HỌC VIỆN NÔNG NGHIỆP VIỆT NAM

**KHOA CÔNG NGHỆ THÔNG TIN**



**KHÓA LUẬN TỐT NGHIỆP**

ĐỀ TÀI: TÌM HIỂU THUẬT TOÁN K-MEANS VÀ ỨNG DỤNG

**HÀ NỘI – 2021**

HỌC VIỆN NÔNG NGHIỆP VIỆT NAM

**KHOA CÔNG NGHỆ THÔNG TIN**



**KHÓA LUẬN TỐT NGHIỆP**

ĐỀ TÀI: TÌM HIỂU THUẬT TOÁN K-MEANS VÀ ỨNG DỤNG

**NGƯỜI THỰC HIỆN: DƯƠNG BÍCH THỦY**

**KHÓA: 61**

**NGÀNH: CÔNG NGHỆ THÔNG TIN**

**CHUYÊN NGÀNH: TIN HỌC**

**NGƯỜI HƯỚNG DẪN: THS. NGUYỄN XUÂN THẢO**

**HÀ NỘI – 2021**

Mục lục

[LỜI CẢM ƠN 5](#_Toc90932646)

[Bảng từ viết tắt và thuật ngữ 6](#_Toc90932647)

[Danh mục hình vẽ 7](#_Toc90932648)

[Danh mục bảng biểu 7](#_Toc90932649)

[Phần I: Mở đầu 8](#_Toc90932650)

[1.1. Đặt vấn đề 8](#_Toc90932651)

[1.2. Mục đích và yêu cầu 8](#_Toc90932652)

[1.2.1. Mục đích 8](#_Toc90932653)

[1.2.2. Yêu cầu 9](#_Toc90932654)

[Phần II: Tổng quan tình hình nghiên cứu trong và ngoài nước 9](#_Toc90932655)

[2.1. Tình hình nghiên cứu trong nước 9](#_Toc90932656)

[2.2. Tình hình nghiên cứu ở nước ngoài 11](#_Toc90932657)

[Phần III: Nội dung và phương pháp nghiên cứu 12](#_Toc90932658)

[3.1. Địa điểm và thời gian nghiên cứu (thực tập) 12](#_Toc90932659)

[3.2. Nội dung nghiên cứu 12](#_Toc90932660)

[3.3. Phương pháp nghiên cứu: 13](#_Toc90932661)

[Phần IV: Kết quả nghiên cứu và thảo luận 13](#_Toc90932662)

[4.1. Tổng quan về phân cụm dữ liệu 13](#_Toc90932663)

[4.1.1. Khái niệm chung 13](#_Toc90932664)

[4.1.2. Phân cụm dữ liệu 14](#_Toc90932665)

[4.1.3. Một số kỹ thuật tiếp cận trong phân cụm dữ liệu 20](#_Toc90932666)

[4.1.4. Các vấn đề thường gặp khi phân cụm 20](#_Toc90932667)

[4.2. Thuật toán K-Means trong bài toán phân cụm dữ liệu 21](#_Toc90932668)

[4.2.1. Giới thiệu chung 21](#_Toc90932669)

[4.2.2. Một số ứng dụng 22](#_Toc90932670)

[4.2.3. Thuật toán K-Means 23](#_Toc90932671)

[4.2.4. Một số phương pháp lựa chọn số cụm k 30](#_Toc90932672)

[4.3. Ứng dụng K-Means vào bài toán phân khúc khách hàng 33](#_Toc90932673)

[4.3.1. Bài toán phân khúc khách hàng 33](#_Toc90932674)

[4.3.2. Dữ liệu nghiên cứu 34](#_Toc90932675)

[4.3.2. Các công nghệ sử dụng trong nghiên cứu bài toán 35](#_Toc90932676)

[4.3.3. Kết quả nghiên cứu 35](#_Toc90932677)

[Phần V: Kết luận và đề nghị 50](#_Toc90932678)

[5.1. Kết luận 50](#_Toc90932679)

[5.2. Đề nghị 50](#_Toc90932680)

[Phần VI. Tài liệu tham khảo 52](#_Toc90932681)

# LỜI CẢM ƠN

Trước tiên với tình cảm sâu sắc và chân thành nhất, em xin bày tỏ lòng biết ơn đến tất cả các cá nhân và tổ chức đã tạo điều kiện hỗ trợ, giúp đỡ em trong suốt quá trình nghiên cứu và hoàn thành báo cáo thực tập tốt nghiệp.

Đặc biệt, em xin chân thành cảm ơn ThS. Nguyễn Xuân Thảo đã tận tình hướng dẫn, cung cấp tài liệu, định hướng, truyền đạt nội dung nghiên cứu một cách dễ hiểu, giúp em có thể hoàn thành báo cáo một cách hoàn thiện nhất.

Đồng thời cũng xin gửi lời cảm ơn Ban lãnh đạo Học viện Nông nghiệp Việt Nam, các thầy, cô giáo khoa Công nghệ thông tin cũng như bộ môn Mạng và Hệ thống thông tin đã tạo điều kiện thuận lợi để em hoàn thành tốt công việc nghiên cứu khoa học của mình.

Do điều kiện thời gian không cho phép cũng như những hạn chế về kiến thức, báo cáo thực tập tốt nghiệp này không tránh khỏi những sơ suất và thiếu sót. Kính mong nhận được sự chỉ bảo, đóng góp ý kiến từ các quý thầy cô để em có thể hoàn thiện báo cáo một cách trọn vẹn và hoàn thiện bản thân, qua đó nâng cao năng lực và kiến thức để phục vụ công việc sau này.

Dương Bích Thủy

# Bảng từ viết tắt và thuật ngữ

|  |  |  |  |
| --- | --- | --- | --- |
| **STT** | **Viết tắt** | **Thuật ngữ tiếng Anh** | **Giải ngữ** |
| 1 |  | Categorical | Phân loại (kiểu dữ liệu) |
| 2 |  | Centroid | Trọng tâm (cụm) |
| 3 |  | Classification | Phân lớp |
| 4 |  | Cluster | Cụm |
| 5 |  | Clustering | Phân cụm |
| 6 |  | Dimensionality Reduction | Giảm chiều dữ liệu |
| 7 |  | Distance Measures | Độ đo khoảng cách |
| 8 |  | Feature | Đặc trưng; một cột trong bảng dữ liệu |
| 9 | init | Initiation | Khởi tạo |
| 10 |  | Normalization | Phương pháp biến đổi làm thay đổi hình dạng phân phối của dữ liệu |
| 11 |  | Numerical | Số (kiểu dữ liệu) |
| 12 |  | Outliers | Bất thường, phần tử ngoại lai |
| 13 |  | Outliers Detection | Phát hiện bất thường |
| 14 |  | Preprocessing | Tiền xử lý |
| 15 |  | Regression | Hồi quy |
| 16 |  | Scale | Phương pháp thay đổi khoảng giá trị của dữ liệu |
| 17 |  | Segmentation | Phân đoạn |
| 18 |  | Standardization | Phương pháp biến đổi đưa các mẫu dữ liệu khác nhau về một kiểu dữ liệu chung. |
| 19 | SSE | Sum Of Squared Errors | Tổng bình phương sai số |
| 20 |  | Unsupervised Learning | Học không giám sát |
| 21 | WCSS | Within-Cluster Sum of Squares | Bình phương khoảng cách trung bình giữa tất cả các điểm trong cụm đến tâm cụm |
| 22 | GIS | Geographic Information System | Hệ thống thông tin địa lý |
| 23 |  | Web Mining | Khai phá dữ liệu |

# Danh mục hình vẽ

[Hình 1: Hình dạng các loại cụm 16](#_Toc90931803)

[Hình 2: Các bước trong quá trình phân cụm 18](#_Toc90931804)

[Hình 3: Lưu đồ thuật toán K-means. 26](#_Toc90931805)

[Hình 4: Mô tả ví dụ 28](#_Toc90931806)

[Hình 5: Giả mã thuật toán K-Means 28](#_Toc90931807)

[Hình 6: Kết quả thuật toán K-means khi chọn a) K=1, b) K=2, c) K=3. 30](#_Toc90931808)

[Hình 7: Đồ thị của Elbow method. 32](#_Toc90931809)

[Hình 8: Minh họa phân tích Silhouette cho k cụm lần lượt là k=2 (góc trên bên trái), k=3 (góc trên bên phải), k=4 (góc dưới bên trái), k=5 (góc dưới bên phải). 34](#_Toc90931810)

[Hình 9: Mô tả một phần dữ liệu Wholesale customers. 36](#_Toc90931811)

[Hình 10: Mô tả 5 hàng đầu tiên của dữ liệu Wholesale customers 37](#_Toc90931812)

[Hình 11: Kích thước và thông tin về dữ liệu Wholesale customers. 37](#_Toc90931813)

[Hình 12: Thông tin thống kê của dữ liệu Wholesale. 38](#_Toc90931814)

[Hình 13: Dữ liệu Wholesale sau khi tiền xử lý bước 1. 38](#_Toc90931815)

[Hình 14: Dữ liệu thống kê sau khi chuẩn hóa. 39](#_Toc90931816)

[Hình 15: Dữ liệu Wholesale sau khi giảm chiều. 40](#_Toc90931817)

[Hình 16: Một phần dữ liệu Wholesale sau khi giảm chiều. 41](#_Toc90931818)

[Hình 17: Đồ thị Elbow chọn ra số lượng cụm tối ưu 42](#_Toc90931819)

[Hình 18: Số điểm Silhouette cho từng cụm. 42](#_Toc90931820)

[Hình 19: Trực quan hóa 2 cụm dữ liệu Wholesale 44](#_Toc90931821)

[Hình 20: Biểu đồ cột thể hiện kích thước các cụm 45](#_Toc90931822)

[Hình 21: Biểu đồ tròn thể hiện tỷ lệ giữa các cụm 45](#_Toc90931823)

[Hình 22: Lượng chi tiêu cho các sản phẩm đối với từng cụm 46](#_Toc90931824)

[Hình 23: Chi tiêu của các kênh và vùng khách hàng 47](#_Toc90931825)

[Hình 24: Ma trận tương quan giữa các sản phẩm 49](#_Toc90931826)

# Danh mục bảng biểu

[Bảng 1. Dữ liệu của ví dụ 1 25](#_Toc90932223)

[Bảng 2. Khoảng cách Manhattan của Ví dụ 1 (bước 1) 26](#_Toc90932224)

[Bảng 3. Khoảng cách Manhattan của Ví dụ 1 (bước 2) 26](#_Toc90932225)

# Phần I: Mở đầu

## 1.1. Đặt vấn đề

Trong bối cảnh cạnh tranh ngày càng gay gắt, giữa các doanh nghiệp đang diễn ra một cuộc chạy đua giành giật thị trường và khách hàng, thì việc hiểu rõ nhu cầu của khách hàng, phát triển các sản phẩm phù hợp với các đối tượng khách hàng khác nhau trở nên vô cùng quan trọng, cấp thiết. Theo nguyên lý Pareto, 20% khách hàng sẽ mang lại 80% doanh số. Tuy nhiên, một loại sản phẩm đồng nhất không thể thỏa mãn được nhu cầu và sở thích của tất cả khách hàng. Do đó, doanh nghiệp cần xác định được những khách hàng quan trọng nhất để chăm sóc đặc biệt. Những tập khách hàng này được gọi là VIP, Priority hoặc khách hàng cao cấp, tùy từng doanh nghiệp có cách gọi khác nhau. Việc phân chia khách hàng thành các nhóm khác nhau dựa trên nhu cầu mua sắm sẽ giúp doanh nghiệp kinh doanh hiệu quả hơn, marketing nhắm đến đúng tập khách hàng hơn, và khách hàng được phục vụ tốt hơn.

Hiện nay, có nhiều phương pháp giúp doanh nghiệp thực hiện việc phân cụm khách hàng mục tiêu dựa trên những hiểu biết về hành vi (behavior), thói quen (habits), sở thích (preferences) của khách hàng tiềm năng. Trong phạm vi nghiên cứu, em lựa chọn phương pháp phân cụm theo thuật toán K-Means, đây là thuật toán quan trọng và được sử dụng phổ biến trong các nghiên cứu hiện nay (Chapman và Feit, 2019) để thực hiện đề tài “***Tìm hiểu thuật toán K-Means và ứng dụng***” sử dụng phương pháp phân cụm K-means để giải quyết bài toán phân khúc khách hàng.

## 1.2. Mục đích và yêu cầu

### 1.2.1. Mục đích

Mục đích nghiên cứu của khóa luận là vận dụng lý thuyết về thuật toán K-Means và phân khúc thị trường, thông qua ngôn ngữ lập trình Python để phân nhóm các đối tượng mặt hàng bán buôn dựa trên chi tiêu hàng năm cho các danh mục sản phẩm, qua đó xác định phân đoạn thị trường nào là mục tiêu, giải pháp marketing nào cần ưu tiên cho nhà quản lý để thúc đẩy doanh số bán hàng.

### 1.2.2. Yêu cầu

Những yêu cầu cần làm khi thực hiện đề tài:

- Tìm hiểu về phân cụm và thuật toán K-Means

- Có hiểu biết sơ bộ về ngôn ngữ lập trình Python và các thư viện trong Python (numpy, pandas, sklearn, seaborn, matplotlib,…)

- Phân tích kết quả và đánh giá chất lượng phân cụm.

# Phần II: Tổng quan tình hình nghiên cứu trong và ngoài nước

## 2.1. Tình hình nghiên cứu trong nước

Bài toán phân cụm đã được nghiên cứu và đề xuất từ cách đây hơn nửa thế kỷ: Thuật toán K-Means, thuật toán phân cụm kiểu phân tầng (1967), thuật toán phân cụm dựa trên đồ thị (1973), thuật toán phân cụm mờ (1981), thuật toán phân cụm dựa trên mô hình mật độ (DBSCAN, 1996). Với mỗi loại thuật toán đã đề xuất luôn tồn tại những ưu điểm và nhược điểm riêng. Với K-Means, đó là kết quả phân cụm phụ thuộc vào các điểm trọng tâm khởi tạo ban đầu, với thuật toán DBSCAN và phương pháp phân cụm thứ bậc là độ phức tạp của thuật toán cao, đối với phương pháp sử dụng đồ thị là khó xác định tham số,… Trong số những thuật toán cơ bản nêu trên, K-Means là thuật toán được sử dụng nhiều nhất.

Một trong những hướng nghiên cứu thu hút các nhà nghiên cứu trong những năm gần đây là hướng nghiên cứu về thuật toán phân cụm nửa giám sát. Xuất phát từ những thuật toán cơ bản kể trên, giả thiết rằng chúng ta có thêm một số thông tin chẳng hạn như có một số ràng buộc giữa các phần tử trong tập dữ liệu hay đã có trước một số phần tử được gán nhãn sẵn xác định chúng ở cụm nào. Ý tưởng cơ bản là phát triển các thuật toán có thể tích hợp các thông tin bổ trợ nhằm tăng chất lượng của quá trình phân cụm. Hàng loạt các công trình nghiên cứu nhằm cải thiện K-Means và ứng dụng trong đa lĩnh vực như marketing, mạng máy tính, nhận dạng, v.v... đã được công bố:

- Một cải tiến thuật toán K-Means song song sử dụng phương pháp lấy mẫu (*Kỷ yếu Hội thảo Khoa học Quốc gia CITA, 2017*), cải tiến thuật toán phân cụm dữ liệu PK-means, sau đó trình bày thuật toán cải tiến SK-meansMR và thực nghiệm đánh giá chất lượng, tốc độ của thuật toán trên dữ liệu lớn.

- Giải thuật K-Means xử lý song song với mô hình MapReduce (*Chuyên san Kỹ thuật và Công nghệ, 2015*), đạt được hiệu suất cao hơn khi phân loại tự động dữ liệu lớn và chứng tỏ tính hiệu quả và tính chính xác của giải thuật so với thuật toán K-Means ban đầu.

- Thuật toán phân cụm K-Means\* (*Tạp chí Khoa học và Công nghệ ĐH Thái Nguyên, 2017*), kết hợp giữa phương pháp minmax và phương pháp sử dụng seed cho kết quả tương đối tốt và có thể áp dụng trong thực tế đối với một số dạng bài toán không thể tìm đủ các dữ liệu đã gán nhãn cho các lớp ban đầu.

- Cải tiến thuật toán K-Means và ứng dụng hỗ trợ sinh viên chọn chuyên ngành theo học chế tín chỉ (*Tạp chí Khoa học công nghệ và Thực phẩm 15, 2018*), đề xuất phương pháp phân cụm trên tập dữ liệu nhóm điểm đặc trưng cho mỗi chuyên ngành. Kết quả cho thấy thời gian thực hiện thuật toán sau khi khử nhiễu giảm đáng kể và gợi ý lựa chọn chuyên ngành của sinh viên khá chính xác, giúp sinh viên có quyết định chắc chắn và hợp lý hơn khi lựa chọn chuyên ngành phù hợp với khả năng của mình.

- Ứng dụng thuật toán K-Means trong phân cụm khách hàng mục tiêu (*Tạp chí Khoa học Kỹ thuật Mỏ - Địa chất Tập 61, Kỳ 5, 2020*), thực hiện trên 272 khách hàng với các đặc điểm về độ tuổi, thu nhập và điểm chi tiêu. Kết quả nghiên cứu đã chia thành 2 cụm khách hàng mục tiêu, hứa hẹn sẽ giúp việc chăm sóc, tiếp thị khách hàng hiệu quả hơn; giúp đơn vị kinh doanh có những chiến lược marketing phù hợp giảm chi phí và tăng hiệu quả.

- Sử dụng thuật toán K-means trong bài toán phân loại đám mây điểm LiDAR (*Tạp Chí Khoa học và Công nghệ Việt Nam, 16*), đề cập đến việc phân loại dữ liệu LiDAR sử dụng thuật toán K-means. Tuy nhiên, thử nghiệm mới chỉ sử dụng một thuộc tính (độ cao của dữ liệu điểm) trong phân cụm, chưa giải quyết được bài toán có nhiều thuộc tính.

- Giao thức định tuyến dựa trên K-Means có thích ứng năng lượng cho mạng cảm biến không dây (*Tuyển tập Hội nghị Khoa học thường niên năm 2014*), đề xuất một cách tiếp cận mới KMR cho bài toán định tuyến trong mạng WSNs, và so sánh nó với một số giao thức khác. KMR tiến hành phân chia các nút cảm biến vào trong các cụm sử dụng thuật toán phân cụm K-means.

- Dùng đặc trưng Gabor kết hợp AdaBoost và K-means trong bài toán nhận dạng mặt người (*Tạp chí Khoa học ĐHSP TPHCM, 2013*), cho kết quả khả quan hơn so với phương pháp ban đầu với cơ sở dữ liệu cơ sở dữ liệu đa dạng.

## 2.2. Tình hình nghiên cứu ở nước ngoài

Để cải tiến chất lượng của K-Means, năm 2002, Basu và các cộng sự đề xuất phương pháp Seed-Kmeans với giả thiết nếu được cung cấp đầy đủ mỗi cụm một vài các phần tử đã được gán nhãn, các trọng tâm ban đầu ở bước khởi tạo sẽ được sử dụng bởi chính các thông tin đó và sẽ làm cho chất lượng của thuật toán tăng lên. Năm 2009, M. A. Hasan và các cộng sự đề xuất thuật toán ROBIN bằng cách đề xuất sử dụng phương pháp min-max để ước lượng các trọng tâm ở bước khởi tạo và cũng làm cho chất lượng phân cụm của K-Means được tăng lên. Trong năm 2016, Yoder và Priebe giới thiệu phương pháp phân cụm nửa giám sát cho K-Means++, trong đó tác giả đã sử dụng phương pháp tìm các trọng tâm theo thuật toán Lloyd và kết hợp với một số dữ liệu đã được gán nhãn trước.

Các phương pháp được để chọn ra số cụm k cũng đã được các nhà nghiên cứu đề xuất, điển hình trong “Selection of K in K-means clustering” (*Trung tâm Kỹ thuật Sản xuất ĐH Cardiff, Anh, 2005*) hay “Learning the k in k-means” (*Khoa Khoa học máy tính, ĐH California, Mỹ, 2004*) đã đưa ra cách chọn k bằng G-means và X-means. Cùng với đó là các kỹ thuật cải tiến thuật toán K-means và đưa vào ứng dụng trong các bài toán dự đoán kết quả học tập sinh viên “Application of k-Means Clustering algorithm for prediction of Students’ Academic Performance” (*Tạp chí Quốc tế về Khoa học Máy tính và An toàn Thông tin, 2010*), phân đoạn ảnh màu “Determination of Number of Clusters in K-Means Clustering and Application in Colour Image Segmentation” (*Khoa Khoa học Máy tính và Kỹ thuật Phần mềm Đại học Monash, Úc*), phân loại lưu lượng mạng “Network Traffic Classification Using K-means Clustering” (*IMSCCS 2007*) đã thu được nhiều kết quả tốt.

Qua khảo sát tất cả những công trình nói trên, ở những mức độ khác nhau, đã giúp em có một số tư liệu và kiến thức cần thiết để hình thành những hiểu biết chung, giúp tiếp cận và đi sâu nghiên cứu về thuật toán K-Means và ứng dụng của nó trong phân khúc khách hàng.

# Phần III: Nội dung và phương pháp nghiên cứu

## 3.1. Địa điểm và thời gian nghiên cứu (thực tập)

Nghiên cứu được thực hiện tại Công ty TNHH Thương mại dịch vụ và Phát triển Trần Vỹ trong khoảng thời gian từ tháng 08/2021 đến tháng 12/2021.

## 3.2. Nội dung nghiên cứu

Để hoàn thành các yêu cầu, khóa luận tập trung nghiên cứu các nội dung chính như sau:

- Tìm hiểu thuật toán phân cụm K-means

- Phân tích, xử lý bảng dữ liệu khách hàng

- Ứng dụng K-means vào phân cụm dữ liệu khách hàng

- Đánh giá chất lượng phân cụm

## 3.3. Phương pháp nghiên cứu:

- Phương pháp phân loại và hệ thống hóa lý thuyết

- Phương pháp phân tích và tổng hợp lý thuyết

- Phương pháp thu thập số liệu thứ cấp

# Phần IV: Kết quả nghiên cứu và thảo luận

## 4.1. Tổng quan về phân cụm dữ liệu

### 4.1.1. Khái niệm chung

Khám phá tri thức (phát hiện tri thức) trong cơ sở dữ liệu là một quy trình nhận biết các mẫu hoặc các mô hình trong dữ liệu với các tính năng: phân tích, tổng hợp, hợp thức, khả ích và có thể hiểu được.

Khai phá dữ liệu là một bước trong quá trình khám phá tri thức, gồm các thuật toán khai thác dữ liệu chuyên dùng dưới một số quy định về hiệu quả tính toán chấp nhận được để tìm ra các mẫu hoặc các mô hình trong dữ liệu. Nói cách khác, mục tiêu của khai phá dữ liệu là tìm kiếm các mẫu hoặc mô hình tồn tại trong cơ sở dữ liệu nhưng ẩn trong khối lượng lớn dữ liệu.

Phân cụm (clustering/segmentation) là một kỹ thuật trong khai phá dữ liệu, nhằm tìm kiếm, phát hiện các cụm, các mẫu dữ liệu tự nhiên tiềm ẩn trong tập dữ liệu lớn, từ đó cung cấp thông tin, tri thức hữu ích cho việc ra quyết định. Kỹ thuật này yêu cầu sắp xếp các đối tượng theo từng cụm dữ liệu tự nhiên, tức là số lượng và tên cụm chưa được biết trước. Các đối tượng được gom cụm sao cho mức độ tương tự giữa các đối tượng trong cùng một cụm là lớn nhất và mức độ tương tự giữa các đối tượng nằm trong các cụm khác nhau là nhỏ nhất.

### 4.1.2. Phân cụm dữ liệu

Phân loại là một trong những hành vi nguyên thuỷ nhất của con người nhằm nắm giữ lượng thông tin khổng lồ họ nhận được hằng ngày, vì sự xử lý mọi thông tin như một thực thể đơn lẻ là không thể. Phân cụm dữ liệu nhằm mục đích chính là khai phá cấu trúc của mẫu dữ liệu để thành lập các nhóm dữ liệu từ tập dữ liệu lớn, theo đó, cho phép người ta đi sâu vào phân tích và nghiên cứu cho từng cụm dữ liệu này nhằm khai phá và tìm kiếm các thông tin tiềm ẩn, hữu ích phục vụ cho ra quyết định.

#### 4.1.2.1. Khái niệm

Phân cụm là phương pháp phân loại một tập dữ liệu (hay còn gọi là tập luyện), trong đó quá trình tạo ra các cụm không sử dụng bất kỳ kiến thức tiên nghiệm nào về chỉ số lớp của các cá thể thuộc tập luyện.

Khi cho một tập luyện S gồm N phần tử:

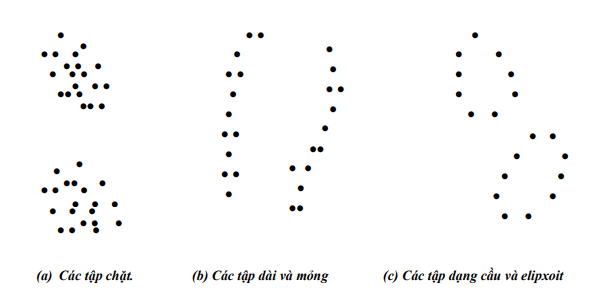
Thì quá trình phân cụm có thể được phát biểu như sau:

Tìm các miền con: S1, S2 · · · , SK của tập luyện S sao cho mỗi xi, i= sẽ chỉ thuộc vào duy nhất một miền con xác định trên, nghĩa là:

S1 ∪ S2 ∪ · · · ∪ SK = S

Si ∩ Sj = ∅ với ∀i ≠ j

Khi đó ta nói tập luyện có N phần tử trên đã được phân thành K cụm khác nhau. Đồng thời, các véctơ thuộc cùng một cụm Si thì "gần nhau" hơn, còn các véctơ thuộc các cụm khác nhau thì không "gần nhau", trong đó tiêu chuẩn độ đo sự gần gũi giữa các véctơ của các cụm sẽ được lựa chọn thích hợp với kiểu dáng hình thành của tập luyện cũng như với các thành phần được xác định (hay còn gọi là các biến) của các véctơ dạng mẫu trong tập luyện. Thêm vào đó, các véctơ trong một cụm là tương tự nhau hơn so với các véctơ thuộc một cụm khác. Lượng hoá thuật ngữ “tương tự” và “không tương tự” phụ thuộc rất nhiều vào loại của cụm. Chẳng hạn, loại cụm chặt thì có một số độ đo phù hợp, trong khi loại cụm có hình dáng dài và mỏng lại phù hợp hơn với các độ đo loại khác. Với định nghĩa trên, mỗi véctơ chỉ thuộc về một cụm riêng nên loại phân cụm này thỉnh thoảng còn được gọi là chặt hay rõ (hard or crisp).



Hình : Hình dạng các loại cụm

Không giống như phân lớp, phân cụm không đòi hỏi phải cho trước các chỉ số lớp của các véctơ dạng mẫu của tập luyện, cũng vì đặc điểm này nên phân cụm còn có thể được sử dụng như một bước tiền xử lí cho các thuật toán phân loại.

#### 4.1.2.2. Các bước cơ bản để phân cụm

#### Chọn lựa đặc trưng

Các đặc trưng phải được chọn lựa một cách hợp lý để có thể “mã hoá” nhiều nhất thông tin liên quan đến công việc quan tâm. Mục tiêu chính là phải giảm thiểu sự dư thừa thông tin giữa các đặc trưng. Các đặc trưng cần được tiền xử lý trước khi dùng trong các bước sau.

#### Chọn độ đo khoảng cách

Đây là một độ đo chỉ ra mức độ tương tự hay không tương tự giữa hai véctơ đặc trưng. Phải đảm bảo rằng tất cả các véctơ đặc trưng góp phần như nhau trong việc tính toán độ đo gần gũi và không có đặc trưng nào át hẳn đặc trưng nào. Điều này được đảm nhận bởi quá trình tiền xử lý. Các phương pháp cổ điển để đo khoảng cách là khoảng cách Euclide và Manhattan, được định nghĩa như sau:

Khoảng cách Euclid:

(1)

Khoảng cách Manhattan:

(2)

Trong đó, x và y là hai vectơ có độ dài n.

#### Tiêu chuẩn phân cụm

Điều này phụ thuộc vào sự giải thích của chuyên gia cho thuật ngữ “dễ nhận thấy” dựa vào loại của các cụm được chuyên gia cho rằng đang ẩn giấu dưới tập dữ liệu. Chẳng hạn, một cụm loại chặt (compact) của các véctơ đặc trưng trong không gian ℓ-chiều có thể dễ nhận thấy theo một tiêu chuẩn, trong khi một cụm loại “dài và mỏng” lại có thể đươc dễ nhận thấy bởi một tiêu chuẩn khác. Tiêu chuẩn phân loại có thể được diễn đạt bởi hàm chi phí hay một vài loại quy tắc khác.

#### Thuật toán phân loại

Cần lựa chọn một sơ đồ thuật toán riêng biệt nhằm làm sáng tỏ cấu trúc cụm của tập dữ liệu.

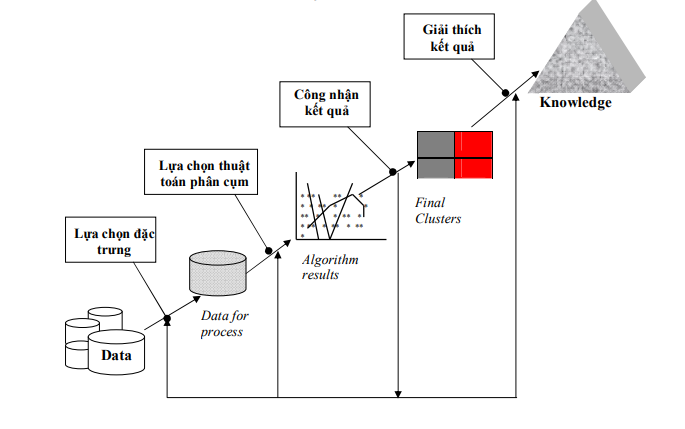
#### Công nhận kết quả

Khi đã có kết quả phân loại thì ta phải kiểm tra tính đúng đắn của nó. Điều này thường được thực hiện bởi việc dùng các kiểm định phù hợp.

#### Giải thích kết quả

Trong nhiều trường hợp, chuyên gia trong lĩnh vực ứng dụng phải kết hợp kết quả phân loại với bằng chứng thực nghiệm và phân tích để đưa ra các kết luận đúng đắn. Trong một số trường hợp, nên có cả bước khuynh hướng phân cụm; trong bước này có các kiểm định khác nhau để chỉ ra một dữ liệu có hay không một cấu trúc phân cụm. Ví dụ như tập dữ liệu có thể hoàn toàn ngẫu nhiên, vì vậy mọi cố gắng phân cụm đều vô nghĩa.

Các lựa chọn khác nhau của các đặc trưng, độ đo gần gũi, tiêu chuẩn phân cụm có thể dẫn tới các kết quả phân cụm khác nhau. Do đó, việc lựa chọn một cách hợp lý nhất hoàn toàn dựa vào kiến thức và kinh nghiệm của chuyên gia. Tính chủ quan (của chuyên gia) là một thực tế mà ta phải chấp nhận.



Hình : Các bước trong quá trình phân cụm

#### 4.1.2.2. Ý nghĩa thực tiễn của phân cụm dữ liệu

Phân cụm có ý nghĩa rất quan trọng trong hoạt động của con người. Ngay từ nhỏ, con người đã học cách làm thế nào để phân biệt được giữa mèo và chó, giữa động vật và thực vật và liên tục đưa vào sơ đồ phân loại trong tiềm thức của mình. Phân cụm được ứng dụng rộng rãi trong nhiều lĩnh vực, bao gồm nhận dạng mẫu, phân tích dữ liệu, xử lí ảnh, nghiên cứu thị trường, địa lí, địa chất, y học,... Phân cụm có thể được sử dụng như một công cụ độc lập giúp hình thành các đặc trưng của mỗi cụm trong sự phân bố của tập luyện và từ đó tập trung vào một tập riêng biệt của các cụm để giúp cho việc phân tích đạt kết quả.

#### Phân cụm dữ liệu phục vụ cho biểu diễn dữ liệu gene

Phân cụm là một trong những phân tích được sử dụng thường xuyên nhất trong biểu diễn dữ liệu gene (Yeung et al., 2003; Eisen at al., 1998). Dữ liệu biểu diễn gene là một tâp hợp các phép đo được lấy từ DNA microarray (còn gọi là DNA chip hay gene chip) là một tấm thủy tinh hoặc nhựa trên đó có gắn các đoạn DNA thành các hàng siêu nhỏ. Các nhà nghiên cứu sử dụng các con chip như vậy để sàng lọc các mẫu sinh học nhằm kiểm tra sự có mặt hàng loạt trình tự cùng một lúc. Các đoạn DNA gắn trên chip được gọi là probe (mẫu dò). Trên mỗi điểm của chip có hàng ngàn phân tử probe với trình tự giống nhau.

Dữ liệu biểu diễn gen có thể được phân cụm theo hai cách. Cách thứ nhất là nhóm các mẫu gen giống nhau, ví dụ như gom các dòng của ma trận D, cách khác là nhóm các mẫu khác nhau trên các hồ sơ tương ứng.

#### Phân cụm dữ liệu phục trong sức khỏe tâm lý

Phân cụm dữ liệu áp dụng trong nhiều lĩnh vực sức khỏe tâm lý, bao gồm cả việc thúc đẩy và duy trì sức khỏe, cải thiện cho hệ thống chăm sóc sức khỏe, và công tác phòng chống bệnh tật và người khuyết tật (Clatworthy et al., 2005). Trong sự phát triển hệ thống chăm sóc sức khỏe, phân cụm dữ liệu được sử dụng để xác định các nhóm của người dân mà có thể được hưởng lợi từ các dịch vụ cụ thể (Hodges và Wotring, 2000). Trong thúc đẩy y tế, nhóm phân tích được sử dụng để lựa chọn nhắm mục tiêu vào nhóm sẽ có khả năng đem lại lợi ích cho sức khỏe cụ thể từ các chiến dịch quảng bá và tạo điều kiện thuận lợi cho sự phát triển của quảng cáo. Ngoài ra, phân cụm dữ liệu được sử dụng để xác định các nhóm dân cư bị rủi ro do phát triển y tế và các điều kiện những người có nguy cơ nghèo.

Phân cụm dữ liệu đối với hoạt đông nghiên cứu thị trường Trong nghiên cứu thị trường, phân cụm dữ liệu được sử dụng để phân đoạn thị trường và xác định mục tiêu thị trường (Chrisoppher, 1969; Saunders, 1980, Frank and Green, 1968). Trong phân đoạn thị trường, phân cụm dữ liệu thường được dùng để phân chia thị trường thành những cụm mang ý nghĩa, chẳng hạn như chia ra đối tượng nam giới từ 21-30 tuổi và nam giới ngoài 51 tuổi, đối tượng nam giới ngoài 51 tuổi thường không có khuynh hướng mua các sản phẩm mới.

#### Phân cụm dữ liệu đối với hoạt động Phân đoạn ảnh

Phân đoạn ảnh là việc phân tích mức xám hay màu của ảnh thành các lát đồng nhất (Comaniciu and Meer, 2002). Trong phân đoạn ảnh, phân cụm dữ liệu thường được sử dụng để phát hiện biên của đối tượng trong ảnh.

Phân cụm dữ liệu là một công cụ thiết yếu của khai phá dữ liệu, khai phá dữ liệu là quá trình khám phá và phân tích một khối lượng lớn dữ liệu để lấy được các thông tin hữu ích (Berry and Linoff, 2000). Phân cụm dữ liệu cũng là một vấn đề cơ bản trong nhận dạng mẫu (pattern recognition).

Ngoài ra, phân cụm dữ liệu còn được ứng dụng trong một số lĩnh vực như:

- Thương mại: Trong thương mại, phân cụm có thể giúp các thương nhân khám phá ra các nhóm khách hàng quan trọng có các đặc trưng tương đồng nhau và đặc tả họ từ các mẫu mua bán trong cơ sở dữ liệu khách hàng;

- Thư viện: Phân loại các cụm sách có nội dung và ý nghĩa tương đồng nhau để cung cấp cho độc giả, cũng như đặt hàng với nhà cung cấp;

- Sinh học: Trong sinh học, phân cụm được sử dụng để xác định các loại sinh vật, phân loại các gen với chức năng tương đồng và thu được các cấu trúc trong các mẫu.

- Bảo hiểm: Nhận dạng nhóm tham gia bảo hiểm có chi phí yêu cầu bồi thường trung bình cao, xác định gian lận trong bảo hiểm thông qua các mẫu cá biệt;

- Phân tích dữ liệu không gian: Do sự đồ sộ của dữ liệu không gian như dữ liệu thu được từ các hình ảnh chụp từ vệ tinh các thiết bị y học hoặc hệ thống thông tin địa lý (GIS),…làm cho người dùng rất khó để kiểm tra các dữ liệu không gian một cách chi tiết. Phân cụm có thể trợ giúp người dùng tự động phân tích và xử lý các dữ liệu không gian như nhận dạng và chiết xuất các đặc tính hoặc các mẫu dữ liệu quan tâm có thể tồn tại trong cơ sở dữ liệu không gian;

- Lập quy hoạch đô thị: nhận dạng các nhóm nhà theo kiểu, vị trí địa lí, giá trị ... nhằm cung cấp thông tin cho quy hoạch đô thị;

- Nghiên cứu địa chấn: phân cụm để theo dõi các tâm động đất nhằm cung cấp thông tin cho việc nhận dạng các vùng nguy hiểm;

- Địa lý: Phân lớp các động vật và thực vật và đưa ra đặc trưng của chúng;

- Web Mining: Phân cụm có thể khám phá các nhóm tài liệu quan trọng, có nhiều ý nghĩa trong môi trường Web. Các lớp tài liệu này trợ giúp cho việc khám phá tri thức từ dữ liệu,…

Như vậy, phân cụm dữ liệu là một phương pháp xử lý thông tin quan trọng và phổ biến, nó nhằm khám phá mỗi liên hệ giữa các mẫu dữ liệu bằng cách tổ chức chúng thành các cụm tương tự.

### 4.1.3. Một số kỹ thuật tiếp cận trong phân cụm dữ liệu

Các kỹ thuật phân cụm có rất nhiều cách tiếp cận và các ứng dụng trong thực tế, nó đều hướng tới hai mục tiêu chung đó là chất lượng của các cụm khám phá được và tốc độ thực hiện của thuật toán. Hiện nay, các kỹ thuật phân cụm có thể phân loại theo các phương pháp tiếp cận chính như sau:

- Phân cụm phân họach (Partitioning Methods) với một số thuật toán điển hình như: K-Means, K-Medoids, CLARA, CLARANS,… ;

- Phân cụm phân cấp (Hierarchical Methods) với một số thuật toán điển hình như: AGNES, DIANA, BIRCH (1996), CURE (1998), CHAMELEON (1999);

- Phân cụm dựa trên mật độ (Density-Based Methods) với một số thuật toán điển hình như DBSCAN, DENCLUE, CLIQUE, OPTICS;

- Phân cụm dựa trên lưới (Grid-Based Methods) với một số thuật toán điển hình như STING, WAVECLUSTER, CLIQUE;

- Phân cụm dựa trên mô hình (Model-Based Clustering Methods) với một số thuật toán điển hình như EM, COBWEB, CLASSIT, AutoClass;

- Phân cụm có dữ liệu ràng buộc (Binding data Clustering Methods).

### 4.1.4. Các vấn đề thường gặp khi phân cụm

Một vấn đề thường gặp trong phân cụm là hầu hết các dữ liệu cần cho phân cụm đều có chứa nhiễu do quá trình thu thập thiếu chính xác hoặc không đầy đủ, vì vậy cần phải xây dựng chiến lược cho bước tiền xử lí dữ liệu nhằm khắc phục hoặc loại bỏ nhiễu trước khi chuyển sang giai đoạn phân tích cụm dữ liệu. Nhiễu ở đây được hiểu là các dữ liệu không chính xác, không tường minh, hoặc thiếu thông tin về một số thuộc tính... Một trong những kỹ thuật xử lí nhiễu phổ biến là việc thay thế giá trị các thuộc tính của dữ liệu bị nhiễu bằng giá trị thuộc tính tương ứng. Ngoài ra, phát hiện phần tử ngoại lai cũng là một trong những hướng nghiên cứu quan trọng trong phân cụm, chức năng chính của nó là xác định một nhóm nhỏ các dữ liệu không tuân theo các mô hình của tập dữ liệu đang được xét, nhằm tránh sự ảnh hưởng của chúng tới kết quả phân cụm.

Mục tiêu của phân cụm là xác định được bản chất của cấu trúc cụm trong học không giám sát (unsupervised learning). Để thực hiện được việc này cần phân tích bài toán thực tế nhằm tìm ra các tiêu chuẩn cho việc tạo thành các cụm tốt theo một ý nghĩa nào đó.

Theo các nghiên cứu đến thời điểm hiện nay thì chưa có một phương pháp phân cụm tổng quát nào có thể giải quyết trọn vẹn cho tất cả các dạng cấu trúc cơ sở dữ liệu. Hơn nữa, đối với các phương pháp phân cụm cần có cách thức biểu diễn cấu trúc của cơ sở dữ liệu, với mỗi cách thức biểu diễn khác nhau sẽ có tương ứng một thuật toán phân cụm phù hợp. Vì vậy phân cụm dữ liệu vẫn đang là một vấn đề khó và mở, vì phải giải quyết nhiều vấn đề cơ bản một cách trọn vẹn và phù hợp với nhiều dạng dữ liệu khác nhau, đặc biệt đối với dữ liệu hỗn hợp đang ngày càng tăng trong các hệ quản trị dữ liệu và cũng là một trong những thách thức lớn trong lĩnh vực khai phá dữ liệu.

## 4.2. Thuật toán K-Means trong bài toán phân cụm dữ liệu

### 4.2.1. Giới thiệu chung

Thuật toán K-means do McQueen giới thiệu trong tài liệu “J. Some Methods for Classification and Analysis of Multivariate Observations” năm 1967, sau đó được J.A. Hartigan và M.A.Wong phát triển vào khoảng năm 1975.

K-Means là một thuật toán dùng trong các bài toán phân loại/nhóm n đối tượng thành k nhóm dựa trên đặc tính/thuộc tính của đối tượng (k là số nguyên dương, nhỏ hơn n).

Thuật toán này sử dụng phương pháp tạo và cập nhật trung tâm để phân nhóm các điểm dữ liệu cho trước vào các nhóm khác nhau. Đầu tiên chúng sẽ tạo ra các điểm trung tâm ngẫu nhiên. Sau đó gán mỗi điểm trong tập dữ liệu vào trung tâm gần nó nhất. Sau đó chúng sẽ cập nhật lại trung tâm và tiếp tục lặp lại các bước đã kể trên. Điều kiện dừng của thuật toán là khi các trung tâm không thay đổi trong 2 vòng lặp kế tiếp nhau. Tuy nhiên, việc đạt được một kết quả hoàn hảo là rất khó và rất tốn thời gian, vậy nên thường người ta sẽ cho dừng thuật toán khi đạt được 1 kết quả gần đúng và chấp nhận được.

K-Means là một trong số những phương pháp học không giám sát cơ bản nhất, là thuật toán rất quan trọng và được sử dụng phổ biến trong kỹ thuật phân cụm.

### 4.2.2. Một số ứng dụng

K-Means là một thuật toán đa năng có thể được ứng dụng vào bất kì quy trình phân loại và chia nhóm nào. Một số ví dụ về các trường hợp sử dụng là:

- Phân đoạn thị trường theo hành vi (Behavioral segmentation): Trong giao dich ngân hàng, việc phân loại dữ liệu khách hàng là đặc biệt quan trọng, thường được dựa vào đó để đưa ra các chính sách chung cho toàn hệ thống hay là có những chính sách chăm sóc đến từng khách hàng. Một vài cách phân loại dựa trên hành vi người dùng như sau:

* Phân đoạn theo lịch sử thanh toán (chi tiêu)
* Phân đoạn theo các hoạt động trên ứng dụng di động, trang web hoặc nền tảng ATM
* Xác định tính cách cá nhân khách hàng dựa trên sở thích, mối quan tâm của họ (thông qua lịch sử mua sắm)
* Tạo hồ sơ khách hàng dựa trên dữ liệu giám sát hoạt động người dùng

- Phân loại sáng chế của cục sở hữu khoa học kĩ thuật (Inventory categorization):

* Nhóm các sáng chế theo hoạt động kinh doanh
* Nhóm các sáng chế dựa trên khối ngành sản xuất

- Phân loại cảm biến theo chức năng:

* Phát hiện các loại hoạt động trong cảm biến chuyển động
* Gom nhóm hình ảnh
* Phân tách âm thanh
* Chia nhóm trong theo dõi sức khỏe

- Phát hiện bất thường (outliers detection)

Đặc biệt, việc theo dõi sự thay đổi của những điểm dữ liệu bị theo dõi (track points), có thể là một số cá nhân trong dữ liệu khách hàng của ngân hàng, có thể được sử dụng để theo dõi xu hướng hoạt động. Điều này thường được sử dụng nhiều trong theo dõi hoạt động mua sắm, góp phần giúp cho các tập đoàn lớn như Nike, Zara, Uniqlo... nắm bắt được người dùng và đưa ra những chính sách góp phần nâng cao doanh số cũng như thúc đẩy hoạt động mua sắm.

### 4.2.3. Thuật toán K-Means

Mục đích của thuật toán K-Means là sinh ra k cụm dữ liệu {C1, C2,…,Ck} từ một tập dữ liệu chứa n đối tượng trong không gian d chiều Xi =(x1,x2,…xn), (i= ) sao cho hàm tiêu chuẩn:

(3)

đạt giá trị tối thiểu.

Trong đó:

‘mi’ là trọng tâm của cụm Ci

*‘||xi-mj||’* là khoảng cách giữa đối tượng *xi* và trọng tâm *mi*.

*‘k’* là số tâm cụm.

#### 4.2.3.1. Ý tưởng và ví dụ minh họa

Về nguyên lý, có n đối tượng, mỗi đối tượng có m thuộc tính, ta phân chia được các đối tượng thành k nhóm dựa trên các thuộc tính của đối tượng bằng việc áp dụng thuật toán này. Coi mỗi thuộc tính của đối tượng (đối tượng có m thuộc tính) như một toạ độ của không gian m chiều và biểu diễn đối tượng như một điểm của không gian m chiều.

Trong đó:

ai (i =) là đối tượng thứ i.

xij (i =; j=) là thuộc tính thứ j của đối tượng thứ i.

Phương thức phân loại/ phân nhóm dữ liệu thực hiện dựa trên khoảng cách Euclid nhỏ nhất giữa đối tượng đến phần tử trung tâm của các nhóm.

Trọng tâm của một cụm là một véctơ, trong đó giá trị của mỗi phần tửcủa nó là trung bình cộng của các thành phần tương ứng của các đối tượng vectơ.

#### Các bước của giải thuật

Thuật toán K-Means lặp đi lặp lại quá trình phân các ví dụ vào cụm có tâm gần nhất, sau đó là điều chỉnh tâm cụm, cho tới khi điều kiện hội tụ được thỏa mãn. Tham số đầu vào của thuật toán là số cụm k, và tham số đầu ra của thuật toán là các trọng tâm của các cụm dữ liệu.

Thuật toán K-means được thực hiện qua các bước chính sau:

Cho tập các điểm dữ liệu X = {x1, x2, x3, …, xn} và tập các tâm cụm M = {m1, m2, …, mk}.

**Input:** Số cụm k và các trọng tâm cụm {mj}kj=1.

**Output:** Các cụm C[i] (1≤ i≤  k) và hàm tiêu chuẩn E đạt giá trị tối

thiểu.

Begin

Bước 1. Khởi tạo ngẫu nhiên K tâm (centroid) ban đầu cho K cụm (cluster) trong không gian Rd (d là số chiều dữ liệu). Mỗi cụm được đại diện bằng các tâm của cụm.

Bước 2. Tính toán khoảng cách: Đối với mỗi điểm Xi (1≤i≤n), tính toán khoảng cách của nó tới mỗi trọng tâm mj (j=. Sau đó tìm trọng tâm gần nhất đối với mỗi điểm.

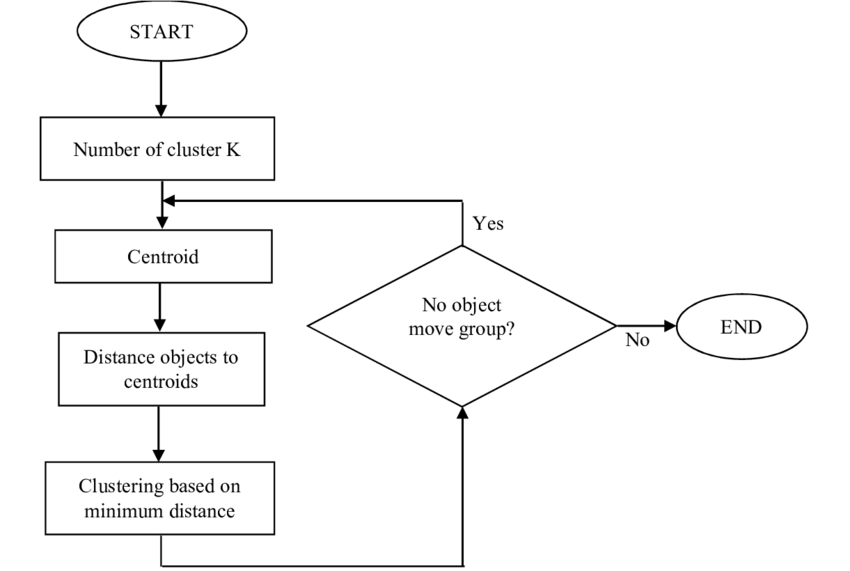
Bước 3. Điều chỉnh lại tâm cho các cụm: Đối với mỗi j=tọa độ của tâm cụm mới bằng tọa độ trung bình của tất cả các ví dụ trong cụm đó.

(4)

Trong đó, ‘ci’ là số điểm dữ liệu trong cụm thứ i.

Bước 4. Kiểm tra điều kiện dừng: nếu thuật toán chưa hội tụ, quay lại bước 2. Điều kiện dừng có thể được định nghĩa bằng số vòng lặp giới hạn, hoặc khi không có sự thay đổi về vị trí của tâm cụm.

End.



Hình : Lưu đồ thuật toán K-means.

#### Ví dụ 1

Quá trình phân cụm bằng thuật toán K-Means có thể minh họa bằng việc xử lí một tập luyện hai chiều X gồm 6 vectơ dạng mẫu sau:

Ta thực hiện các thuật toán với giả thiết cho trước số cụm K = 2 với tập luyện như trên theo các bước sau:

Bảng . Dữ liệu của ví dụ 1

|  |  |  |
| --- | --- | --- |
|  | x | y |
| X1 | 1 | 2 |
| X2 | 2 | 2 |
| X3 | 2 | 3 |
| X4 | 3 | 3 |
| X5 | 3 | 4 |
| X6 | 2 | 4 |

Bước 1: Chọn 2 véctơ dạng mẫu bất kì làm tâm của hai cụm khởi tạo đầu tiên.m1(1) = x1 = (1, 2)

m2(1) = x6 = (2, 4)

Bước 2: Tính khoảng cách từ các điểm đến các tâm cụm (ở đây dùng khoảng cách Manhattan):

Bảng .Khoảng cách Manhattan của Ví dụ 1 (bước 1)

|  |  |
| --- | --- |
| D(x1, m1) = | D(x1, m2) = |
| D(x2, m1) = | D(x2, m2) = |
| D(x3, m1) = | D(x3, m2) = |
| D(x4, m1) = | D(x4, m2) = |
| D(x5, m1) = | D(x5, m2) = |
| D(x6, m1) = | D(x6, m2) = |

Với việc khảo sát tương tự trên các véctơ của tập luyện X đối với hai véctơ x1 và x6 ở trên ta thu được kết quả phân cụm sau bước lặp đầu tiên là:

C1 = {x1 , x2} và

C2 = {x3 , x4 , x5 , x6}

Bước 3: Cập nhật trọng tâm cụm:

M1(2) =

M2(2) =

Do M1(1), m2(1) ≠ m1(2), m2(2), nên thuật toán chưa dừng, vì vậy chúng ta quay lại bước 2 và lặp lại các thao tác như trên.

Tính khoảng cách:

Bảng .Khoảng cách Manhattan của Ví dụ 1 (bước 2)

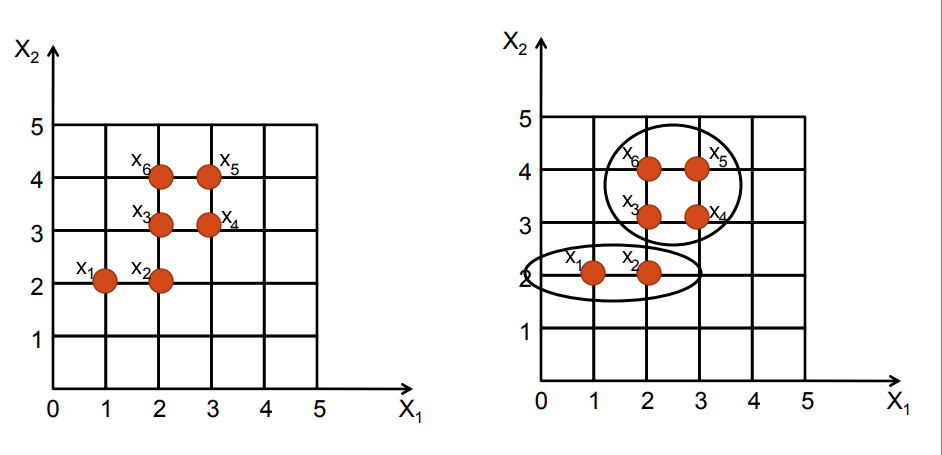
|  |  |
| --- | --- |
| D(x1, m1) = | D(x1, m2) = |
| D(x2, m1) = .5 | D(x2, m2) = |
| D(x3, m1) = | D(x3, m2) = |
| D(x4, m1) = | D(x4, m2) = |
| D(x5, m1) = | D(x5, m2) = |
| D(x6, m1) = | D(x6, m2) = |

Ta thu được kết quả phân cụm sau bước lặp thứ 2 là:  
C1 = {x1 , x2 } và

C2 = {x3 , x4 , x5 , x6}.

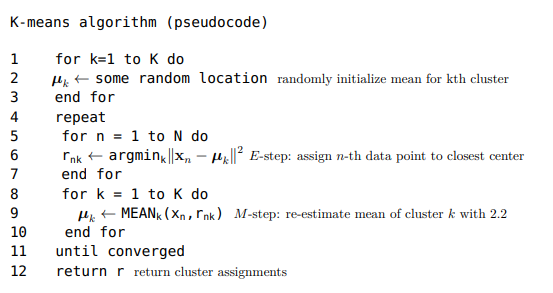
Sau lần lặp 2 không có sự phân bố lại các đối tượng giữa các cụm (điều kiện dừng lý tưởng). Giải thuật hội tụ và kết quả của quá trình phân cụm là:

C1 = {m1 , m2 } và C2 = {m3 , m4 , m5 , m6}.



Hình : Mô tả ví dụ 1

#### Giả mã



Hình : Giả mã thuật toán K-Means

#### 4.2.3.3. Độ phức tạp

Thuật toán K-Means có độ phức tạp tính toán là O(tKn). Trong đó n là số mẫu trong CSDL, K là số cluster, t là số lần lặp. Thông thường t, K << n, nên thuật toán này có hiệu quả tương đối với các CSDL lớn [Han, 2001].

#### 4.2.3.4. Nhận xét

Thuật toán K-means trên đã được chứng minh là hội tụ và có độ phức tạp tính toán không cao, vì vậy việc phân tích phân cụm là khá đơn giản nên có thể áp dụng đối với tập dữ liệu lớn.

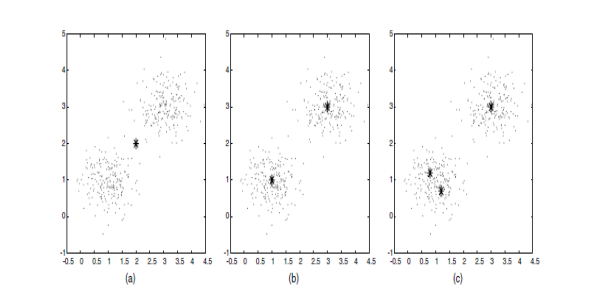
Thuật toán K-means có ưu điểm là dễ dàng cài đặt, tuy nhiên hiệu quả của thuật toán này phụ thuộc vào số lượng các tâm cụm ban đầu được chọn và thứ tự các véctơ dạng mẫu được xử lý trong thuật toán. Nó cũng phụ thuộc vào dạng biểu diễn hình học của tập dữ liệu được phân tích vì không phù hợp với việc khai phá các dữ liệu gồm các cluster có hình dạng không lồi. Có thể đưa thêm nhiều cải tiến vào K-means để được thuật toán hiệu quả hơn, như thay đổi cách chọn các mẫu khởi đầu, cách tính tiêu chuẩn,...

Từ đây ta có thể thấy rằng, mặc dù thuật toán K-means áp dụng rộng rãi trong nhiều lĩnh vực, nhưng nó có ba nhược điểm chính:

• Đối với những bộ dữ liệu có hình dạng phức tạp hoặc mất cân bằng thì thuật toán không hội tụ về qui luật phân chia tổng quát. Chẳng hạn như dữ liệu có dạng đường viền hình tròn bao ngoài một hình tròn ở bên trong nó; dữ liệu hình trôn ốc; dữ liệu có phân phối dẹt; dữ liệu bị mất cân bằng phân phối giữa các cụm.

• Trong trường hợp một số trọng tâm được khởi tạo quá lệch so với các điểm trọng tâm tự nhiên trong tập dữ liệu đầu vào thì các điểm này không có bất kỳ cơ hội nào được xét đến trong toàn bộ quá trình thực hiện thuật toán và vì vậy chúng được gọi là các điểm chết, dẫn đến kết quả phân cụm của K-means là rất thấp, nghĩa là các cụm dữ liệu được khám phá rất lệch so với các cụm trong thực tế.

• Thuật toán K-Means cần được xác định trước tham số đầu vào là số cụm K. Khi K được chọn "tốt" thì thuật toán K-Means có thể tìm kiếm được chính xác tâm cụm như đã chỉ ra ở hình (2.3b). Nếu không nó sẽ dẫn đến một kết quả phân cụm không đúng như mô tả trong hình (Hình 5a và 5c), nơi một số tâm cụm không ở đúng vị trí mà đáng ra chúng nên ở đó, thay vào đó chúng lại ở một trong những điểm biên giữa các cụm khác nhau hoặc ở tại các điểm lệch so với tâm cụm. Trên thực tế, có một số phương pháp được đưa vào để chọn tham số đầu vào nhưng chưa tìm ra một giải pháp tối ưu nhất, nên giải pháp thường được sử dụng là thử nghiệm với nhiều giá trị đầu vào k khác nhau rồi sau đó chọn ra giá trị tốt nhất.



Hình : Kết quả thuật toán K-means khi chọn a) K=1, b) K=2, c) K=3.

#### Các yếu tố quan trọng cần xem xét khi sử dụng K-Means

1. Số cụm k: Chúng ta cần phải xác định trước số cụm cho thuật toán, vì bộ dữ liệu chưa được gán nhãn nên dường như chúng ta không có thông tin nào về số lượng cụm hợp lý. Chúng ta chỉ có thể thực hiện phương pháp thử và sai (trial and error) và xác định số cụm thông qua một phương pháp chẳng hạn như Elbow.

2. Tâm cụm (centroids): Vị trí tâm của cụm sẽ bị phụ thuộc vào điểm khởi tạo ban đầu của chúng. Những vị trí khởi tạo khác nhau có thể dẫn tới cách phân cụm khác nhau, mặc dù thuật toán có cùng thiết lập số cụm.

3. Phần tử ngoại lai (outliers): Sự hình thành cụm rất nhạy cảm với sự hiện diện của các yếu tố ngoại lai. Các yếu tố ngoại lai kéo cụm về phía chính nó, khiến cho tâm cụm bị chệch và do đó dự báo cụm không còn chuẩn xác. Chính vì thế chúng ta cần phải loại bỏ outliers trước khi huấn luyện thuật toán.

4. Độ đo khoảng cách (Distance Measures): Sử dụng các độ đo khoảng cách khác nhau có thể cho ra các cụm khác nhau.

5. Thuật toán K-Means không hoạt động với dữ liệu phân loại (categorical).

6. Quá trình có thể không hội tụ theo số lần lặp đã cho, nên luôn luôn cần kiểm tra sự hội tụ.

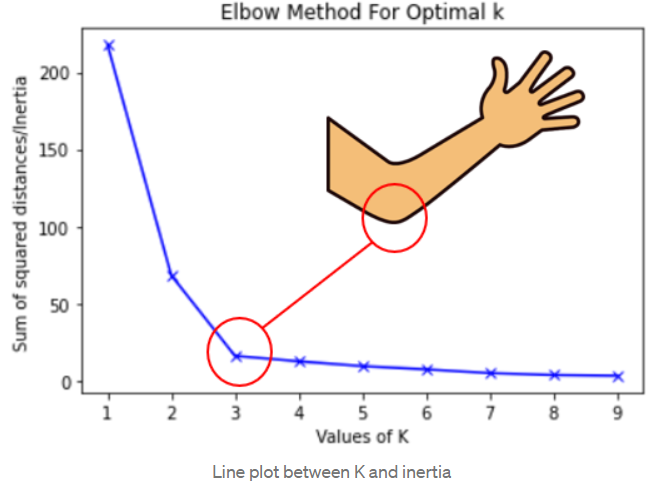
### 4.2.4. Một số phương pháp lựa chọn số cụm k

Ở các phần trước, chúng ta đã biết bước đầu tiên khi triển khai thuật toán cần phải xác định số cụm k muốn phân từ tập dữ liệu. Tuy nhiên việc quyết định xem cần phân ra bao nhiêu cụm để tối ưu nhất và các cụm khi kết thúc thuật toán có được đánh giá là phù hợp, chính xác là cực kỳ quan trọng và cần có phương pháp cụ thể. Nếu quá trình chọn số cụm k chỉ được thực hiện dựa trên kinh nghiệm phân tích, kiến thức chuyên môn, và mục đích kinh doanh mà không dựa trên đặc tính của dữ liệu thì khả năng cao việc ứng dụng phân cụm sẽ không mang lại giá trị mong đợi vì các cluster có thể không phản ánh tốt các quy luật, các mối quan hệ giữa những đối tượng quan sát trong tự nhiên mà đang tiềm ẩn trong tập dữ liệu.

#### 4.2.4.1. Phương pháp khuỷu tay (Elbow method)

Phương pháp Elbow được coi là phương pháp xác định số cụm k cho K-Means phổ biến nhất, được minh họa dưới dạng đồ thị đường cong với trục hoành là số k cụm, trục tung là WCSS (Within-Cluster Sum of Squares - tổng biến thiên bình phương khoảng cách giữa các điểm dữ liệu với tâm trong cụm, còn được gọi là SSE). Khi các điểm dữ liệu trong cùng một cụm có đặc điểm gần giống nhau thì chứng tỏ cụm đó “chất lượng” và ngược lại.

Thông thường, số k thường chạy từ 1 đến 10, với mỗi k sẽ có 1 chỉ số WCSS tương ứng. Khi vẽ đồ thị minh họa các cặp WCSS – k, đồ thị này trông giống như một cái khuỷu tay. Khi số lượng cụm tăng lên, giá trị WCSS sẽ bắt đầu giảm. Giá trị WCSS lớn nhất khi K = 1. Số k tối ưu chính là điểm mà ở đó WCSS giảm đều, đồ thị đi ngang hoặc giảm không đáng kể.



Hình : Đồ thị của Elbow method.

Trong đồ thị minh họa phía trên, k = 3 sẽ là số cụm ban đầu thích hợp cho tập dữ liệu.

#### 4.2.4.2. Phương pháp bóng mờ (Average Silhouette method)

Phương pháp phân tích điểm Silhouette cũng là một trong những cách thức phổ biến để đo lường mức độ tối ưu khi một quan sát hay một điểm dữ liệu được phân vào cụm bất kỳ bằng cách đo khoảng cách giữa các cụm. Sau đó, nó cung cấp một cách đánh giá thông số như số lượng cụm bằng cách cho điểm silhouette (silhouette score). Điểm số này là một số liệu đo lường mức độ gần của mỗi điểm trong một cụm với các điểm trong các cụm lân cận.

Với mỗi mẫu trong tập dữ liệu, điểm Silhouette được tính bằng công thức (5):

(5)

Trong đó, 𝑝 là khoảng cách giữa mẫu và tâm của cụm gần nhất.

𝑞 là khoảng cách giữa mẫu và tâm cụm của chính nó.

Giá trị điểm Silhouette có phạm vi là [-1, 1], cụ thể như sau:

- Điểm +1 : Điểm gần +1 cho biết mẫu nằm gọn trong cụm, cụm này dày đặc và tách biệt với các cụm khác.

- Điểm 0 : Điểm 0 cho biết mẫu nằm trên hoặc rất gần ranh giới quyết định giữa hai cụm lân cận, các cụm chồng chéo lên nhau.

- Điểm -1 : Điểm âm chỉ ra rằng các mẫu có thể đã được chỉ định vào các cụm sai.

Ngoài ra, theo Daniel T. Larose, tác giả cuốn Data Mining and Predictive Analytics thì:

- Điểm Silhouette từ 0.5 trở lên cho thấy có thể cluster này sát với thực tế.

- Điểm Silhoutte nằm trong khoảng từ 0.25 đến 0.5 thì cần thêm kiến thức chuyên môn, kinh nghiệm để đánh giá thêm khả năng cluster có trong thực tế.

- Điểm Silhouette dưới 0.25, thì không nên tin tưởng cluster, và cần đi tìm nhiều bằng chứng khác[[1]](#footnote-1).

Bóng của mỗi trường hợp K khác nhau được biểu diễn trên biểu đồ với x là giá trị bóng của mỗi điểm, y thể hiện mật độ số điểm tương ứng với giá trị bóng đó.

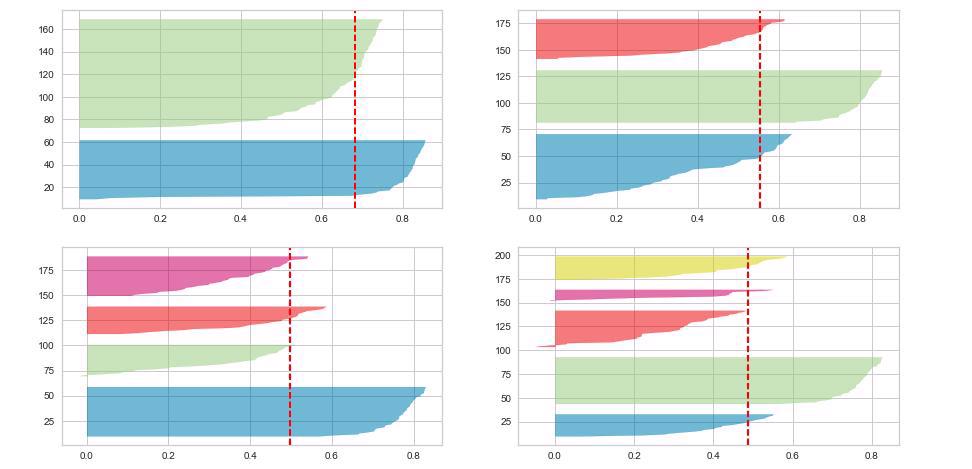
Quy trình triển khai Average silhouette method để tìm số cụm k tối ưu như sau:

1. Triển khai thuật toán phân cụm với các số cụm k thay đổi (ví dụ từ 1 đến 10)

2. Với mỗi giá trị k, tính giá trị average silhouette

3. Vẽ đồ thị average silhouette theo các giá trị k.

4. Vị trí có average silhouette lớn nhất là số cụm k cần tìm.



Hình : Minh họa phân tích Silhouette cho k cụm lần lượt là k=2 (góc trên bên trái), k=3 (góc trên bên phải), k=4 (góc dưới bên trái), k=5 (góc dưới bên phải).

Nhìn đồ thị có thể thấy giá trị k = 4 và k = 5 có vẻ không tối ưu đối với dữ liệu đã cho vì có điểm Silhouette dưới trung bình (vạch nét đứt đỏ) và kích thước ô hình bóng các cụm không đồng đều. Ngược lại, giá trị 2 và 3 cho k có vẻ là giá trị tối ưu. Điểm Silhouette cho mỗi cụm đều trên trung bình. Ngoài ra, độ dày của ô hình bóng đại diện cho mỗi cụm cũng là một điểm quyết định . Đối với ô có k = 3 (trên cùng bên phải), độ dày đồng đều hơn ô có k = 2 (trên cùng bên trái) với độ dày của một cụm nhiều hơn nhiều so với độ dày khác. Do đó, người ta có thể chọn số lượng cụm tối ưu là 3.

## 4.3. Ứng dụng K-Means vào bài toán phân khúc khách hàng

### 4.3.1. Bài toán phân khúc khách hàng

Bài toán phân cụm có đầu vào là tập dữ liệu không có nhãn và đầu ra là các cụm dữ liệu đã được phân chia. Do đó, mục tiêu của bài toán phân cụm là những cụm dữ liệu được phân chia bởi thuật toán. Trong các mô hình kinh doanh, doanh nghiệp sẽ chia nhỏ tệp khách hàng ra thành những nhóm đối tượng khác nhau để có thể áp dụng những chiến lược kinh doanh cụ thể cho từng nhóm đối tượng. Điều này giúp cho khách hàng được tiếp cận với các sản phẩm thật sự phù hợp với bản thân họ. Sự phù hợp đó sẽ kéo doanh thu và lợi nhuận tăng lên. Vấn đề đặt ra là làm sao có thể chia nhỏ tệp khách hàng đó ra khi mà số lượng đơn hàng là rất lớn và chúng ta không thể ngồi để phân tích từng vị khách.

Mục tiêu của các thuật toán phân cụm là từ tập dữ liệu khổng lồ đó, làm sao để biết được có những nhóm dữ liệu đặc trưng nào trong đó hay từng dữ liệu trong đó thuộc vào nhóm nào? Đó là điều mà thuật toán phân cụm sẽ đi tìm câu trả lời. Mục tiêu của dự án này là mô tả tốt nhất sự thay đổi của các kiểu khách hàng khác nhau (types of customers) mà nhà phân phối bán buôn tương tác. Khi đó nhà phân phối sẽ có được cái nhìn sâu sắc về cách cấu trúc tốt nhất dịch vụ giao hàng của họ để đáp ứng nhu cầu của từng khách hàng.

### 4.3.2. Dữ liệu nghiên cứu

Tập dữ liệu nghiên cứu được lấy từ UCI Machine Learning Repository, thể hiện khách hàng của một nhà phân phối bán buôn. Nó bao gồm chi tiêu hàng năm (annual) tính bằng đơn vị tiền tệ (đvtt) cho nhiều danh mục sản phẩm đa dạng. Các thông tin được tập hợp thành 440 bản ghi và 8 cột: Channel, Region, Fresh, Milk, Grocery, Frozen, Detergents\_Paper, Delicassen, trong đó:

- FRESH: chi tiêu hàng năm cho các sản phẩm tươi sống;

- MILK: chi tiêu hàng năm cho các sản phẩm sữa;

- GROCERY: chi tiêu hàng năm cho các sản phẩm tạp hóa;

- FROZEN: chi tiêu hàng năm cho các sản phẩm đông lạnh;

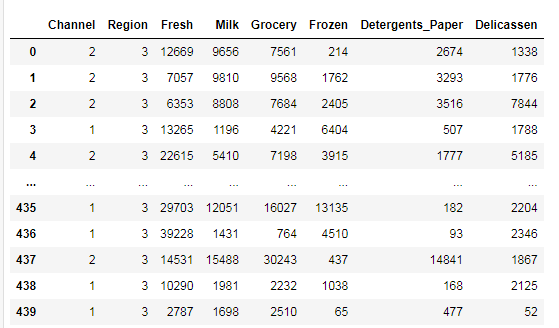
- DETERGENTS\_PAPER: chi tiêu hàng năm cho chất tẩy rửa và sản phẩm giấy;

- DELICASSEN: chi tiêu hàng năm cho và các sản phẩm đồ ăn nhanh;

- CHANNEL: Kênh khách hàng;

- REGION: Khu vực khách hàng.

Dưới đây là mô tả một phần dữ liệu.



Hình : Mô tả một phần dữ liệu Wholesale customers.

### 4.3.2. Các công nghệ sử dụng trong nghiên cứu bài toán

Nghiên cứu được thực hiện thông qua ngôn ngữ lập trình Python, với các thư viện hỗ trợ như:

- **Pandas**: viết tắt cho panel data (bảng dữ liệu), là một thư viện mã nguồn mở giúp phân tích và thao tác dữ liệu nhanh, mạnh mẽ, linh hoạt và dễ sử dụng, được xây dựng dựa trên ngôn ngữ lập trình Python.

- **Numpy**: là một thư viện lõi phục vụ cho khoa học máy tính của Python, hỗ trợ cho việc tính toán các mảng nhiều chiều, có kích thước lớn với các hàm đã được tối ưu áp dụng lên các mảng nhiều chiều đó.

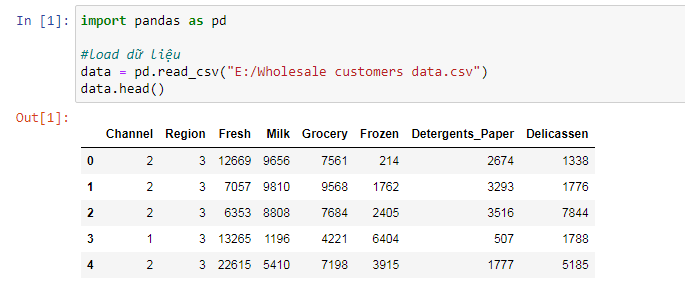
- **Matplotlib**: là một thư viện vẽ đồ thị rất mạnh mẽ hữu ích cho những người làm việc với Python, phục vụ công việc trực quan hóa dữ liệu.

- **Scikit-learn (sklearn):** là thư viện mạnh mẽ nhất dành cho các thuật toán học máy được viết trên ngôn ngữ Python. Thư viện này cung cấp một tập các công cụ xử lý các bài toán machine learning và statistical modeling gồm: classification, regression, clustering, và dimensionality reduction.

### 4.3.3. Kết quả nghiên cứu

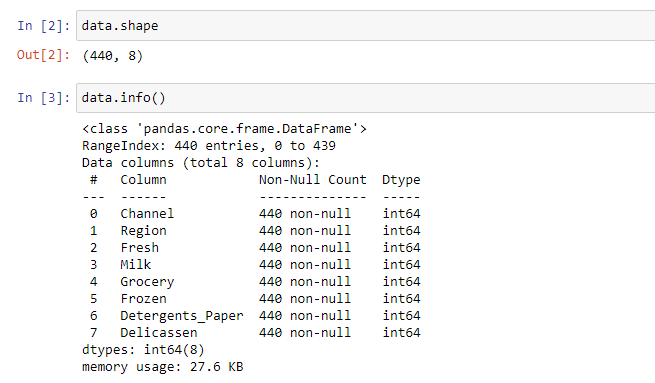
#### Tìm hiểu dữ liệu

Trước tiên, ta đọc dữ liệu và xem qua 5 hàng đầu tiên của dữ liệu:



Hình : Mô tả 5 hàng đầu tiên của dữ liệu Wholesale customers

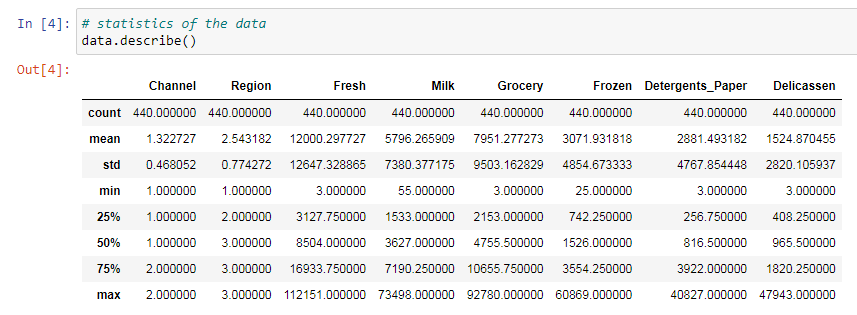
Đây là những thông tin chi tiết về chi tiêu của khách hàng đối với các sản phẩm khác nhau như Sữa, Hàng tạp hóa, Đồ đông lạnh, Chất tẩy rửa, v.v. Chúng ta sẽ phải phân khúc khách hàng dựa trên các thông tin được cung cấp.



Hình : Kích thước và thông tin về dữ liệu Wholesale customers.

Bộ dữ liệu có kích thước 440 bản ghi × 8 cột, trong đó có 2 cột về khu vực và 6 danh mục sản phẩm: 'Fresh', 'Milk', 'Grocery', 'Frozen', 'Detergents\_Paper' và 'Delicatessen'. Không có giá trị NULL, toàn bộ đều định dang số (numerical – int64).

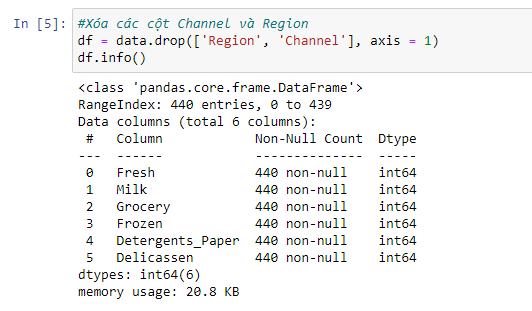
Một số số liệu về thống kê liên quan đến dữ liệu:



Hình : Thông tin thống kê của dữ liệu Wholesale.

#### Tiền xử lý dữ liệu

Thuật toán K-Means tiêu chuẩn không áp dụng trực tiếp cho dữ liệu phân loại. Cụ thể, không gian tìm kiếm của các biến Channel, Region là rời rạc (chỉ mang giá trị 1 hoặc 2), không cung cấp thông tin về chi tiêu, không phù hợp với mục tiêu bài toán đặt ra, nên cần phải loại bỏ các cột này.



Hình : Dữ liệu Wholesale sau khi tiền xử lý bước 1.

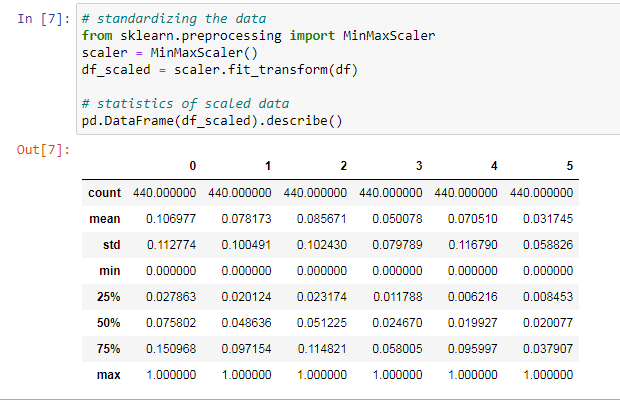
Sau khi xóa 2 cột Channel và Region, kích thước dữ liệu thu gọn thành 440×6.

Có thể thấy rằng độ lớn của dữ liệu không đồng đều do sự chênh lệch chi tiêu của người mua đối với các mặt hàng. Ví dụ, giá trị **mean** của mặt hàng Fresh lên tới khoảng 12000 đvtt trong khi Delicassen chỉ khoảng 1524 đvtt. Điều này tác động tới tính hiệu quả của thuật toán, ví dụ thời gian thực hiện, quá trình hội tụ, hay thậm chí ảnh hưởng cả tới độ chính xác của thuật toán. Chính vì vậy, phải tiến hành điều chỉnh dữ liệu để các đặc tính cùng có chung một tỉ lệ (data scaling). Và thường để các đặc tính có giá trị trong khoảng [0, 1], kết quả sẽ giúp cho thuật toán hội tụ nhanh.

Có nhiều cách để điều chỉnh dữ liệu, như chuẩn hóa (standardization), hoặc bình thường hóa (normalization). Trong báo cáo này sử dụng chuẩn hóa Min-Max, tích hợp trong MinMaxScaler() của gói preprocessing trong scikit-learn, một thư viện mã nguồn mở của Python. Chuẩn hóa Min-Max được thể hiện bởi công thức:

(6)

với x là giá trị ban đầu, x’ là giá trị sau khi chuẩn hóa, min(x) là giá trị nhỏ nhất của đặc trưng và max(x) là giá trị lớn nhất của đặc trưng.



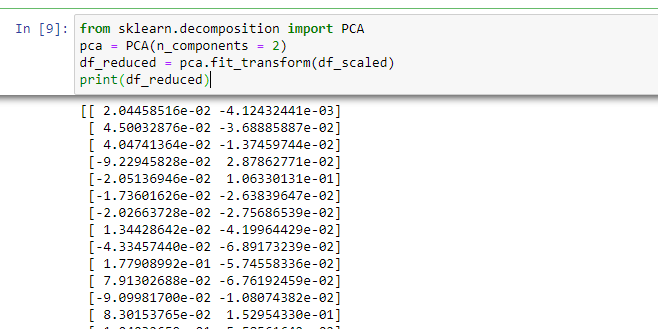
Hình : Dữ liệu thống kê sau khi chuẩn hóa.

Sau khi scale thì các giá trị cũng đã khá tương đồng nhau do đều được đưa về khoảng giá trị [0,1], với min = 0 và max = 1.

#### Giảm chiều dữ liệu

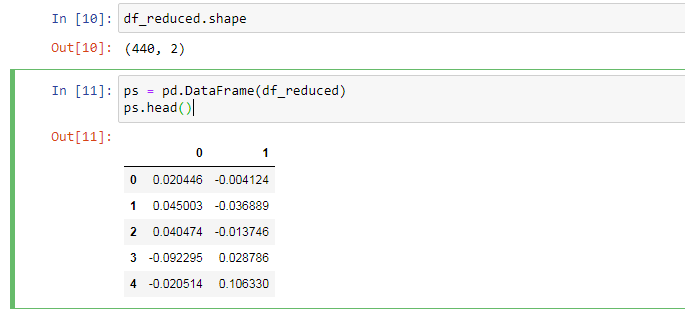
Chúng ta có thể dễ dàng trực quan hóa các cụm trong không gian 1 chiều, 2 chiều hay thậm chí là 3 chiều. Tuy nhiên với bộ dữ liệu 6 cột tương đương với 6 chiều dữ liệu thì đây thật sự là một bài toán khó. Do đó phải áp dụng một phương pháp giảm chiều dữ liệu để dễ dàng trực quan hơn. Principal Component Analysis (PCA) là một trong số các phương pháp như thế. Ý tưởng cơ bản là nó cho phép chúng ta giảm kích thước (các đặc trưng) của tập dữ liệu xuống bất kỳ con số nào nhỏ hơn số lượng đặc trưng hiện tại. Một công dụng khác của PCA là nén dữ liệu làm tiết kiệm thời gian tính toán, do kích thước ít hơn dẫn đến mất ít thời gian hơn để train và test mô hình.

Sau khi import PCA từ thư viện sklearn, tạo một mô hình PCA và truyền tham số là số thành phần n\_component là 2 (chiều), rồi fit vào df\_scaled ở bước scaling.



Hình : Dữ liệu Wholesale sau khi giảm chiều.

Đúng như lý thuyết, kích thước dữ liệu sau khi áp dụng giảm chiều đã giảm xuống còn 440×2, các phần tử của dữ liệu theo đó cũng chuyển thành véctơ 2 chiều.

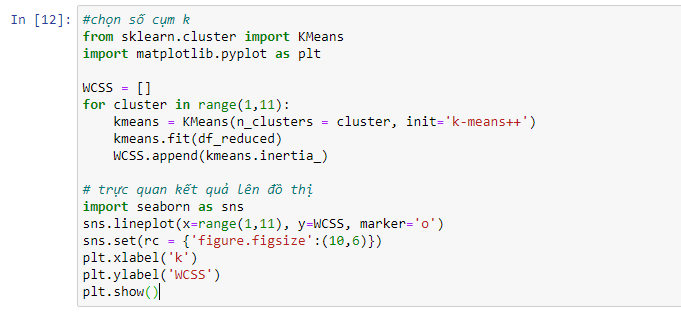


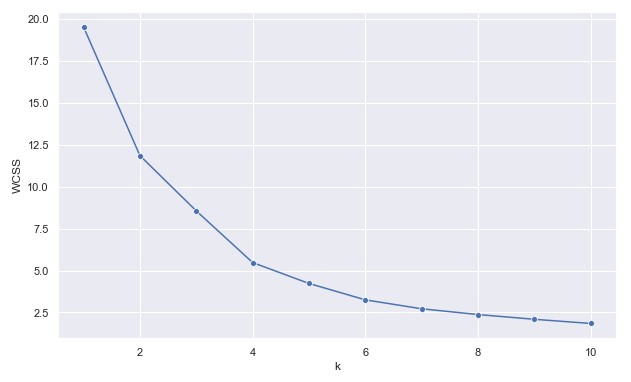
Hình : Một phần dữ liệu Wholesale sau khi giảm chiều.

Bước tiếp theo mô hình sẽ được huấn luyện dựa trên các đặc trưng mới do PCA() tạo ra.

#### Triển khai K-Means với bộ dữ liệu

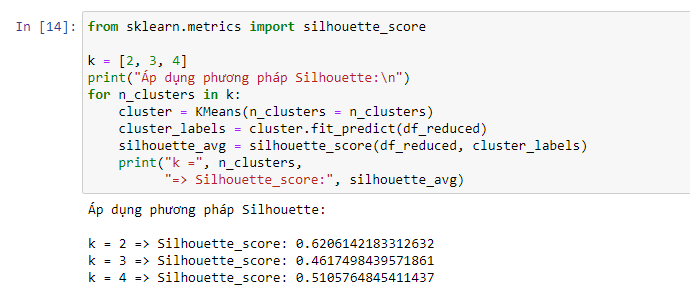
Chọn ra số lượng cluster tối ưu bằng phương pháp Elbow.





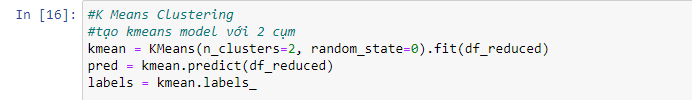
Hình : Đồ thị Elbow chọn ra số lượng cụm tối ưu

Ý tưởng ở đây là chọn một số cụm mà chúng ta không thấy nhiều sự khác biệt đối với WCSS. Nhìn vào điểm khuỷu tay ở đồ thị trên, chúng ta có thể chọn bất kỳ số lượng cụm nào trong khoảng từ 2 đến 4, vì khi số lượng cụm lớn hơn 2 thì tốc độ suy giảm của hàm biến dạng dường như không đáng kể so với trước đó. Để khẳng định số cụm tối ưu nhất, áp dụng phương pháp Silhouette cho số cụm là 2, 3 hoặc 4.



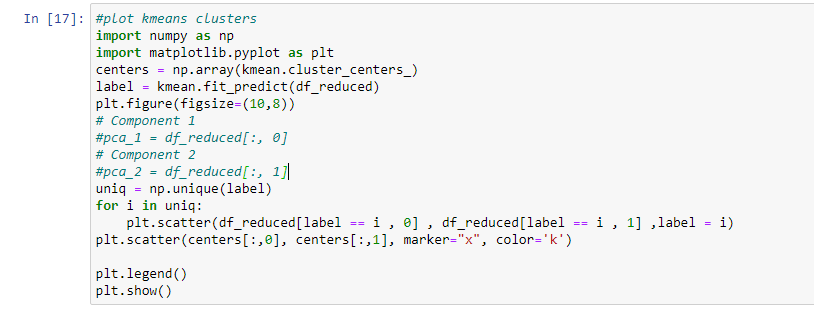
Hình : Số điểm Silhouette cho từng cụm.

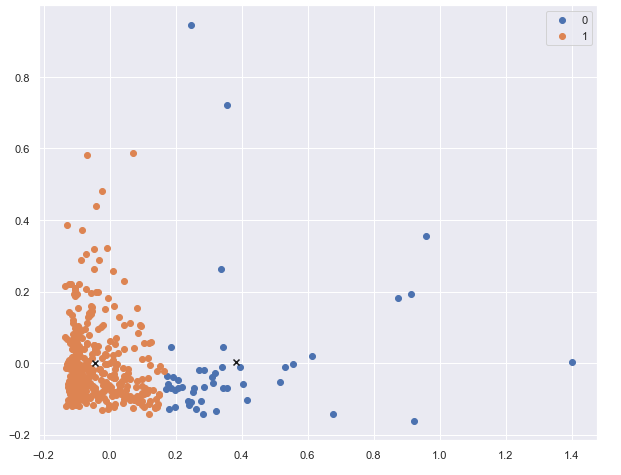
Số cụm k=2 có Silhouette Score cao hơn cả (0.6206), nên ta chọn số cụm tối ưu nhất là 2, sau đó fit vào mô hình:



Một hạn chế của thuật toán K-Means là nó nhạy cảm với việc khởi tạo các trọng tâm hoặc các điểm trung bình. Vì vậy, nếu khởi tạo một điểm trọng tâm ở “xa” vị trí thật sự của nó, có thể cuối cùng sẽ không có điểm nào được liên kết với điểm đó và đồng thời, nhiều hơn một điểm trọng tâm có thể được khởi tạo vào cùng một cụm dẫn đến phân cụm kém và ngược lại. Để khắc phục nhược điểm này, chúng ta sử dụng K-means++. Thuật toán này đảm bảo khởi tạo centroid thông minh hơn và tăng tốc độ hội tụ của thuật toán, cải thiện chất lượng của phân cụm. Ngoài phần khởi tạo sử dụng K-Means++, phần còn lại của thuật toán sẽ vẫn là thuật toán K-means tiêu chuẩn. Do đó trong đoạn code trên tồn tại tham số “init=’k-means++’”.

Để có thể nhìn rõ sự phân bố của các phần tử dữ liệu và các cụm, cần phải trực quan hóa các cụm mà K-Means tạo ra. Vì chỉ có 2 thành phần chính, chúng ta sẽ có một biểu đồ 2 chiều với 2 cụm được biểu diễn cùng với các các trọng tâm bởi thuộc tính “cluster\_centers\_”.



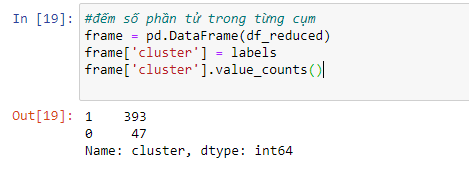


Hình : Trực quan hóa 2 cụm dữ liệu Wholesale

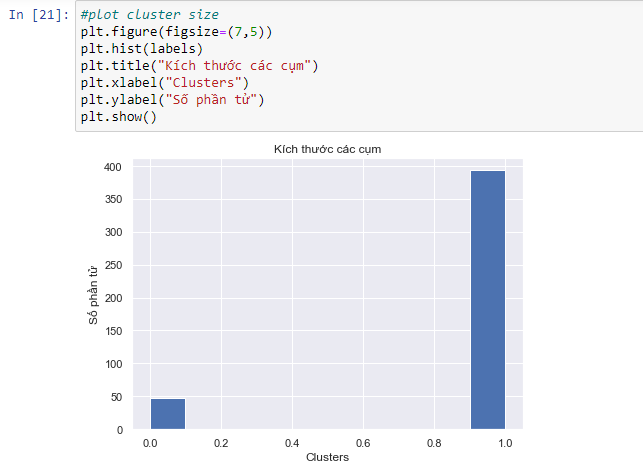
#### Quan sát và phân tích kết quả

##### Đánh giá từ kết quả phân cụm

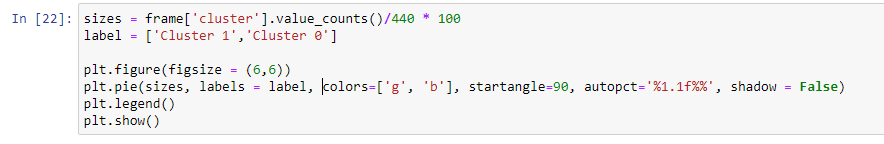
Xem mỗi cụm có bao nhiêu phần tử bằng cách sử dụng value\_counts().

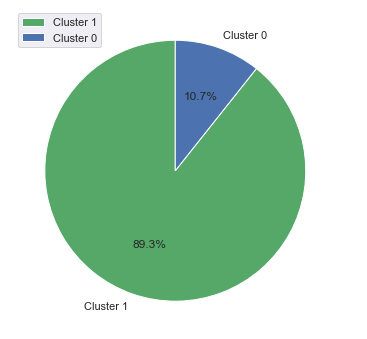


Như vậy, có 393 điểm dữ liệu thuộc cluster 1 và 47 điểm thuộc cluster 0. Có thể biểu diễn dưới dạng biểu đồ cột cho trực quan:



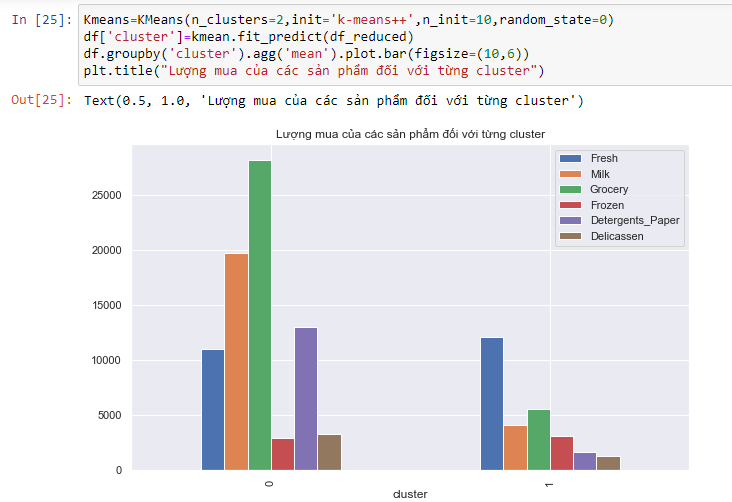
Hình : Biểu đồ cột thể hiện kích thước các cụm





Trong đó, tỷ lệ đối tượng thuộc cụm 1 chiếm 89.3% tổng số giá trị quan sát, số còn lại thuộc cụm 0, chiếm thiểu số (10.7%).

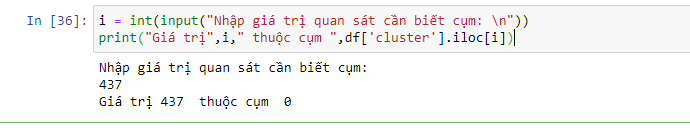
Hình : Biểu đồ tròn thể hiện tỷ lệ giữa các cụm



Hình : Lượng chi tiêu cho các sản phẩm đối với từng cụm

Qua biểu đồ cột trên có thể thấy rằng, cụm 0 dành chi tiêu cao hơn cho các sản phẩm Milk, Grocery, Detergents\_Paper, trong khi cụm 1 chi tiêu nhỉnh hơn cho các sản phẩm Fresh và Frozen so với cụm 0. Điều này cho thấy cụm 0 có thể là các nhà phân phối số lượng lớn như các đại lý, cửa hàng tiện lợi hoặc siêu thị, trong khi với lượng tiêu dùng nêu trên thì có thể cụm 1 đại diện cho các nhà hàng ăn do tiêu thụ nhiều đồ tươi và đồ đông lạnh.

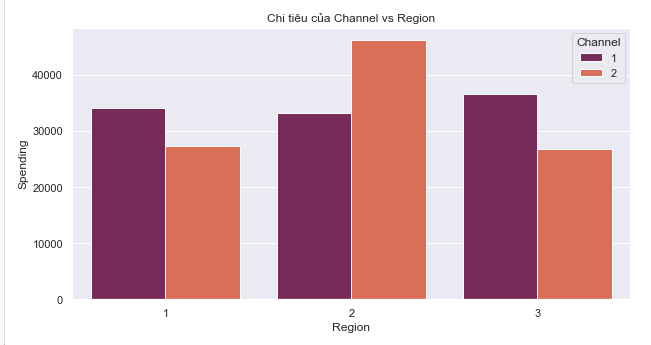
Ngoài ra cũng có thể xem giá trị quan sát ở vị trí nào đó thuộc cụm nào bằng cách sử dụng trình truy cập (accessor) .iloc[] để truy xuất một hàng hoặc cột nào đó theo chỉ số nguyên được nhập vào từ bàn phím:



##### Đánh giá chung

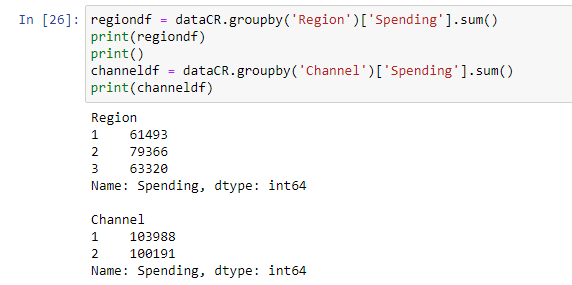
Để đánh giá lượng tiêu thụ từ các kênh khách hàng và khu vực khách hàng, tạo 01 cột mới đặt tên “Spending” thể hiện tổng chi tiêu cho các danh mục sản phẩm. Từ biểu đồ cột biểu diễn chi tiêu của Channel và Region, chúng ta có thể đánh giá xem kênh khách hàng hay vùng nào chi tiêu nhiều hơn.





Hình : Chi tiêu của các kênh và vùng khách hàng

Xét trên biểu đồ và số liệu cụ thể bên dưới, vùng 2 chi tiêu nhiều cho các sản phẩm nhất (79366 đvtt), trong đó kênh 2 ở vùng này dẫn đầu các kênh với chi tiêu khoảng 45000. Ở cả 3 vùng, các kênh 1 chi tiêu khá đồng đều nhau. Cả kênh 1 và 2 có số liệu chi tiêu khá xấp xỉ nhau, tuy nhiên kênh 1 có nhỉnh hơn một chút.



Bản đồ nhiệt (heatmap) biểu diễn sự tương quan giữa các đặc trưng dựa trên hệ số tương quan Pearson (r) của hai mẫu và :

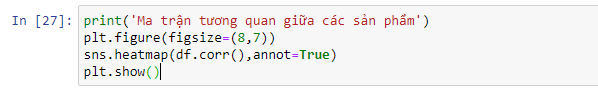
(7)

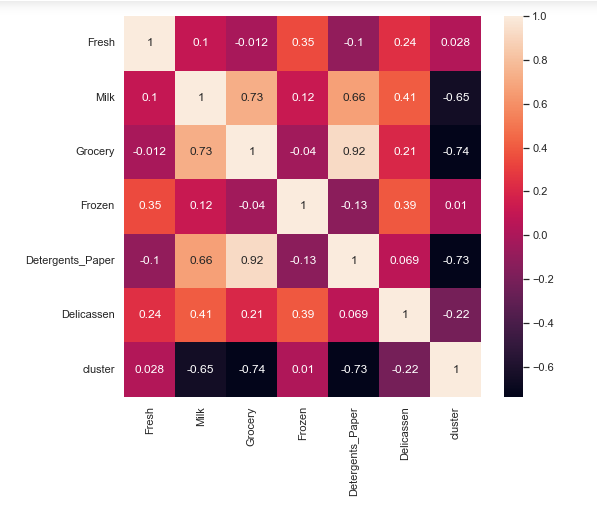
Trong đó, và .

- Nếu |r| nằm trong đoạn [0,5; 1], thì nó được cho là tương quan mạnh (nếu r nằm trong đoạn [-1; -0.5], thì tương quan theo chiều tuyến tính ngược, có nghĩa là nếu biến tăng thì giá trị hàm giảm và ngược lại; nếu r nằm trong đoạn [0.5; 1], thì tương quan theo chiều tuyến tính thuận, có nghĩa là nếu biến tăng thì giá trị hàm tăng và ngược lại).

- Nếu |r| nằm trong nửa khoảng [0.3; 0.5), thì nó được gọi là tương quan trung bình.

- Nếu |r| < 0.3, thì nó được gọi là một mối tương quan yếu.





Hình : Ma trận tương quan giữa các sản phẩm

Theo bản đồ này, có thể cố gắng xếp hạng các đặc trưng theo mối tương quan giữa chúng như sau:

- Bậc 1: Tương quan cao

* "Grocery" và "Detergents\_Paper" có tương quan cao với r = 0,92
* "Grocery" và "Milk" có mối tương quan cao với r = 0,73
* "Milk" và "Detergents\_Paper" có tương quan vừa phải với r = 0,66

- Bậc 2: Tương quan trung bình

* "Milk" và "Delicassen" có tương quan trung bình với r = 0,41
* "Delicassen" và "Frozen" có tương quan trung bình với r = 0,39
* "Frozen" và "Fresh" có tương quan trung bình với r = 0,35

Các cặp tính năng còn lại tương quan yếu hoặc gần như không tương quan. Từ ma trận tương quan này và kết quả phân cụm, có thể đưa ra kết luận rằng, các nhóm sản phẩm cùng bậc tương quan thì thường được mua cùng nhau.

Nhà phân phối bán buôn đang xem xét thay đổi dịch vụ giao hàng từ 5 ngày/tuần hiện tại thành 3 ngày/ tuần. Tuy nhiên, nhà phân phối sẽ chỉ thực hiện thay đổi này cho những khách hàng có phản ứng tích cực. Dựa vào phân khúc khách hàng, nhà phân phối có thể xác định những khách hàng có phản ứng tích cực với sự thay đổi trong dịch vụ giao hàng như sau:

Khách hàng ở cụm 1 đang chi tiêu nhiều hơn cho đồ đông lạnh (Frozen) và đồ tươi (Fresh) và rõ ràng họ sẽ rất hài lòng nếu được giao hàng hàng ngày. Vì vậy, nếu nhà phân phối thay đổi dịch vụ giao hàng từ 5 ngày/tuần thành 3 ngày/tuần, khách hàng của cụm 1 sẽ phản ứng rất nặng nề. Vì vậy, nếu nhà phân phối đang xem xét thay đổi dịch vụ giao hàng từ 5 ngày thành 3 ngày/tuần thì trước tiên họ nên thử nghiệm dịch vụ này trên khách hàng cụm 0 và xem phản ứng của họ trước khi có thể áp dụng cho khách hàng cụm 1. Cách làm này sẽ giúp nhà phân phối hạn chế được lượng khách hàng phàn nàn và có thể áp dụng được cơ chế giao hàng mới cho những khách hàng phản ứng tích cực hoặc chí ít là không phản ứng tiêu cực.

# Phần V: Kết luận và đề nghị

## 5.1. Kết luận

Trong khoảng thời gian thực hiện khóa luận, em đã hoàn thiện báo cáo với các mục tiêu đặt ra ban đầu. Cụ thể khóa luận đã đạt được những kết quả sau:

1. Trình bày các kiến thức cơ bản về khai phá tri thức và khai phá dữ liệu. Sau đó giới thiệu phương pháp khai phá dữ liệu phân cụm, trình bày thuật toán phân cụm K-means.

2. Cài đặt và chạy thử nghiệm thuật toán K-means trên Jupyter Notebook của Python trên bộ dữ liệu wholesale customers từ UCI gồm 440 bản ghi và 8 thuộc tính.

3. Dựa vào mô hình phân cụm dữ liệu khách hàng để hỗ trợ nhà phân phối ra quyết định phân bổ các chính sách giao hàng cho từng nhóm khách hàng.

## 5.2. Đề nghị

1. Tiếp tục nghiên cứu các thuật toán khai phá dữ liệu phân cụm khác như thuật toán EM, FCM, DBSCAN, … để so sánh được hiệu quả của các thuật toán này khi áp dụng vào mô hình phân khúc khách hàng.

2. Tiếp tục nghiên cứu các công cụ khai phá dữ liệu khác như Oracle Data Mining, R, … và so sánh ưu nhược điểm của các công cụ này.

3. Bổ sung thêm dữ liệu tập huấn để mô hình phân cụm có độ tin cậy cao hơn và hoạt động hiệu quả hơn.

4. Tìm hiểu nhu cầu thực tế, từ đó cải tiến chương trình, cài đặt lại bài toán theo các thuật toán đã nghiên cứu để làm việc tốt hơn với các cơ sở dữ liệu lớn.

Phân cụm dữ liệu và ứng dụng của phân cụm dữ liệu là hướng nghiên cứu cần thiết, quan trọng. Tuy nhiên đây cũng là mảng rất rộng, bao hàm nhiều phương pháp, kỹ thuật, và hình thành nhiều nhóm khác nhau. Trong quá trình nghiên cứu, thực hiện luận văn mặc dù đã cố gắng tập trung nghiên cứu và tham khảo nhiều tài liệu, bài báo, tạp chí khoa học trong và ngoài nước, nhưng do trình độ còn có nhiều giới hạn, không thể tránh khỏi thiếu sót và hạn chế. Em rất mong được sự chỉ bảo đóng góp nhiều hơn nữa của các thầy, cô giáo, các nhà khoa học để dần tiến tới những dự án lớn hơn về xây dựng, phát triển các kỹ thuật, ứng dụng của phân cụm dữ liệu trong tương lai.

# Phần VI. Tài liệu tham khảo

**Tiếng Việt:**

[1]. Doan Nguyen (2018). K-Means Clustering [trực tuyến], truy cập 14/09/2021, từ: <<https://ml-for-vietnameses.readthedocs.io/vi/latest/k-means-clustering.html>>.

[2]. Thuật toán K-Means cho bài toán phân cụm khách hàng, Megastore handbook, truy cập 20/09/2021, từ: < <https://handbook.magestore.com/books/machine-learning-in-retail/page/thu%E1%BA%ADt-to%C3%A1n-k-means-cho-b%C3%A0i-to%C3%A1n-ph%C3%A2n-c%E1%BB%A5m-kh%C3%A1ch-h%C3%A0ng>>.

[3]. K-means Clustering, blog DataiSG, truy cập 20/09/2021, từ: <[https://dataisg.org/tutorial/machine-learning/k-means-clustering](https://dataisg.org/tutorial/machine-learning/k-means-clustering/)/>.

[4]. Các phương pháp đánh giá trong thuật toán Clustering, BigDataUni, truy cập 21/09/2021, từ < <https://bigdatauni.com/tin-tuc/cac-phuong-phap-danh-gia-trong-thuat-toan-clustering.html>>.

[5]. Pham Dinh Khanh (2021), Các bước của thuật toán k-Means Clustering [online], truy cập 10/09/2021, từ < <https://phamdinhkhanh.github.io/deepai-book/ch_ml/KMeans.html>>.

[6]. Phạm Kiên Trung\*, Nguyễn Đức Thắng, Lê Văn Chiến, Nguyễn Văn Thưởng (2020). Ứng dụng thuật toán K-Means trong phân cụm khách hàng mục tiêu. *Tạp chí Khoa học Kỹ thuật Mỏ - Địa chất Tập 61*, trang 145 – 150.

[7]. Phan Đức Việt (2016). Bức tranh tổng quan về thuật toán phân cụm [online], truy cập 10/08/2021, từ <<https://techmaster.vn/posts/33893/buc-tranh-tong-quan-ve-thuat-toan-phan-cum>>.

**Tiếng Anh:**

[1]. Dr. Michael J. Garbade (2018). Understanding K-means Clustering in Machine Learning [online], viewed 14 September 2021, from: <<https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1>>.

[2]. Juan E. Rolon (2017). Customer Segmentation Analysis via Unsupervised Learning [online], viewed 25 September 2021, from: <<http://www.juanrolon.com/customer-segmentation.html>>.

[3]. Abhay Gandotra (2017). Customer Segmentation using K-means in R [online], viewed 26 September 2021, from <<https://medium0.com/@abhay.gandotra/customer-segmentation-using-k-means-in-r-41a541dd5b57>>.

[4]. Mohammed Muzammil (2021). Understanding K – Means Clustering WIth Customer Segmentation Usecase [online], viewed 1 October 2021, from <<https://www.analyticsvidhya.com/blog/2021/07/understanding-k-means-clustering-using-customer-segmentation/>>.

[5]. Pranshu Sharma (2021). K Means Clustering Simplified in Python, viewed 13 September 2021, from <<https://www.analyticsvidhya.com/blog/2021/04/k-means-clustering-simplified-in-python/>>.

[6]. Rajesh Sharma (2021). Detect and Remove the Outliers using Python [online], viewed 23 September 2021, from < <https://www.geeksforgeeks.org/detect-and-remove-the-outliers-using-python/>>.

[7]. Daniel T. Larose (2015). Data Mining and Predictive Analytics. 2nd ed., John Wiley & Sons, Inc., Hoboken, New Jersey.

[8]. Ch. Seshadri Rao, B. Meena, G. Sri Devi (2020). Analyzing wholesale customer data with machine learning techniques. *Journal of Information and Computational Science, Volume 10 Issue 5*, page 502-508.

[9]. Meher Krishna Patel (2017). Clustering with dimensionality reduction, Machine Learning Guide.

1. Mục “Interpretation of Average Silhouette Value”, trang 855. [↑](#footnote-ref-1)