các điểm dữ liệu và trọng tâm cụm tương ứng của chúng.

Tóm lại, phân cụm K-means là một phương pháp được sử dụng để nhóm các điểm dữ liệu tương tự lại với nhau dựa trên các tính năng hoặc thuộc tính của chúng, nhằm tối ưu hóa việc phân phối tổng thể các điểm dữ liệu trong và giữa các cụm. Nó được sử dụng rộng rãi trong các ứng dụng khác nhau, bao gồm phân đoạn hình ảnh, phân đoạn khách hàng, phát hiện bất thường và phân cụm tài liệu.

1. Cách thức hoạt động



Thuật toán phân cụm K-means hoạt động thông qua quy trình lặp để nhóm các điểm dữ liệu tương tự thành K cụm dựa trên các chức năng hoặc thuộc tính của chúng. Tổng quan từng bước về cách thức hoạt động của thuật toán K-means:

Khởi tạo: Quyết định số cụm (K) bạn muốn tạo. Chọn ngẫu nhiên K điểm dữ liệu từ tập dữ liệu làm trọng tâm cụm ban đầu (trung tâm).

Nhiệm vụ: Tính khoảng cách giữa mỗi điểm dữ liệu và tất cả các trọng tâm. Chỉ định từng điểm dữ liệu cho tâm gần nó nhất, dựa trên thước đo khoảng cách, chẳng hạn như khoảng cách Euclide. Điều này tạo thành các cụm ban đầu.

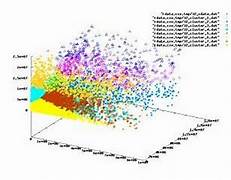
Cập nhật: Đối với mỗi cụm, hãy tính giá trị trung bình của tất cả các điểm dữ liệu trong cụm đó. Cập nhật trọng tâm thành giá trị trung bình mới này.

Lặp lại: Tiếp tục lặp lại các bước 2 (Chỉ định) và 3 (Cập nhật) cho đến khi vị trí trọng tâm không thay đổi đáng kể giữa các lần lặp lại hoặc đạt đến số lần lặp lại tối đa được xác định trước.

Thuật toán hội tụ khi sự thay đổi vị trí trọng tâm giữa các lần lặp liên tiếp giảm xuống dưới ngưỡng được xác định trước hoặc khi đạt đến số lần lặp tối đa. Tại thời điểm này, thuật toán (lý tưởng nhất) đã tìm thấy phân cụm tối ưu của tập dữ liệu, với mỗi điểm dữ liệu được gán cho trọng tâm gần nhất của nó.

Điều quan trọng cần lưu ý là thuật toán K-means nhạy cảm với lựa chọn trọng tâm ban đầu và các lần khởi tạo khác nhau có thể dẫn đến các cụm cuối cùng khác nhau. Để giảm thiểu vấn đề này, bạn có thể chạy thuật toán nhiều lần với các lần khởi tạo khác nhau và chọn kết quả phân cụm có tổng bình phương khoảng cách thấp nhất giữa các điểm dữ liệu và trọng tâm tương ứng của chúng. Ngoài ra, bạn có thể sử dụng các kỹ thuật khởi tạo cải tiến như K-means++ để chọn các trọng tâm ban đầu tốt hơn.

1. Các ứng dụng của thuật toán K-means



K-means là một thuật toán linh hoạt với nhiều ứng dụng trên nhiều miền khác nhau. Một số ứng dụng phổ biến của phân cụm K-means bao gồm:

* Phân khúc khách hàng: Các doanh nghiệp sử dụng phân cụm K-means để phân tích dữ liệu khách hàng và xác định các phân khúc khách hàng khác nhau dựa trên hành vi, sở thích hoặc nhân khẩu học của họ. Điều này giúp tiếp thị có mục tiêu, cải thiện sự hài lòng của khách hàng và cá nhân hóa các đề xuất sản phẩm.
* Phân đoạn và nén hình ảnh: Có thể áp dụng phân cụm K-means để phân vùng hình ảnh thành các vùng khác nhau dựa trên màu hoặc cường độ pixel, dẫn đến phân đoạn hình ảnh. Nó cũng có thể được sử dụng để giảm không gian màu cho quá trình nén hình ảnh bằng cách biểu thị từng pixel bằng màu trung tâm gần nhất của nó, do đó giảm số lượng màu cần thiết để biểu thị hình ảnh.
* Phát hiện bất thường: K-means phân cụm có thể được sử dụng để xác định các điểm dữ liệu ngoại lai hoặc bất thường không thuộc về bất kỳ cụm nào. Những ngoại lệ này có thể đại diện cho các giao dịch gian lận, xâm nhập mạng hoặc sản phẩm bị lỗi, tùy thuộc vào miền.
* Phân cụm tài liệu: Trong khai thác văn bản, phân cụm K-means có thể được sử dụng để nhóm các tài liệu hoặc bài viết tương tự dựa trên nội dung của chúng. Điều này giúp tổ chức các bộ sưu tập lớn văn bản, khám phá các chủ đề liên quan và cải thiện việc truy xuất thông tin.
* Giảm kích thước: Có thể kết hợp phân cụm K-means với các kỹ thuật như Phân tích thành phần chính (PCA) để giảm kích thước của dữ liệu nhiều chiều trong khi vẫn giữ được cấu trúc thiết yếu của dữ liệu.
* Phân tích dữ liệu không gian địa lý: Có thể sử dụng phân cụm K-means để phân tích dữ liệu không gian địa lý, chẳng hạn như xác định các khu vực sử dụng đất tương tự hoặc nhóm các vị trí dựa trên các thuộc tính khác nhau như nhân khẩu học, tỷ lệ tội phạm hoặc điểm ưa thích.
* Sinh học và di truyền học: Trong tin sinh học, phân cụm K-means được áp dụng để phân tích dữ liệu biểu hiện gen và xác định các nhóm gen đồng biểu hiện, có thể cung cấp thông tin chuyên sâu về các quá trình tế bào hoặc cơ chế bệnh tật.

Đây chỉ là một vài ví dụ về nhiều ứng dụng có thể có của phân cụm K-means. Do tính đơn giản, khả năng mở rộng và dễ triển khai, phân cụm K-means được sử dụng rộng rãi trong các lĩnh vực khác nhau để phân tích dữ liệu khám phá, nhận dạng mẫu và trích xuất tính năng.

1. Nhược điểm

Mặc dù phổ biến và được sử dụng rộng rãi, K-means phân cụm có một số nhược điểm:

Lựa chọn K: Việc xác định số cụm (K) tối ưu có thể là một thách thức. Chọn sai giá trị cho K có thể dẫn đến kết quả phân cụm kém. Các kỹ thuật như phương pháp khuỷu tay và phân tích hình bóng có thể hữu ích, nhưng chúng không phải lúc nào cũng dứt khoát trong việc tìm ra K tốt nhất.

Lựa chọn trọng tâm ban đầu: K-mean nhạy cảm với các vị trí trọng tâm ban đầu. Các khởi tạo khác nhau có thể dẫn đến các cụm cuối cùng khác nhau. Để giảm thiểu vấn đề này, bạn có thể chạy thuật toán nhiều lần với các lần khởi tạo khác nhau hoặc sử dụng các kỹ thuật khởi tạo cải tiến như K-means++.

Cực tiểu cục bộ: Thuật toán có thể hội tụ đến cực tiểu cục bộ thay vì cực tiểu toàn cầu, dẫn đến việc phân cụm dưới mức tối ưu. Chạy thuật toán nhiều lần với các lần khởi tạo khác nhau có thể giúp giải quyết vấn đề này.

Giả định về kích thước và mật độ cụm bằng nhau: K-means giả định rằng các cụm có kích thước, hình dạng và mật độ tương tự nhau. Nó hoạt động tốt nhất với các cụm hình cầu và có kích thước đồng đều. Khi xử lý các cụm có hình dạng bất thường hoặc có kích thước khác nhau, K-means có thể không hoạt động tốt.

Nhạy cảm với các ngoại lệ: Các ngoại lệ trong tập dữ liệu có thể tác động đáng kể đến vị trí của trọng tâm, dẫn đến kết quả phân cụm kém. Xử lý trước dữ liệu để loại bỏ hoặc xử lý các giá trị ngoại lệ có thể giúp cải thiện hiệu suất của thuật toán.

Khả năng phân tách tuyến tính: K-means có thể không hoạt động tốt với dữ liệu có thể phân tách phi tuyến tính. Các kỹ thuật phân cụm khác, chẳng hạn như DBSCAN hoặc phân cụm theo cấp bậc, có thể phù hợp hơn cho những trường hợp như vậy.

Thiếu hỗ trợ dữ liệu phân loại: K-means dựa vào số liệu khoảng cách, thường không phù hợp với dữ liệu phân loại. Các kỹ thuật phân cụm khác, chẳng hạn như chế độ K hoặc nguyên mẫu K, được thiết kế đặc biệt để xử lý dữ liệu phân loại.

Khả năng mở rộng: Mặc dù K-means có thể xử lý các tập dữ liệu lớn, nhưng nó có thể trở nên tốn kém về mặt tính toán đối với các tập dữ liệu rất lớn hoặc dữ liệu nhiều chiều. Các kỹ thuật như Mini-Batch K-means hoặc sử dụng phép tính gần đúng có thể giúp tăng tốc thuật toán.

Bất chấp những nhược điểm này, phân cụm K-means vẫn là một lựa chọn phổ biến cho nhiều tác vụ phân cụm do tính đơn giản, dễ triển khai và dễ hiểu của nó.

Ví dụ về thuật toán K-Means

thuật toán K-means sử dụng để phân nhóm khách hàng của một cửa hàng bán xe

Phương pháp phân loại khách hàng không còn là một phương pháp mới lạ trong kinh doanh, hầu hết các doanh nghiệp hiện nay đều sử dụng phương pháp này để tiếp cận khách hàng một cách hiệu quả. RFM là một trong những mô hình được các công ty sử dụng rộng rãi **RFM (Recency – Frequency – M (Monetary Value):**là một phần của Marketing Analysis và được sử dụng để phân tích giá trị khách hàng (Customer Value), từ đó giúp doanh nghiệp có thể phân tích ra từng nhóm khách hàng mà mình đang có, từ đó có những chiến dịch marketing hoặc chăm sóc đặc biệt.mô hình này giúp các công ty biết khách hàng trung thành của họ, những khách hàng không hài lòng với dịch vụ và sản phẩm của họ và những khách hàng cũ đã rời bỏ họ.

Tuy nhiên việc áp dụng mô hình RFM vào lĩnh vực bán lẻ sẽ tồn tại nhiều đặc điểm riêng biệt cần được nghiên cứu. Kết hợp mô hình RFM với thuật toán học máy K-means được mong đợi mang lại hiệu quả lớn cho các doanh nghiệp trong việc phân loại khách hàng. Vấn đề lớn nhất của ngành bán lẻ xe đạp là các nhà bán lẻ vẫn còn sử dụng mô hình phân khúc khách hàng truyền thống, vì vậy họ gặp khó khăn trong việc đáp ứng nhu cầu khách hàng và tìm ra nguồn khách hàng tiềm năng. Đối với ngành bán lẻ, việc sử dụng mô hình RFM cùng thuật toán K-means được kỳ vọng sẽ cung cấp sự hiểu biết khách hàng tốt hơn nhằm đưa ra quyết định hiệu quả hơn. Ví dụ này sẽ hướng đến việc tạo lập mô hình phân khúc khách hàng hiệu quả kết hợp phương pháp học máy cho lĩnh vực bán lẻ xe đạp của 1 công ty .

Phân loại khách hàng dự trên 1 số tiêu chí như dự trên

1.dữ liệu nhân khẩu học: giới tính độ tuổi, gia đình, qui mô hộ gia đình

2. dữ liệu địa lý: khu vực cư trú, nơi làm việc

3. dữ liệu tâm lý học: tầng lớp xã hội, lối sống, phong cách, nhân cách

4. dữ liệu bán hàng: cho biết hành vi mua sắm, số lượng bán hàng, số lượt truy cập, tần suất truy cập, lượt truy cập gần nhất

5. dữ liệu hành vi(hàng vi khác ngoài mua sắm): dữ liệu thu được từ giỏ hàng sử dụng RFID để biết những gì khách hàng đã bỏ vào giỏ hàng của họ

Việt phân loại khách hàng cũng có thể biểu diễn với nhiều mô hình khai thác dữ liệu như:

1 dựa trên mô hình liên kết. vd qui tắc liên kết, chuỗi Markov

2 dựa trên mô hình phân loại( cây quyết định (decision tree)), phân cụm

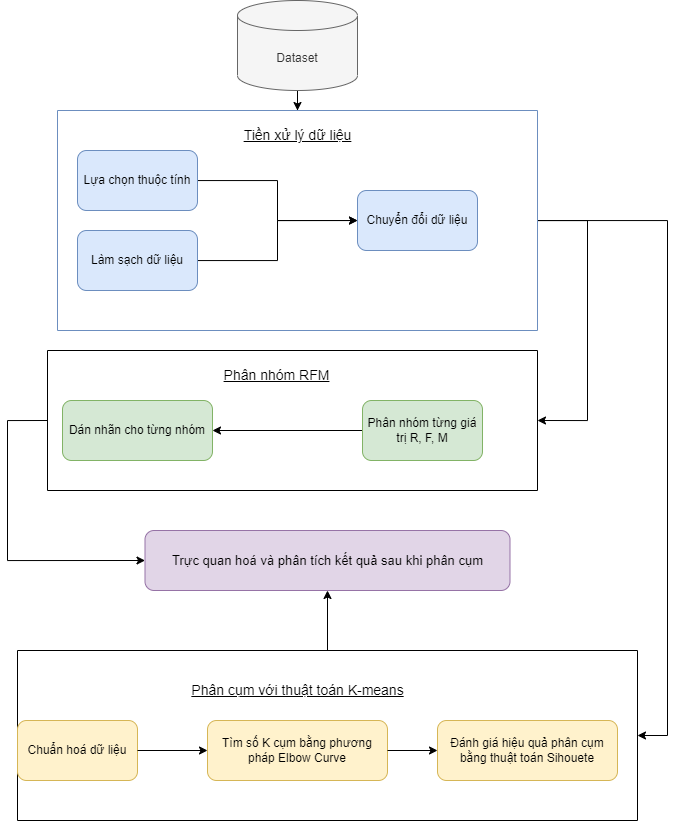
3 Dự đoán và khám phá(Neural Network)

Sau khi áp dụng các phương pháp phân loại khách hàng, các nhóm khách hàng sẽ được gán nhãn tương ứng với những đặc trưng của nó( khách hàng tiềm năng, khách hàng trung thành, khách hàng vip, khách hàng có khả năng rời bỏ). Sau khi phân loại khách hàng riêng biệt, doanh nghiệp có thể tùy chỉnh kế hoạch tiếp thị, xác định xu hướng, lập kế hoạch phát triển sản phẩm, chiến dịch quảng cáo và cung cấp các sản phẩm phù hợp với từng phân khúc khách hàng tương ứng từ đó nâng cao doanh số cho doanh nghiệp.

Quy trình quy trình thuật toán với 4 giai đoạn chính như sau:

1. Giai đoạn 1 thu thập dữ liệu cho việc tính toán;
2. Giai đoạn 2 là giai đoạn tiền xử lý dữ liệu, đây là giai đoạn có nhiều bước để chọn ra các thuộc tính phù hợp cho việc xử lý (số lần mua hàng, khối lượng đơn hàng, lần mua hàng gần nhất
3. Giai đoạn 3 thiết lập các giá trị RFM, từ bộ dữ liệu đã được làm sạch và chuẩn hoá ở Giai đoạn 2, nghiên cứu tiến hành chấm điểm RFM của từng khách hàng thông qua Lần mua hàng gần nhất, Số lần mua hàng, Tổng số tiền.với số lượng nhóm khách hàng là 5

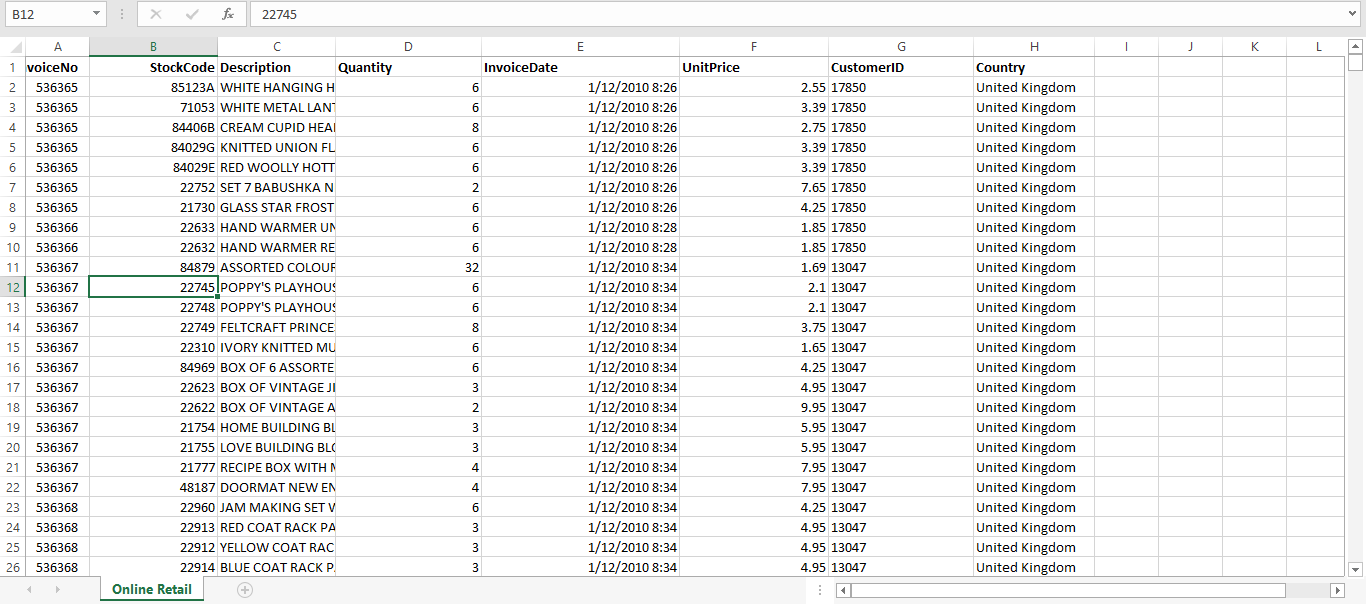
4) Giai đoạn 4 Dán nhãn và phân cụm khách hàng, từ bộ điểm RFM của từng khách hàng nghiên cứu tiếp tục phân loại khách hàng dựa trên điểm số của từng người thông qua 2 biến chính là R và F, từ đó rút ra được những kết luận, giải pháp cho từng nhóm khách hàng.



Chi tiết cách làm ví dụ:

1.1 Thu thập dữ liệu

Sử dụng bộ dữ liệu về mua bán xe.



* 1. Tiền xử lý dữ liệu

Lựa chọn các thuộc tính

Nguyên cứu và tiến hành quan sát, sàng lọc để tìm ra các thuộc tính cần thiết trong bộ dữ liệu. Tập trung khai thác những dữ liệu như Customer, InvoiceDate

Làm sạch dữ liệu

Sửa hoặc xoá dữ liệu không chính xác, dữ liệu sai, trùng lặp hoặc không đầy đủ trong tập dữ liệu

Biến đổi dữ liệu

Chuyển hoá dữ liệu là quá trình biến đổi hình thức, cấu trúc hoặc giá trị của dữ liệu nhằm tổ chức dữ liệu nhằm tổ chức dữ liệu 1 cách trực quan hiệu quả hơn. Tạo nên 1 tập dữ liệu dễ dàng sử dụng cho cả người và máy tính có đủ thuộc tính dữ liệu cần thiết cho việc phân tích.

* 1. Tính RFM tổng thể

R Lần gần đây, F Tần suất, M Giá trị

Phân cấp khách hàng bằng phương pháp phân tổ dữ liệu, số cụm được chọn là 5 để thuận tiện cho việc đánh giá.

Các mốc 20%, 40%, 60%, 80%,100%

Sau đó dán nhãn để phân biệt nhóm khách hàng khác nhau.

* 1. Sử dụng thuật toán K-Means phân loại khách hàng
* Lựa chọn số lượng nhóm (clusters): Lựa chọn số lượng nhóm (clusters) cần phân loại khách hàng. Số lượng nhóm này phải được lựa chọn sao cho phù hợp với mục đích phân loại và tối ưu hóa hiệu suất của mô hình. Dùng phương pháp Elbow Curve để xác định số lượng clusters tối ưu bằng cách tính toán tổng bình phương khoảng cách giữa các điểm dữ liệu và centroid của mỗi cluster, và sau đó biểu diễn các giá trị này trên đồ thị.
* Thực hiện thuật toán k-means: Thực hiện thuật toán k-means trên ma trận RFM với số lượng nhóm được lựa chọn. Kết quả của k-means sẽ là các nhóm khách hàng được phân loại dựa trên RFM của họ. Sử dụng phương pháp Silhouette để đánh giá sự tách biệt của các cluster bằng cách tính toán giá trị silhouette cho mỗi điểm dữ liệu. Giá trị này thể hiện mức độ tách biệt của mỗi điểm dữ liệu giữa cluster của nó so với các cluster khác.
* Phân tích và đánh giá kết quả: Phân tích và đánh giá kết quả phân loại khách hàng dựa trên các nhóm được tạo ra.

**So sánh thuật toán k-mean với các thuật toán khác:**

Kmeans là 1 thuật toán phân cụm phổ biến nhất. Trong đó còn các thuật toán phân cụm khác như:

Hierarchical clustering: phân cụm dữ liệu theo cấp độ, bắt đầu từ việc phân chia thành các cluster con nhỏ và sau đó kết hợp các cluster con này thành các cluster lớn hơn.

DBSCAN: phân cụm dữ liệu bằng cách phát hiện các khu vực có mật độ cao của điểm dữ liệu.

Mean shift clustering: tìm kiếm các điểm cực đại của hàm mật độ của dữ liệu, đại diện cho các trung tâm của các cluster, và gom các điểm dữ liệu vào cùng một cluster.

Spectral clustering: sử dụng đặc trưng của dữ liệu để phân chia chúng thành các cluster bằng cách tách dữ liệu thành các thành phần riêng lẻ và sử dụng các giá trị riêng và vector riêng của ma trận đặc trưng để xác định các cluster.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Thuật toán** | **Độ chính xác** | **Tốc độ** | **Khả năng mở rộng** | **Khả năng giải thích** | **Khả năng chống nhiễu** | **Tính linh hoạt** |
| K-Means | Trung bình | Nhanh | Tốt | Kém | Kém | Thấp |
| Hierarchical | Cao | Chậm | Tệ | Tốt | Tốt | Cao |
| DBSCAN | Cao | Nhanh | Tệ | Trung bình | Tốt | Cao |
| Mean shift | Cao | Chậm | Tệ | Kém | Tốt | Cao |
| Spectral clustering | Cao | Chậm | Kém | Kém | Kém | Thấp |

K-Means: Đơn giản, tốc độ nhanh nhưng kém chính xác, khó giải thích kết quả và không chống nhiễu tốt.

Hierarchical: Chính xác và dễ giải thích nhưng chậm và khó mở rộng cho lớn dữ liệu.

K-means

 Dễ giải thích kết quả hơn - Không cần xác định trước số cluster - Thuật toán phức tạp và tốc độ chậm hơn. Kém linh hoạt với dữ liệu nhiều chiều

DBSCAN: Chính xác, chống nhiễu tốt, linh hoạt nhưng tốc độ trung bình và khó mở rộng.

K-means  Tốt hơn trong việc loại bỏ nhiễu

 - Không yêu cầu xác định số cluster trước

 - Linh hoạt và phù hợp với dữ liệu nhiều hình thái

- Thuật toán phức tạp và tốc độ chậm hơn

 - Khó thiết lập các tham số

Mean shift: Chính xác, chống nhiễu tốt nhưng chậm, khó giải thích và khó mở rộng.

K-means: - Tốt hơn trong việc loại bỏ nhiễu

- Không yêu cầu xác định số cluster trước

- Linh hoạt và phù hợp với dữ liệu nhiều hình thái

- Thuật toán phức tạp và tốc độ rất chậm

- Khó thiết lập các tham số và khó giải thích kết quả

Spectral clustering: Chính xác cao nhưng chậm, khó giải thích, không chống nhiễu tốt và kém linh hoạt.

K-means: - Độ chính xác cao hơn

- Không yêu cầu xác định số cluster trước

- Thuật toán rất phức tạp, tốc độ rất chậm

- Khó giải thích kết quả và kém chống nhiễu