LỜI CAM ĐOAN

Tôi xin cam đoan:

1. Những nội dung trong luận văn này là do tôi thực hiện dưới sự hướng dẫn trực tiếp của Thầy **PGS. TS Nguyễn Đình Thuân**.
2. Mọi tham khảo trong luận văn đều được trích dẫn rõ ràng tên tác giả, tên công trình, thời gian công bố.

Mọi sao chép không hợp lệ, vi phạm quy chế đào tạo tôi xin chịu hoàn toàn trách nhiệm.

Tp. Hồ Chí Minh, ngày 15 tháng 07 năm 2022

Học viên thực hiện luận văn

Trần Thành Nguyên

LỜI CÁM ƠN

Em xin dành lời cảm ơn chân thành và sâu sắc nhất đến Thầy **PGS. TS Nguyễn Đình Thuân** người đã truyền cảm hứng về mảng khai phá dữ liệu, khuyến khích và chỉ dẫn tận tình cho em trong từng bước từ khi bắt đầu cho đến khi hoàn thành luận văn của mình.

Em cũng xin dành lời cảm ơn chân thành đến quý Thầy Cô Học viện Bưu Chính Viễn Thông Cơ Sở Thành Phố Hồ Chí Minh đã truyền đạt kiến thức vô cùng quý giá và tạo điều kiện thuận lợi cho em trong suốt thời gian học tập và nghiên cứu tại trường.

Tôi cũng xin chân thành cảm ơn Viễn thông Tây Ninh đã tạo điều kiện cho tôi tìm hiểu thông tin, cung cấp dữ liệu và hỗ trợ tôi trong suốt quá trình thực hiện luận văn.

Cuối cùng em xin gửi lời cám ơn đến Cha Mẹ, vợ con, gia đình, người thân, bạn bè và đồng nghiệp đã quan tâm, ủng hộ trong suốt quá trình học tập cao học.

Tp. Hồ Chí Minh, ngày 15 tháng 07 năm 2022

Học viên thực hiện luận văn

Trần Thành Nguyên

MỤC LỤC

DANH MỤC CÁC THUẬT NGỮ, CHỮ VIẾT TẮT

Viết tắt	Tiếng Anh	Tiếng Việt
CI	Cluster Index	Độ phụ thuộc
KPDL	Data Mining	Khai phá dữ liệu
CSDL	Database	Cơ sở dữ liệu
KPTT	Knowledge Discovery	Khám phá tri thức
CURE	Clustering Using REpresentatives	Phân cụm bằng cách sử dụng đại diện
BIRCH	Balance Iterative Reducing and Clustering using Hierarchies	Cân bằng Giảm lặp lại và Phân cụm bằng cách sử dụng Cấu trúc phân cấp
ROCK	Robust Clustering Algorithm for Categorical Attributes	Thuật toán phân cụm mạnh mẽ cho các thuộc tính phân loại

DANH SÁCH CÁC BẢNG

DANH SÁCH CÁC HÌNH VẼ VÀ ĐỒ THỊ

MỞ ĐẦU

Với sự bùng nổ công nghệ như hiện nay, có rất nhiều giải pháp công nghệ được nghiên cứu và triển khai nhằm phục vụ nhu cầu của cá nhân và doanh nghiệp. Trong đó Data Mining (Khai phá dữ liệu - KPDL) là một trong những lĩnh vực quan trọng nhất trong công nghệ. KPDL là quá trình chọn lọc, xử lý dữ liệu thô, sắp xếp, phân loại các tập hợp dữ liệu lớn qua đó để xác định các mẫu và xây dựng các mối quan hệ của dữ liệu để giải quyết các vấn đề bằng cách phân tích dữ liệu. Việc ứng dụng KPDL cho phép các đơn vị, doanh nghiệp có thể dự đoán trước được xu hướng trong tương lai.

Trong lĩnh vực viễn thông, một môi trường có nhiều sự cạnh tranh về số lượng thuê bao, chất lượng dịch vụ trong mảng di động (cuộc gọi thoại, sms, data…) như hiện nay. Các doanh nghiệp viễn thông cần phải nhanh chóng ứng dụng các giải pháp mới, và nhất là khai phá dữ liệu trên tập hành vi sử dụng dịch vụ di động của khách hàng để hoạch định rõ các chiến lược kinh doanh khác nhau trên từng tập khách hàng.

Trong bối cảnh hiện tại, các công ty nhận thấy rằng họ phải có được cái nhìn "từ toàn cảnh đến chi tiết" về khách hàng của mình từ nhu cầu, sở thích, hành vi, thái độ, nhận thức, … của khách hàng. Sau đó, các hoạt động sản xuất, sales, marketing phải tinh chỉnh sao cho thỏa mãn các nhu cầu của khách hàng. Đây sẽ là lợi thế cạnh tranh cần hướng đến.

Vấn đề đặt ra là đối với từng nhóm khách hàng cụ thể, các doanh nghiệp viễn thông cần có cơ chế, chính sách, và chiến lược kinh doanh khác nhau để giữ chân, và đáp ứng được nhu cầu sử dụng dịch vụ của từng nhóm khách hàng để mang lại chất lượng phục vụ tốt nhất cho từng nhóm khách hàng.

Là một người đang công tác trong lĩnh vực viễn thông, vì vậy để hỗ trợ cho công việc hiện tại, và để giúp công ty xác định rõ từng phân khúc khách hàng sử

dụng dịch vụ di động của Vinaphone Tây Ninh. Nên xin đề xuất đề tài nghiên cứu về "Xác định số cụm tối ưu vào bài toán phân khúc khách hàng sử dụng dịch vụ di động tại VNPT Tây Ninh".

Ngành viễn thông và thông tin di động là một trong các ngành nghề kinh tế - kỹ thuật quan trọng của đất nước nhằm đảm bảo an ninh thông tin quốc phòng của quốc gia. Trong một môi trường cạnh tranh khốc liệt giữa các nhà cung cấp mạng di động như hiện nay, để đáp ứng được các loại sản phẩm, dịch vụ thích hợp tới từng khách hàng thì các nhà quản lý tiếp thị cần phải xác định được những phân khúc khách hàng và mục tiêu cốt lõi mà doanh nghiệp muốn thu hút khách hàng.

Khi mà phân khúc khách hàng hiệu quả thì qua đó doanh nghiệp có thể dễ dàng giới thiệu, khuyến nghị, tiếp thị các sản phẩm, dịch vụ phù hợp nhất với những nhu cầu, mong muốn đối với từng nhóm khách hàng.

Do đó mục tiêu chính của bài luận này là tìm hiểu các thuật toán phân cụm, các phương pháp xác định số cụm tối ưu và sau đó ứng dụng vào bài toán phân khúc khách hàng sử dụng dịch vụ di động tại Vinaphone Tây Ninh. Các nội dung cụ thể của đề tài bao gồm:

- Nghiên cứu các bài báo về bài toán phân cụm.
- Nghiên cứu các tài liệu về thuật toán phân cụm: K-means, K-medoids.
- Nghiên cứu các toán về lựa chọn số cụm tối ưu: Elbow method, Average silhouette method.
- Nghiên cứu các bài báo, thuật toán về các phương pháp đánh giá số lượng cụm: Độ đo bóng (Silhouette), Độ đo Davies – Bouldin, Độ đo Dunn.
- Ứng dụng các thuật toán vào tập dữ liệu khách hàng sử dụng dịch vụ di động tại Vinaphone Tây Ninh, tiến hành đánh giá và chọn phân khúc khách hàng tối ưu nhất.

- Tổng kết các kết quả nghiên cứu liên quan trước đây và sau đó đánh giá hiệu quả của các phương pháp. Tiến hành áp dụng thực tế để kiểm tra và đánh giá kết quả.

Nội dung đề tài bao gồm 4 chương:

- **Chương 1**: Tổng quan
- **Chương 2**: Cơ sở lý luận
- **Chương 3**: Áp dụng các thuật toán xác định số cụm tối ưu vào bài toán phân khúc khách hàng sử dụng dịch vụ di động tại VNPT Tây Ninh
- **Chương 4**: Đánh giá - kết luận và hướng phát triển

Chương 1: TỔNG QUAN

Nội dung ở chương này sẽ xoay quanh chủ đề về bài toán phân khúc khách hàng tại VNPT Tây Ninh, nêu được một cách tổng quan về các phương pháp nghiên cứu cũng như các đối tượng cần nghiên cứu trong luận văn, và quan trọng là trả lời được câu hỏi tại sao cần phải xác định số cụm tối ưu trong bài toán phân khúc khách hàng. Các nội dụng sẽ trình bày bao gồm:

- Tổng quan về bài toán phân khúc khách hàng sử dụng dịch vụ di động.
- Tại sao phải xác định số cụm tối ưu vào bài toán phân khúc khách hàng.
- Các đối tượng trong phạm vi nghiên cứu.
- Các phương pháp nghiên cứu bài toán phân khúc khách hàng.

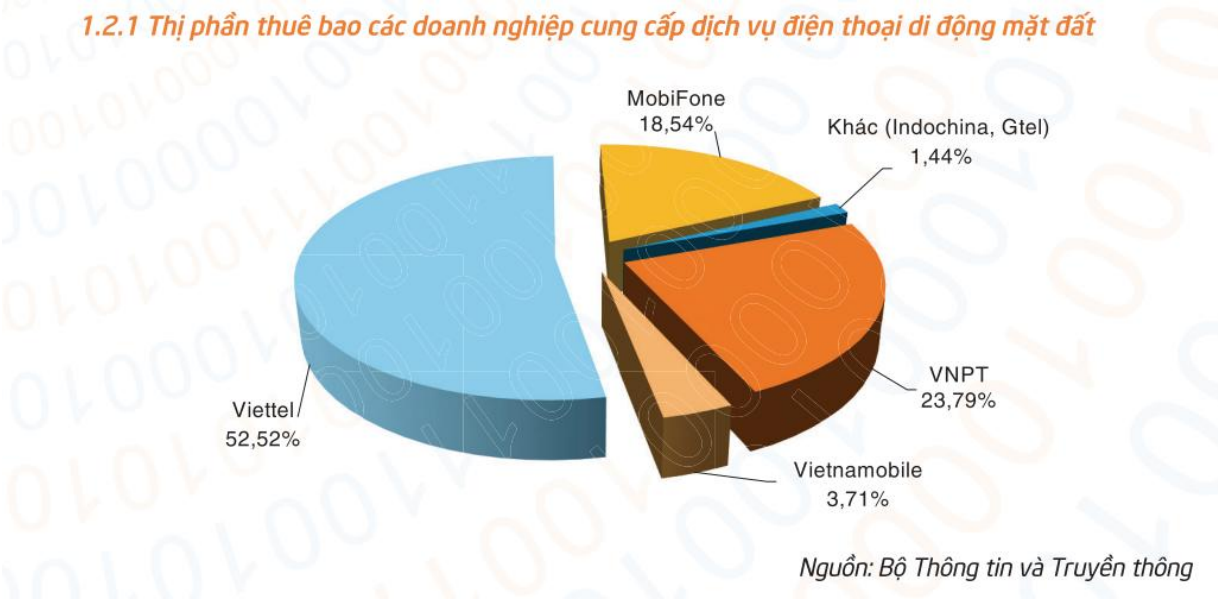
1.1 Bài toán phân khúc khách hàng dựa trên hành vi sử dụng dịch vụ di động

VNPT hiện là Tập đoàn Bưu chính Viễn thông hàng đầu tại Việt Nam được thành lập vào năm 1996, Công ty Dịch vụ Viễn thông là một công ty trực thuộc Tập đoàn Bưu chính Viễn thông Việt Nam (VNPT) hoạt động trong lĩnh vực thông tin di động, cung cấp các dịch vụ GSM, 3G, 4G, nhắn tin,... và nhiều lĩnh vực về công nghệ thông tin khác, và có tên cho mảng dịch vụ di động là Vinaphone. VNPT là một công ty cung cấp dịch vụ về mạng viễn thông và di động đầu tiên ở Việt Nam, có thể nói VNPT đã đặt nền móng cho sự phát triển chung của ngành, và qua đó cũng đóng góp vai trò then chốt trong việc đưa Việt Nam trở thành 1 trong 10 quốc gia có tốc độ phát triển Bưu chính Viễn thông nhanh nhất toàn cầu.

Tuy vậy, ở thời điểm mới xuất hiện thì thông tin di động vẫn còn là khái niệm xa lạ đối với đa số người tiêu dùng, số lượng thuê bao của mạng di động này không nhiều do vùng phủ sóng hạn chế(độ phủ của các trạm BTS còn ít) và giá cước cũng như thiết bị đầu cuối(điện thoại di động) còn đắt đỏ.

Kể từ khi Viettel bắt đầu tham gia cung cấp dịch vụ thông tin di động vào năm 2004 thì sự bùng nổ của thị trường thông tin di động Việt Nam mới bắt đầu diễn ra. Và

nhờ sự cạnh tranh đó giá cước di động Việt Nam đã giảm hơn 3 lần trong 20 năm qua. Kết quả của việc cạnh tranh khốc liệt giữa các nhà mạng đã giúp cho Việt Nam trở thành nước có mức cước thuộc hàng rẻ nhất thế giới, mạng lại lợi ích cho người tiêu dùng.



Hình 1.1: Thị phần viễn thông Việt Nam tính đến năm 2021(Nguồn: Sách Trắng công nghệ thông tin và Truyền thông 2021)[1]

Chính vì sự cạnh tranh khốc liệt, và tỷ lệ rời dịch vụ của khách hàng ngày càng có xu hướng tăng và việc giữ chân khách hàng khó khăn hơn trước, các công ty mạng viễn thông ngày nay phải liên tục phát triển các dịch vụ, sản phẩm mới một cách linh hoạt để đáp ứng các nhu cầu thay đổi liên tục của khách hàng. Các doanh nghiệp viễn thông cần phải nhanh chóng ứng dụng các giải pháp mới, và nhất là khai phá dữ liệu trên tập hành vi sử dụng dịch vụ di động của khách hàng để hoạch định rõ các chiến lược kinh doanh khác nhau trên từng tập khách hàng. Một trong các công cụ được sử dụng đó là phân khúc khách hàng.

Và vì thế "Phân khúc khách hàng" được coi là một công cụ marketing mang tính "khác biệt". Nó cho phép các tổ chức hiểu hơn về khách hàng của mình xây dựng các

chiến lược marketing, sales "khác biệt" theo các đặc điểm, tính chất, hành vi của từng khách hàng[2].

Hiện nay có nhiều phương pháp để phân khúc khách hàng như:

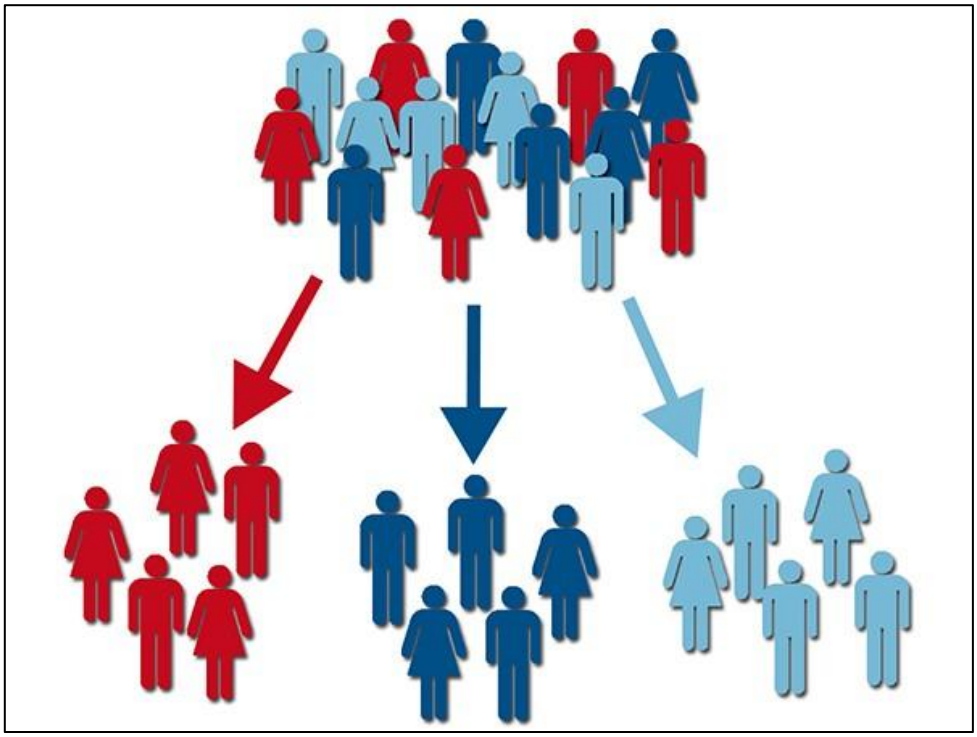
- Phân khúc theo nhân khẩu học: giới tính, tuổi, nghề nghiệp, trình độ học vấn, tình trạng hôn nhân, thu nhập hàng tháng, tình trạng bất động sản

- Phân khúc theo địa lý

- Phân khúc theo hành vi (sử dụng và thanh toán): phân khúc theo lần gần nhất sử dụng dịch vụ, số lần sử dụng trong tuần, tháng, số tiền mỗi lần sử dụng hay tổng số tiền sử dụng trong tháng ,…

- Phân khúc theo giá trị: phân khúc theo giá trị trung bình mỗi lần sử dụng dịch vụ, tổng giá trị sử dụng dịch vụ mỗi tháng; số dư trong tài khoản;…

- Phân khúc theo nhu cầu sử dụng: phân khúc theo các nhu cầu sử dịch dịch vụ gọi thoại, sms, data, các dịch vụ gia tăng,…



Hình 1.2: Phân khúc khách hàng

Ngành viễn thông không có đủ thông tin khách hàng cá nhân hay dữ liệu nhân khẩu học dồi dào. Vì thế, luận văn này chỉ tập trung vào phân khúc theo này vi sử dụng dịch vụ, và phân khúc theo giá trị mỗi lần sử dụng dịch vụ của khách hàng

1.2 Tại sao cần xác định số cụm tối ưu vào bài toán phân khúc khách hàng

1.2.1 Tại sao phải phân khúc khách hàng

Trong lĩnh vực viễn thông khi sử dụng một phương pháp tiếp thị, ưu đãi, chính sách khuyến mãi chung chung cho tất cả các khách hàng, cho dù đó là chiến lược thông minh nhất thì cũng có thể không mang lại kết quả như mong muốn. Bất kể nỗ lực tiếp thị của doanh nghiệp có hiệu quả đến đâu đối với một số khách hàng, chúng vẫn có thể thất bại khi áp dụng với những người khác. Đây là lúc doanh nghiệp cần áp dụng phân khúc khách hàng. Nếu làm đúng, nó có thể mang lại các lợi ích sau cho doanh nghiệp:

- Các chiến dịch tiếp thị tốt hơn: Phân khúc khách hàng cho phép các doanh nghiệp tạo ra các thông điệp tiếp thị tập trung hơn, tùy chỉnh cho từng phân khúc cụ thể.

- Các đề xuất cải tiến: Có ý tưởng rõ ràng về đối tượng khách hàng và họ muốn nhận được gì khi sử dụng sản phẩm/dịch vụ của bạn. Nó cho phép bạn tinh chỉnh và tối ưu hóa các dịch vụ. Nhờ đó, bạn có thể đáp ứng nhu cầu và mong đợi của khách hàng, từ đó cải thiện sự hài lòng của khách hàng.

- Khả năng mở rộng: Hãy phân khúc khách hàng tiềm năng và khách hàng hiện tại thành các nhóm nhỏ cụ thể. Nhờ đó, doanh nghiệp có thể hiểu rõ hơn về những điều khách hàng có thể quan tâm. Điều này sẽ thúc đẩy việc mở rộng các sản phẩm và dịch vụ mới sao cho phù hợp với đối tượng mục tiêu của doanh nghiệp.

- Giữ chân được nhiều khách hàng hơn: Phân khúc khách hàng có thể giúp doanh nghiệp phát triển những chiến lược giữ chân khách hàng mục tiêu tốt hơn bằng cách xác định những khách hàng trả tiền nhiều nhất của công ty. Từ đó, tạo phiếu mua hàng được cá nhân hóa cho họ hoặc thu hút lại những người đã không mua hàng khá lâu.

- Tối ưu hóa giá cả: Xác định tình trạng xã hội và tài chính của khách hàng. Nó giúp doanh nghiệp dễ dàng định giá phù hợp cho các sản phẩm/dịch vụ mà khách hàng của họ cho là hợp lý.

- Tăng doanh thu: Dành ít thời gian, nguồn lực và nỗ lực tiếp thị vào các phân khúc khách hàng ít sinh lời và dành thêm thời gian vào các phân khúc khách hàng thành công nhất của công ty. Kết quả là, nó làm tăng doanh thu, lợi nhuận cũng như giảm chi phí bán hàng cho doanh nghiệp.

1.2.2 Tại sao phải xác định số cụm tối ưu cho bài toán phân khúc khách hàng

Khi lựa chọn được số lượng phân khúc khách hàng(số cụm tối ưu) đủ tốt sẽ giúp doanh nghiệp giảm chi phí cho các phương pháp tiếp thị, bán hàng. Qua đó, cũng làm tăng thêm doanh thu cũng như lợi nhuận cho doanh nghiệp.

Ngoài ra khi chọn được số phân khúc khách hàng tốt sẽ giúp cho doanh nghiệp giảm bớt thời gian, tập trung được tối đa nguồn lực và phân bổ chi phí một cách hợp lý nhất vào các tập khách hàng tiềm năng.

Sau đây là hệ quả doanh nghiệp sẽ gặp phải phải nếu chọn số lượng phân khúc khách hàng không tối ưu:

- Trường hợp nếu số lượng phân khúc khách hàng quá ít sẽ làm cho doanh nghiệp sẽ phải tiếp cận với tập khách hàng quá lớn, và điều đó sẽ làm tăng chi phí tiếp thị, chính sách, ưu đãi...

- Trường hợp nếu chọn số lượng phân khúc khách hàng quá nhiều: Sẽ làm cho tập khách hàng tiềm năng bị băm nhỏ, làm tăng thời gian tiếp thị cũng như thời gian làm chính sách đối với từng.

1.3 Đối tượng và phạm vi nghiên cứu

Đối tượng nghiên cứu:

- Tập dữ liệu khách hàng sử dụng dịch vụ di động
- Các bài toán phân cụm
- Các bài toán về xác định số cụm tối ưu.

Phạm vi nghiên cứu:

- Đề tài được thực hiện trong phạm vi trên tập dữ liệu khách hàng sử dụng dịch vụ di động của VNPT Tây Ninh

- Các giải thuật phân cụm trong khai phá dữ liệu

- Các thuật toán về xác định số cụm tối ưu

1.4 Phương pháp nghiên cứu

Nghiên cứu các tài liệu, ứng dụng các mô hình lý thuyết và chứng minh bằng thực nghiệm:

- Nghiên cứu các bài báo về bài toán phân cụm

- Nghiên cứu các tài liệu về thuật toán phân cụm: K-means[3], K-medoids[4]

- Nghiên cứu các toán về lựa chọn số cụm tối ưu: Elbow method[5], Average silhouette method.

- Nghiên cứu các học thuật, các bài báo, luận văn về các phương pháp đánh giá số lượng cụm: Độ đo bóng (Silhouette), Độ đo Davies – Bouldin, Độ đo Dunn.

- Ứng dụng các thuật toán vào tập dữ liệu khách hàng sử dụng dịch vụ di động tại Vinaphone Tây Ninh, tiến hành đánh giá và chọn phân khúc khách hàng tối ưu nhất. Tổng kết các kết quả nghiên cứu liên quan trước đây và đánh giá hiệu quả của từng phương pháp. Tiến hành thực nghiệm để kiểm tra và đánh giá kết quả.

Chương 2: CƠ SỞ LÝ LUẬN

Chương này sẽ giới thiệu các kiến thức và nội dung, khái niệm cơ bản về khám phá tri thức và KPDL. Đây là các kiến thức và nền tảng cơ bản để phục vụ cho việc tìm hiểu và xây dựng hệ thống KPDL. Các nội dung cụ thể bao gồm: các giai đoạn của quá trình khám phá tri thức, các công đoạn của quá trình KPDL, các phương pháp KPDL và các kỹ thuật thường áp dụng trong KPDL. Ngoài ra, nội dung chương cũng đi sâu vào giới thiệu về phân cụm dữ liệu, một số khái niệm cần biết trong phân cụm dữ liệu và các yêu cầu cần thiết của phân cụm dữ liệu.

2.1 Tổng quan về khai phá dữ liệu

Trong hai thập kỷ qua, số lượng dữ liệu được lưu trữ trong CSDL cũng như số lượng các ứng dụng về CSDL trong các lĩnh vực kinh doanh và khoa học đã tăng lên rất nhiều lần. Sự bùng nổ về số lượng dữ liệu được lưu trữ này là nhờ sự thành công của mô hình dữ liệu quan hệ cùng với đó là sự phát triển và hoàn thiện của các công cụ truy xuất và thao tác dữ liệu. Trong khi công nghệ lưu trữ dữ liệu phát triển nhanh chóng để theo kịp nhu cầu, thì việc phát triển phần mềm để phân tích dữ liệu vẫn còn rất ít, cho đến gần đây thì các công ty nhận ra rằng ẩn bên trong những khối dữ liệu này là một nguồn tài nguyên đang bị bỏ qua.

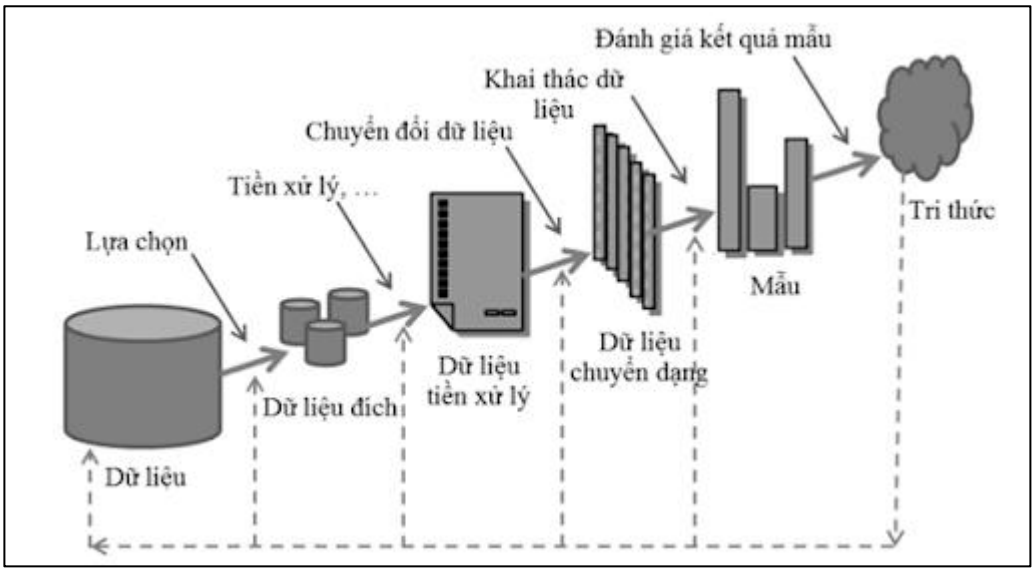
Hiện tại, các hệ thống quản lý CSDL được sử dụng để quản lý các tập dữ liệu này chỉ cho phép người dùng truy cập thông tin hiển thị rõ ràng trong CSDL, tức là dữ liệu. Dữ liệu được lưu trữ trong CSDL chỉ là một phần nhỏ của 'tảng băng thông tin'. Ẩn chứa trong dữ liệu này là kiến thức về một số khía cạnh của hoạt động kinh doanh của họ đang chờ được khai thác và sử dụng để hỗ trợ ra quyết định kinh doanh hiệu quả hơn. Việc trích xuất kiến thức từ các tập dữ liệu lớn này được gọi là Khai phá dữ liệu hoặc Khám phá tri thức trong Cơ sở dữ liệu và được định nghĩa là việc trích xuất những thông tin tiềm ẩn, chưa biết trước đây và có thể hữu ích từ dữ liệu. Thấy rõ được những lợi ích mang lại, nên đã có nhiều nguồn lực tập trung vào KPDL, và kéo theo đó là sự phát triển chung của ngành này.

Một cách ngắn gọn KPDL, còn được gọi là khám phá tri thức trong cơ sở dữ liệu (Knowledge discovery in databases - KDD), là lĩnh vực khám phá thông tin mới và hữu ích từ một lượng lớn dữ liệu. Khai thác dữ liệu đã được áp dụng trong rất nhiều lĩnh vực, bao gồm cả bán lẻ, tin sinh học và chống khủng bố. Ngoài ra cũng có nhiều thuật ngữ được dùng cũng có ý nghĩa với KPDL như Knowledge extraction (chắt lọc tri thức),data dredging (nạo vét dữ liệu), data/pattern analysis (phân tích dữ liệu/mẫu), Knowledge Mining (khai phá tri thức), data archaeology (khảo cổ dữ liệu), …

2.2 Quá trình khám phá tri thức, khai phá dữ liệu

2.2.1. Khám phá tri thức

Quá trình khám phá tri thức[6], gồm các bước:



Hình 2.1: Quá trình khám phá tri thức

Bước 1. Phát triển và hiểu về ứng dụng (Developing and understanding the application domain): Bước này bao gồm việc học kiến thức có liên quan trước đó và mục tiêu của người dùng cuối mà kiến thức đã khám phá sẽ mang lại cho họ.

Bước 2. Lựa chọn dữ liệu mục tiêu(Creating a target data set): Ở đây, công cụ khai thác dữ liệu chọn một tập hợp con các biến (thuộc tính) và điểm dữ liệu (các mẫu)

sẽ được sử dụng để thực hiện các tác vụ khai phá. Bước này thường bao gồm truy vấn dữ liệu hiện có để chọn tập hợp con mong muốn.

Bước 3. Làm sạch và tiền xử lý dữ liệu(Data cleaning and preprocessing): Dữ liệu sau khi được thu thập sẽ được làm sạch, rút gọn và rời rạc hóa. Phần lớn dữ liệu gốc đều ở dạng hỗn loạn, có thể thiếu thông tin hoặc thông tin sai lệch, do vậy cần được xử lý trước khi đưa vào các mô hình thuật toán. Dữ liệu sau khi được xử lý bước này sẽ nhất quán, sạch sẽ, đầy đủ, được rút gọn và được rời rạc hóa.

Bước 4. Giảm và chiếu dữ liệu(Data reduction and projection): Bước này bao gồm việc tìm kiếm các thuộc tính hữu ích bằng cách áp dụng các phương pháp biến đổi và giảm sai số dữ liệu, đồng thời tìm cách biểu diễn bất biến của dữ liệu.

Bước 5. Chuyển đổi dữ liệu (Data Transformation). Chuyển đổi dữ liệu là một kỹ thuật tiền xử lý dữ liệu thiết yếu phải được thực hiện trên dữ liệu trước khi khai thác dữ liệu để cung cấp các mẫu dễ hiểu hơn. Ở bước này dữ liệu được làm mịn và chuẩn hóa để phục vụ cho các bước sau.

Bước 6. Lựa chọn thuật toán khai thác dữ liệu(Choosing the data mining algorithm). Người khai thác dữ liệu sẽ chọn các phương pháp để tìm kiếm các mẫu trong dữ liệu và quyết định các mô hình và thông số của các phương pháp sẽ được sử dụng để có kết quả phù hợp nhất.

Bước 7. Khai phá dữ liệu(Data mining). Đây là công đoạn quan trọng và tốn phần lớn thời gian của cả quá trình KPTT, ở bước này các chuyên gia KPDL sẽ áp dụng các phương pháp, các thuật toán khai phá(phần lớn là các kỹ thuật của machine learning) để khai phá, trích chọn được các mẫu (pattern) thông tin cần thiết và các mối liên hệ trong dữ liệu.

Bước 8. Đánh giá và biểu diễn tri thức (knowledge representation & evaluation): Ở giai đoạn này, để trình bày một cách trực quan và dễ hiểu, các chuyên gia sẽ sử dụng các kỹ thuật biểu diễn và hiển thị để trực quan hóa các tri thức đã thu thập được dưới

dạng gần gũi với con người như đồ thị, cây, bảng biểu, luật,… cho người dùng. Ngoài ra, ở bước này cũng sẽ đánh giá được những tri thức khai phá theo các tiêu chí đã đề ra.

Kết quả mà KPTT mang lại cho giới kinh doanh là không hề nhỏ, do đó KPTT được xem như là một nhu cầu tất yếu của các doanh nghiệp, tập đoàn lớn. Tuy nhiên về mặt kỹ thuật, để có một kết quả tốt từ KPTT đó thực sự là một khó khăn và thách thức đối với các doanh nghiệp cũng như các chuyên gia. Vì KPTT phải được xây dựng dựa trên các giải thuật mới, định hướng theo nhu cầu của từng doanh nghiệp để nó giải quyết các bài toán về kinh doanh cho doanh nghiệp. Một số kỹ thuật đang được nghiên cứu và sử dụng để KPDL hiện nay như: phân lớp dữ liệu, phân cụm dữ liệu, cây quyết định (CART, CHAID, AID), mạng neuron, phương pháp láng giềng gần nhất(K Nearest Neighbour), các luật suy diễn (suy diễn tiến, suy diễn lùi),…

2.2.2. Quá trình khai phá dữ liệu

KPDL là một bước quan trọng trong quá trình KPTT. Công việc chính của giai đoạn này thực hiện là áp dụng các kỹ thuật khai phá, sau đó sẽ trích chọn ra các mẫu thông tin(pattern), các mối liên hệ với nhau trong dữ liệu. Kết quả sau khi thực hiện giai đoạn này là ta sẽ tìm ra được các dữ kiện thông tin mới, hữu ích ẩn chứa trong CSDL, và từ kết quả có được sẽ dùng để phục vụ cho mô tả và dự đoán. Và đây cũng là giai đoạn duy nhất trong cả qui trình để tìm ra được thông tin mới.

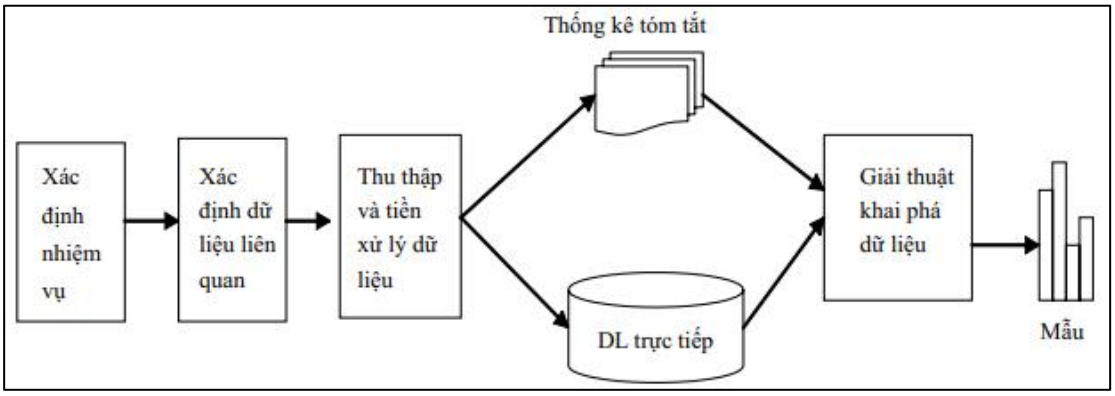
- Mô tả dữ liệu là công việc tóm tắt các văn bản hoặc biểu diễn một cách trực quan dễ hiểu những đặc điểm chung của những thuộc tính dữ liệu mà con người có thể dễ dàng hiểu được.

- Dự đoán là dựa trên những dữ kiện hiện có để từ đó ta có thể đoán ra được các quy luật từ các mối liên hệ giữa các thuộc tính của dữ liệu, và ta có thể rút ra được các pattern(mẫu). Dự đoán được những giá trị mà ta chưa biết hoặc những giá trị trong quá khứ hoặc những giá trị có thể đúng trong tương lai của dữ liệu.

Quá trình KPDL gồm các bước:

- **Bước 1:** Xác định nhiệm vụ: Ở bước này ta cần xác định chính xác, rõ ràng các vấn đề, nhiệm vụ mà ta cần phải giải quyết.

- **Bước 2:** Xác định các dữ liệu, dữ kiện liên quan: Trích chọn các dữ liệu, dữ kiện có liên quan để sử dụng chúng và xây dựng các giải pháp hợp lý.
- **Bước 3:** Thu thập và tiền xử lý dữ liệu: Thu thập dữ liệu để đào tạo mô hình ML là bước cơ bản trong quá học máy. Các dự đoán được cho ra kết quả tốt khi các dữ liệu mà chúng đã được đào tạo đủ tốt. Sau khi được thu thập, dữ liệu sẽ được xử lý trước thành một định dạng mà thuật toán học máy có thể sử dụng được. Nghe tuy rất đơn giản, nhưng khi bắt tay vào thực hiện ta sẽ gặp các vấn đề phát sinh cần phải giải quyết như: trùng lặp dữ liệu, quản lý tập các dữ liệu lớn, phải lặp lại nhiều lần toàn bộ quá trình (nếu như mô hình dữ liệu có thay đổi), v.v..
- **Bước 4:** Tiến hành khai phá bằng thuật toán KPDL: lựa chọn các thuật toán cần thiết và thực hiện việc KPDL để tìm được các mẫu(patterns) có ý nghĩa, các mẫu này được biểu diễn dưới dạng luật kết hợp, luật sản xuất, biểu thức hồi qui, cây quyết định... tương ứng với ý nghĩa của nó.



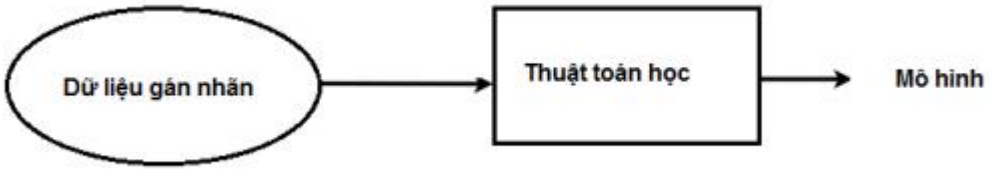
Hình 2.2: Quá trình KPDL

2.3 Các phương pháp khai phá dữ liệu

Nếu theo quan điểm của học máy (Machine Learning), thì các kỹ thuật trong khai phá dữ liệu, bao gồm:

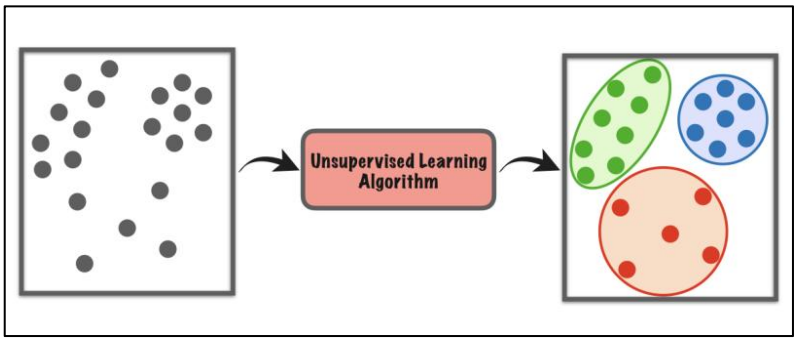
- Học có giám sát (Supervised learning): là một nhóm thuật toán sử dụng dữ liệu được gán nhãn nhằm mô hình hóa mối quan hệ giữa biến đầu vào (x) và biến đầu ra (y).

Hai nhóm bài toán cơ bản trong học có giám sát là classification (phân loại) và regression (hồi quy), trong đó biến đầu ra của bài toán phân loại có các giá trị rời rạc trong khi biến đầu ra của bài toán hồi quy có các giá trị liên tục. Với Supervised Learning, bên cạnh xây dựng các mô hình mạnh, việc thu thập và gán nhãn dữ liệu tốt và hợp lý cũng đóng vai trò then chốt để giải quyết các bài toán trong thực tế.



Hình 2.3: Mô hình học có giám sát

- Học không có giám sát (Unsupervised learning): là một nhóm thuật toán sử dụng dữ liệu không có nhãn. Các thuật toán theo cách tiếp cận này hướng đến việc mô hình hóa được cấu trúc hay thông tin ẩn trong dữ liệu. Hay nói cách khác, sử dụng các phương pháp này thiên về việc mô tả tính chất hay đặc tính của dữ liệu. Thông thường, các thuật toán này dựa trên những thông tin sau: Mối quan hệ tương tự (similarity), Xác suất đồng xuất hiện của các đối tượng, Các phép biến đổi ma trận để trích xuất các đặc trưng,…



Hình 2.4: Mô hình học không giám sát

- Học nửa giám sát (Semi - Supervised learning): Học nửa giám sát là một cách tiếp cận học máy kết hợp một lượng nhỏ dữ liệu được gắn nhãn với một lượng lớn dữ

liệu không được gắn nhãn trong quá trình đào tạo. Học nửa giám sát nằm giữa học tập không giám sát (dữ liệu không được gắn nhãn) và học tập có giám sát (dữ liệu có gắn nhãn). Đây là một trường hợp đặc biệt của việc giám sát yếu..

Nếu căn cứ vào lớp các bài toán cần giải quyết, thì khai phá dữ liệu bao gồm các kỹ thuật áp dụng sau:

- Phân lớp và dự đoán (classification and prediction): Phân lớp là xác định danh mục hoặc các nhãn của một tập dữ liệu huấn luyện. Đầu tiên, một tập dữ liệu được sử dụng làm dữ liệu huấn luyện. Tập dữ liệu huấn luyện bao gồm dữ liệu đầu vào và các kết quả đầu ra(nhãn) tương ứng cung cấp cho thuật toán. Sau khi huấn luyện thì kết quả thu được là mô hình. Các mô hình có thể là cây quyết định, công thức toán học hoặc mạng nơ-ron. Trong Phân lớp, khi dữ liệu chưa được gắn nhãn được cung cấp cho mô hình, nó sẽ tìm ra nhãn cho dữ liệu đó, và đây là mục tiêu của bài toán.

- Luật kết hợp (association rules): là một thủ tục nhằm tìm kiếm các mẫu, mối tương quan, liên kết hoặc cấu trúc nguyên nhân - kết quả từ các tập dữ liệu trong các loại cơ sở dữ liệu khác nhau như cơ sở dữ liệu quan hệ, cơ sở dữ liệu giao dịch và các dạng dữ liệu khác. Luật kết hợp được ứng dụng nhiều trong lĩnh vực như: bán hàng trong kinh doanh, y học, các lĩnh vực về tài chính, chứng khoán, .v.v.

- Phân cụm (clustering/ segmentation): Phân cụm là một Thuật toán dựa trên Học máy không được giám sát bao gồm một nhóm các điểm dữ liệu thành các cụm để các đối tượng thuộc cùng một nhóm. Phân cụm giúp chia dữ liệu thành nhiều tập con. Mỗi tập con này chứa dữ liệu tương tự nhau và các tập con này được gọi là các cụm.

- Khai phá mẫu tuần tự (Sequential Pattern Mining): Mẫu tuần tự là một tập hợp cơ sở dữ liệu có cấu trúc tập phổ biến xảy ra tuần tự với thứ tự cụ thể. Cơ sở dữ liệu trình tự là một tập hợp các thành phần hoặc sự kiện có thứ tự, được lưu trữ có hoặc không có thời gian cụ thể. Mỗi tập hợp chứa một tập hợp các mục bao gồm cùng một giá trị thời gian giao dịch. Trong khi các mô-đun liên kết chỉ ra các mối quan hệ nội bộ giao dịch, các câu hỏi tuần tự thể hiện mối tương quan giữa các giao dịch. Khai thác

theo mô hình tuần tự (SPM) [7] là quá trình phân tách các mô hình tuần tự nhất định có mức hỗ trợ vượt quá ngưỡng hỗ trợ tối thiểu được xác định trước. Ngoài ra, khai thác mẫu tuần tự giúp trích xuất các trình tự phản ánh các hành vi thường xuyên nhất trong cơ sở dữ liệu trình tự, do đó có thể được hiểu là kiến thức miền cho một số mục đích.

- Trực quan hóa (Visualization): trực quan hóa dữ liệu là biểu diễn đồ họa của dữ liệu và thông tin được trích xuất từ khai phá dữ liệu bằng cách sử dụng các yếu tố trực quan như đồ thị, biểu đồ và bản đồ, công cụ trực quan hóa dữ liệu và các kỹ thuật giúp phân tích lượng lớn thông tin và đưa ra quyết định về thông tin đó.

- Tổng hợp (Summarization): Tổng hợp dữ liệu có thể được định nghĩa là việc trình bày một bản tóm tắt / báo cáo dữ liệu được tạo ra một cách dễ hiểu và đầy đủ thông tin. Để chuyển tiếp thông tin về tập dữ liệu, bản tóm tắt được lấy từ toàn bộ tập dữ liệu. Đây là một bản tóm tắt được thực hiện cẩn thận sẽ truyền đạt các xu hướng và mẫu từ tập dữ liệu theo cách đơn giản hóa.

- Mô hình ràng buộc (Dependency modeling): Mô hình ràng buộc bao gồm việc tìm kiếm một mô hình mô tả sự phụ thuộc đáng kể giữa các biến. Mô hình phụ thuộc tồn tại ở hai cấp độ: (1) cấp độ cấu trúc của mô hình cụ thể (thường ở dạng đồ họa) các biến nào phụ thuộc cục bộ vào nhau và (2) cấp độ xác thực của mô hình xác định độ mạnh của các yếu tố phụ thuộc bằng cách sử dụng một số tỉ lệ.

- Đánh giá mô hình (Model Evaluation): Đánh giá mô hình là quá trình sử dụng các chỉ số đánh giá khác nhau để hiểu hiệu suất của mô hình học máy cũng như điểm mạnh và điểm yếu của nó. Đánh giá mô hình là quan trọng để đánh giá hiệu quả của một mô hình trong các giai đoạn nghiên cứu ban đầu, và nó cũng đóng một vai trò trong việc giám sát mô hình.

2.4 Phân cụm dữ liệu

2.4.1 Phân cụm là gì? Mục đích của phân cụm dữ liệu

Phân cụm dữ liệu[8] là việc phân nhóm các đối tượng cụ thể dựa trên các đặc điểm và điểm tương đồng của chúng (thường là các thuộc tính của dữ liệu). Đối với

khai phá dữ liệu, phương pháp này phân chia dữ liệu phù hợp nhất với phân tích mong muốn bằng cách sử dụng một thuật toán nối đặc biệt. Phân tích này cho phép một đối tượng thuộc hoặc không một cụm, được gọi là phân cụm cứng.

Phân cụm dữ liệu được xem là học không giám sát(Unsupervised learning), vì nó phân nhóm các đối tượng không được gắn nhãn và thực hiện công việc phân nhóm chỉ dựa vào đặc tính của các dữ liệu đầu vào thường là dựa vào độ đo mức độ tương đồng của dữ liệu.

Phân cụm được các chuyên gia sử dụng để phân loại khách hàng, phân khúc khách hàng theo những đặc điểm về khách hàng đã xác định từ trước ví dụ sử dụng phân cụm để phân khúc khách hàng dựa theo điểm tín dụng (credit scores) trong ngành tài chính ngân hàng, hay phân khúc khách hàng trong ngành viễn thông, ngành bán lẻ dựa trên mô hình RFM (Recency-Frequency-Monetary Value) để xác định nhóm khách hàng chi tiêu nhiều, đến nhóm khách hàng chi tiêu thấp, khách hàng sử dụng dịch vụ thường xuyên đến khách hàng không sử dụng dịch vụ,… để đánh giá tổng quát.

2.4.2 Các bước cơ bản để phân cụm

- Chọn lựa đặc trưng: là một kỹ thuật cần thiết để giảm vấn đề về kích thước trong tác vụ khai phá dữ liệu. Các đặc trưng cần phải được tiền xử lý(xử lý nhiễu, trùng lặp,…) trước khi được dùng cho các bước tiếp theo. Kết quả phân cụm sẽ khác nhau nếu các đặc trưng được chọn khác nhau. Do đó việc lựa chọn các đặc trưng hợp lý là dựa vào kiến thức và kinh nghiệm của các chuyên gia.

- Chọn độ đo: Ứng với từng phương pháp phân cụm khác nhau mà ta lựa chọn các độ đo phù hợp để cho ra kết quả phù hợp nhất.

- Tiêu chuẩn phân cụm: Ứng với mỗi tập dữ liệu khác nhau sẽ tạo ra các cụm khác nhau và từ đó ta có các tiêu chuẩn phân cụm khác nhau. Từ các hàm chi phí(tính độ đo giữa các cụm) mà ta có thể tính ra được chi phí và chọn ra tiêu chuẩn phân cụm hợp lý.

- Thực thi thuật toán phân cụm: các giải thuật phân cụm khác nhau sẽ được sử dụng ở giai đoạn này, với mục tiêu là làm sáng tỏ các cấu trúc cụm của tập dữ liệu đầu vào.

- Công nhận kết quả: Sau khi thực thi các thuật toán phân cụm và thu được kết quả phân cụm thì ta phải kiểm tra tính đúng đắn và hợp lý của nó. Các kiểm định phù hợp sẽ được sử dụng ở giai đoạn này để lựa chọn và công nhận kết quả.

- Giải thích kết quả: Dựa vào kinh nghiệm thực tế và kết quả phân cụm vừa đạt được, các chuyên gia trong lĩnh vực ứng dụng phải kết hợp những bằng chứng thực nghiệm và kỹ năng phân tích để đưa ra các kết quả đúng đắn và hợp lý nhất.

2.4.3 Các ứng dụng của phân cụm

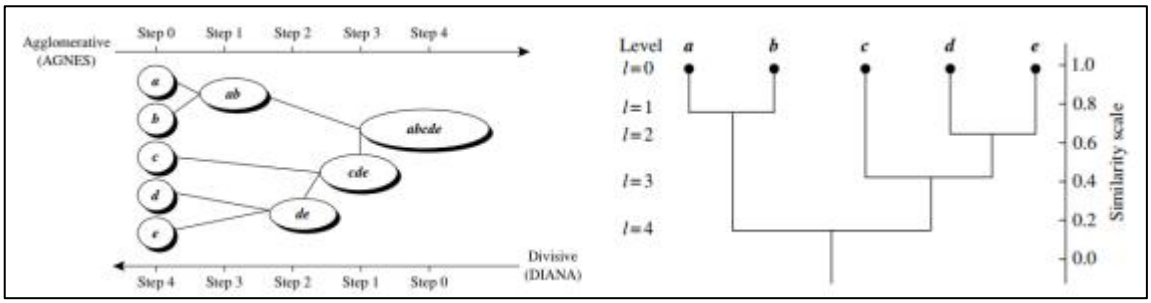
- Hiểu các dữ liệu(Understanding)
  + Gộp nhóm các tài liệu liên quan.
  + Nhóm các gen và protein có chức năng tương tự về mặt sinh học.
  + Phân cụm các cổ phiếu có giá biến động tương tự.
  + …
- Tóm tắt dữ liệu: Giảm kích thước dữ liệu.
- Hỗ trợ giai đoạn tiền xử lý dữ liệu (data processing).
- Nhận dạng mẫu(pattern recognition).
- Phân tích dữ liệu không gian(spatial data analysis).
- Xử lý ảnh(image processing).
- Phân mảnh thị trường(market segmentation).

2.4.4 Các phương pháp phân cụm dữ liệu

a. Phương pháp phân cụm Phân cấp(Hierarchical clustering)

Phân cụm phân cấp (hierarchical clustering) Phân cụm phân cấp, còn được gọi là phân tích cụm phân cấp, là một thuật toán nhóm các đối tượng tương tự thành các nhóm được gọi là cụm. Điểm cuối là một tập hợp các cụm, trong đó mỗi cụm khác biệt

với từng cụm khác, và các đối tượng trong mỗi cụm tương tự nhau. Có hai hướng tiếp cận đối với phương pháp phân cụm phân cấp này: Agglomerative và Divisive.



Hình 2.5: Phân cụm theo cách tiếp cận top-down/bottom-up và dendrogram biểu diễn cây phân cấp đối tượng {a,b,c,d,e}

- Agglomerative clustering: Phương pháp tiếp cận từ dưới lên(bottom-up) Nghĩa là, mỗi đối tượng ban đầu được coi như một cụm đơn nguyên tố (lá). Ở mỗi bước của thuật toán, hai cụm giống nhau nhất được kết hợp thành một cụm (nút) mới lớn hơn. Quy trình này được lặp lại cho đến khi tất cả các điểm chỉ là thành viên của một cụm lớn duy nhất (nút gốc).

- Divisive clustering: Ngược lại với agglomerative, Còn được gọi là cách tiếp cận từ trên xuống. Thuật toán này cũng không yêu cầu xác định trước số lượng cụm. Phân cụm từ trên xuống yêu cầu một phương pháp để tách một cụm chứa toàn bộ dữ liệu và tiến hành bằng cách tách thành các cụm con và thực hiện đệ quy cho đến khi các nút con được tách thành các cụm đơn lẻ.

Ta có thể sử dụng các phương pháp xác định mối liên kết sau để xác định khoảng cách giữa các cụm:

1) Single linkage: Khoảng cách giữa hai cụm được xác định là khoảng cách ngắn nhất giữa hai điểm trong mỗi cụm. Liên kết này có thể được sử dụng để phát hiện các giá trị cao trong tập dữ liệu, những giá trị này có thể là giá trị ngoại lệ vì chúng sẽ được hợp nhất ở cuối.

2) Complete linkage: Khoảng cách giữa hai cụm được xác định là khoảng cách xa nhất giữa hai điểm trong mỗi cụm.

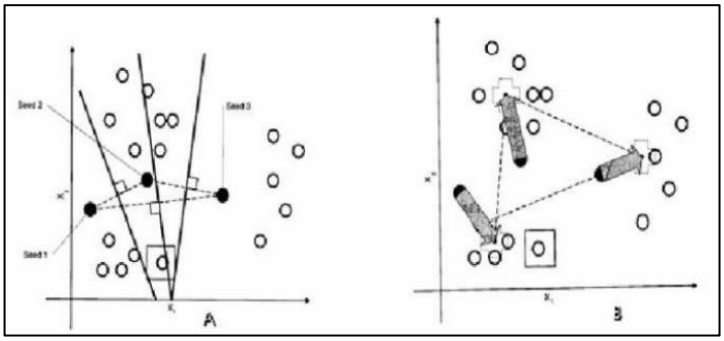
3) Average linkage: Khoảng cách giữa hai cụm được xác định là khoảng cách trung bình giữa mỗi điểm trong một cụm với mọi điểm trong cụm khác.

4) Centroid-linkage: Tìm tâm của cụm 1 và tâm của cụm 2, sau đó tính toán khoảng cách giữa hai trước khi hợp nhất.

Các thuật toán điển hình cho phương pháp phân cụm phân cấp gồm có CURE, BIRCH, ROCK, AGNES, DIANA và Chameleon.

b. Phương pháp phân cụm Phân hoạch(Hierarchical Partitional)

Phương pháp phân hoạch (partitional clustering) là tạo ra các phân vùng khác nhau và sau đó đánh giá chúng theo một số tiêu chí. Chúng cũng được gọi là không phân cấp vì mỗi cá thể được đặt trong chính xác một trong k cụm loại trừ lẫn nhau. Bởi vì chỉ có một tập hợp k cụm là đầu ra của thuật toán phân cụm phân hoạch điển hình, người dùng được yêu cầu nhập số lượng cụm mong muốn (thường được gọi là k).



Hình 2.6: Ví dụ phân hoạch với k=3

Một trong những thuật toán phân cụm phân hoạch được sử dụng phổ biến nhất là thuật toán phân cụm K-Means do có ưu điểm là một giải thuật đơn giản dễ cài đặt, và cho ra kết quả dễ hiểu. Tuy nhiên khả năng chịu nhiễu không tốt, cùng với đó là dễ bị ảnh hưởng bởi các phần tử nhiễu, ngoại lệ, nên đây có thể xem là nhược điểm của thuật toán này.

Thuật toán PAM (Partitioning Around Medoids) tìm kiếm k đối tượng đại diện trong tập dữ liệu (k tâm) và sau đó gán từng đối tượng cho tâm gần nhất để tạo thành các cụm. Mục đích của nó là giảm thiểu tổng điểm không giống nhau giữa các đối

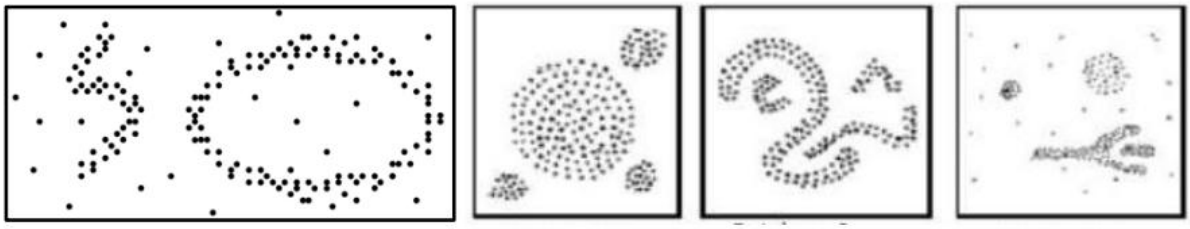
tượng trong một cụm và trung tâm của cùng một cụm (medoid). Nó được biết đến là một phiên bản mạnh mẽ của k-means vì nó được coi là ít nhạy cảm hơn với các ngoại lệ.

Ngoài ra, các giải thuật như CLARA, CLARANS cũng cho ra kết quả phân cụm tốt.

c. Phương pháp phân cụm dựa trên mật độ (Density-based clustering)

Phân cụm dựa trên mật độ đề cập đến một trong những phương pháp học không giám sát phổ biến nhất được sử dụng trong các thuật toán xây dựng mô hình và học máy. Các điểm dữ liệu trong vùng cách nhau bởi hai cụm có mật độ điểm thấp được coi là nhiễu. Môi trường xung quanh có bán kính ε của một đối tượng nhất định được gọi là vùng lân cận ε của đối tượng. Nếu ε vùng lân cận của đối tượng bao gồm ít nhất một số tối thiểu, MinPts của các đối tượng, thì nó được gọi là đối tượng cốt lõi.

Các phương pháp phân cụm dựa trên mật độ là rất tốt vì chúng không chỉ định trước số lượng các cụm. Không giống như các phương pháp phân cụm khác, chúng kết hợp khái niệm về các giá trị ngoại lai và có thể "lọc" chúng ra.



Hình 2.7: Các cụm có hình dạng bất kỳ

Một số thuật toán phổ biến cho phương pháp phân cụm dựa trên mật độ này là: DBSCAN, HDBSCAN, OPTICS, DENCLUE.

d. Phương pháp phân cụm dựa trên lưới(Grid-based Clustering)

Các phương pháp tiếp cận dựa trên mật độ và/hoặc dựa trên lưới phổ biến đối với các cụm khai thác trong một không gian đa chiều rộng lớn, trong đó các cụm được coi là vùng dày đặc hơn so với môi trường xung quanh chúng.

Độ phức tạp tính toán của hầu hết các thuật toán phân cụm ít nhất là tỷ lệ tuyến tính với kích thước của tập dữ liệu. Ưu điểm lớn của phân cụm dựa trên lưới là giảm đáng kể độ phức tạp tính toán, đặc biệt là đối với phân cụm các tập dữ liệu rất lớn.

Cách tiếp cận phân cụm dựa trên lưới khác với các thuật toán phân nhóm thông thường ở chỗ nó không quan tâm đến các điểm dữ liệu mà quan tâm đến không gian giá trị bao quanh các điểm dữ liệu. Nói chung, một thuật toán phân cụm dựa trên lưới điển hình bao gồm năm bước cơ bản sau (Grabusts và Borisov, 2002):

Bước 1: Tạo cấu trúc lưới, tức là phân vùng không gian dữ liệu thành một số ô hữu hạn.

Bước 2: Tính mật độ ô cho mỗi ô.

Bước 3: Sắp xếp các ô theo mật độ của chúng.

Bước 4: Xác định các trung tâm cụm.

Bước 5: Truyền qua các ô lân cận.

e. Phương pháp phân cụm có dữ liệu ràng buộc

Sự phát triển của phân cụm không gian trên cơ sở dữ liệu lớn đã cung cấp nhiều công cụ tiện lợi để phân tích thông tin địa lý, tuy nhiên hầu hết các thuật toán này cung cấp rất ít cách để người dùng xác định các ràng buộc trong thế giới thực cần được thỏa mãn trong quá trình phân nhóm. Để phân cụm không gian hiệu quả hơn, cần phải thực hiện nghiên cứu bổ sung để cung cấp cho người dùng khả năng kết hợp các ràng buộc trong thuật toán phân cụm.

2.4.5 Các thách thức phân cụm

a. Những thách thức chung trong khai phá dữ liệu:

Nhưng thách thức thường gặp trong khai phá dữ liệu[9]:

- *Dữ liệu bị nhiễu và không đầy đủ:* Dữ liệu bị thiếu hoặc không đầy đủ thì khá phổ biến. Việc bỏ qua các trường hợp có giá trị bị thiếu thường dẫn đến thông tin bị mất, điều này đi ngược lại với việc phát triển một mô hình KPDL tốt. Có nhiều phương pháp thống kê để xử lý dữ liệu bị thiếu và xác định các giá trị thuộc tính bị nhiễu.

- *Làm sạch và tiền xử lý dữ liệu:* Trong quá trình này, dữ liệu nhiễu và dữ liệu không liên quan được loại bỏ khỏi bộ sưu tập. Nó điền các giá trị bị thiếu, đồng thời xác định các giá trị ngoại lệ, nó sẽ làm giảm nhiễu và sửa các điểm không nhất quán trong dữ liệu. Làm sạch dữ liệu bao gồm quy trình hai bước lặp đi lặp lại bao gồm: phát hiện sự khác biệt và chuyển đổi dữ liệu.

- *Quá khớp (Overfitting):* Mô hình rất hợp lý, rất khớp với tập huấn luyện nhưng khi đưa ra dự đoán với dữ liệu mới thì lại không phù hợp. Nguyên nhân có thể do ta chưa đủ dữ liệu để đánh giá hoặc do mô hình của ta quá phức tạp. Mô hình bị quá phức tạp khi mà mô hình của ta sử dụng cả những nhiễu lớn trong tập dữ liệu để học, dẫn tới mất tính tổng quát của mô hình.

- *Dữ liệu đa dạng và không đồng nhất:* Các kỹ thuật khai phá dữ liệu trước đây được sử dụng để khai phá các mẫu chưa biết và các mối quan hệ các tập dữ liệu nhỏ, có cấu trúc, đồng nhất. Sự đa dạng là một trong những đặc điểm quan trọng của dữ liệu lớn. Đây là kết quả sự tổng hợp của gần như không giới hạn các nguồn dữ liệu, hệ quả tất yếu của hiện tượng này là sự không đồng nhất của dữ liệu.

- *Thông tin hạn chế:* dữ liệu thu được tuy có nhưng không đầy đủ khiến cho kết quả đầu ra không chính xác.

- *Quy mô dữ liệu:* Dung lượng và quy mô lớn chưa từng có của dữ liệu lớn đòi hỏi các công cụ quản lý và khai phá dữ liệu phải được cải tiến tương ứng. Điểm quan trọng là với quy mô cực lớn thì ta có nhiều cơ hội để khám phá nhiều tri thức hơn trong dữ liệu thông thường (quy mô nhỏ). Những hướng tiếp cận dưới đây nếu được áp dụng hợp lý sẽ đem lại hiệu quả trong khai phá dữ liệu lớn: (1) điện toán đám mây kết hợp với tính toán song song; (2) tương tác người dùng (đồ hoạ - GUI hoặc dựa trên ngôn ngữ) - giúp việc tương tác giữa người dùng và hệ thống trở nên nhanh chóng và hiệu quả.

- *Việc kết hợp các kiến thức nền:* Việc đọc kết quả, thực hiện lựa chọn các đặc trưng, thuộc tính để tiến hành khai phá dữ liệu cần phải có một kiến thức nền tương đối

để đọc kết quả một cách chính xác với thực tế nhất. Vì vậy ngoài các kiến thức chuyên môn về khai phá dữ liệu thì cần phải nắm hoặc phối hợp với các chuyên gia trong lĩnh vực đó để có kết quả đầu ra tốt nhất.

- *Trực quan hóa dữ liệu:* Nhiệm vụ chính ở giai đoạn này là truyền thông và trình bày kết quả thu được một cách rõ ràng và hiệu quả cho người dùng cuối thông qua đồ họa như là các bảng biểu hoặc biểu diễn bằng đồ thị. Bảng biểu thường được dùng khi xem xét hoặc đo lường giá trị của một biến. Kết quả thu được là dữ liệu phức tạp trở thành được thể hiện một cách dễ hiểu hơn. Người sử dụng có thể dễ dàng thực hiện phân tích như tạo phép so sánh dữ liệu.

- *Tốc độ/tính chuyển động liên tục:* Đối với dữ liệu lớn, tốc độ/chuyển động liên tục thực sự quan trọng. Khả năng truy nhập nhanh và khai phá dữ liệu lớn không chỉ là mong muốn chủ quan mà là một nhiệm vụ xử lý đặc biệt đối với các dòng dữ liệu (data stream) (một định dạng phổ biến của dữ liệu lớn) - chúng ta phải hoàn thành việc xử lý/khai phá dòng dữ liệu đó trong một thời gian nhất định, bởi nếu không thì kết quả xử lý/ khai phá đó trở nên ít có giá trị hoặc thậm chí là vô giá trị. Chẳng hạn, ứng dụng đòi hỏi chạy theo thời gian thực như dự đoán động đất, dự đoán thị trường chứng khoán, thị trường ngoại hối...

- *Ngôn ngữ truy vấn khai phá dữ liệu:* Ngôn ngữ truy vấn đóng một vai trò quan trọng trong việc tìm kiếm một cách linh hoạt. Nó sẽ tạo điều kiện thuận lợi cho việc đặc tả các bộ dữ liệu có liên quan để phân tích.

- *Bảo mật dữ liệu riêng tư***:** Dữ liệu riêng tư luôn là vấn đề cần xem xét trong khai phá dữ liệu. Vấn đề này còn nghiêm trọng hơn khi các ứng dụng khai phá dữ liệu lớn thường đòi hỏi các thông tin cá nhân để tạo ra các kết quả có liên quan đến từng cá nhân như các dịch vụ dựa trên địa điểm (chẳng hạn quảng cáo). Hơn nữa, trong các dữ liệu có được từ các phương tiện truyền thông hay mạng xã hội, các thông tin cá nhân của nhiều người thường có liên quan đến nhau và dễ dàng bị "đào xới“ bởi các ứng dụng khai phá dữ liệu. Một ví dụ đơn giản, các giao dịch trong cuộc sống hàng ngày

của chúng ta đang được đưa lên mạng và được lưu vết ở đó: email, tin nhắn, blog, Facebook, mua sắm, thanh toán hoá đơn trực tuyến, số điện thoại, địa chỉ nhà, ngày sinh...

- *Giải thích kết quả*: Đầu ra khai thác dữ liệu có thể yêu cầu các chuyên gia giải thích chính xác kết quả đầu ra.

b. Các thách thức trong phân cụm dữ liệu:

- *Xác định cách tính khoảng cách[10]:* Đối với các thuộc tính số, các thước đo khoảng cách có thể được sử dụng là các phương trình tiêu chuẩn như euclidean, manhattan và thước đo khoảng cách tối đa. Cả ba đều là trường hợp đặc biệt của khoảng cách Minkowski.

- *Xác định số lượng cụm:* Việc xác định số lượng các cụm là một nhiệm vụ khó khăn nếu không biết trước số lượng nhãn lớp. Cần phân tích cẩn thận số lượng các cụm để đưa ra kết quả chính xác. Điều này có thể là thảm họa nếu phương pháp được sử dụng là phân cấp.

- *Thiếu nhãn:* Đối với các tập dữ liệu thực (về bản chất là quan hệ vì chúng có các bộ giá trị và các thuộc tính), việc phân phối dữ liệu phải được thực hiện để hiểu các nhãn lớp ở đâu?

- *Cấu trúc của cơ sở dữ liệu:* Trong thực tế dữ liệu có thể không phải lúc nào cũng chứa các cụm có thể nhận dạng rõ ràng. Ngoài ra, thứ tự sắp xếp các bộ giá trị có thể ảnh hưởng đến kết quả khi một thuật toán được thực thi nếu thước đo khoảng cách được sử dụng không hoàn hảo. Thậm chí việc xác định số lượng các cụm thích hợp sẽ không mang lại kết quả tốt.

- *Các loại thuộc tính trong cơ sở dữ liệu:* Cơ sở dữ liệu có thể không nhất thiết phải chứa các thuộc tính số hoặc phân loại rõ ràng, chúng cũng có thể chứa các loại khác như danh nghĩa, thứ tự, nhị phân, v.v. Vì vậy, các thuộc tính này phải được chuyển đổi thành loại danh mục để làm cho việc tính toán trở nên đơn giản.

- *Chọn tâm cụm khởi tạo ban đầu:* Đối với phương pháp phân cụm phân hoạch, ta thấy rằng hầu hết các thuật toán đề cập đến "k vị trí ban đầu" được chọn ngẫu nhiên. Ngoài ra, nếu các "vị trí ban đầu" không được chọn đúng cách, thì sau một vài lần lặp lại, người ta thấy rằng có thể xuất hiện các cụm không có phần tử.

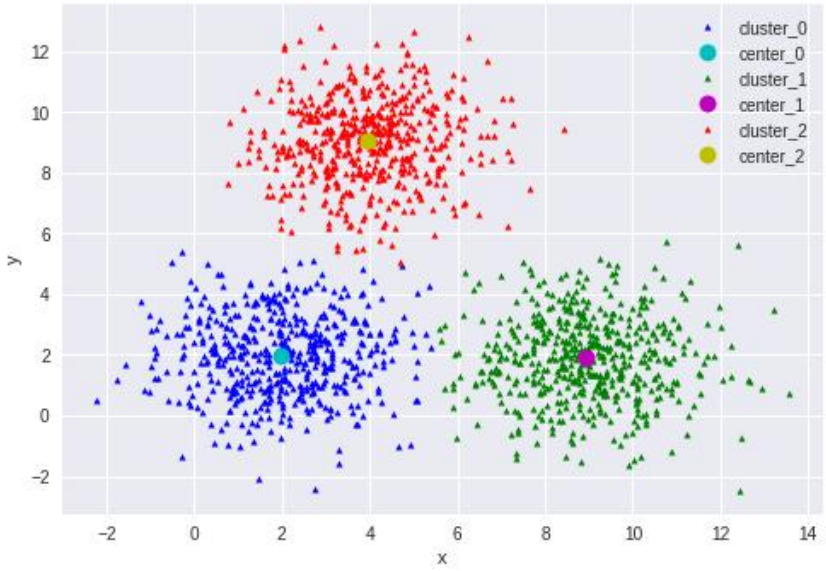
2.5 Thuật toán phân cụm K-Means

2.5.1 Tổng quan về thuật toán

Thuật toán phân cụm K-Means là một thuật toán Học tập không được giám sát, nhiệm vụ của nó nhóm các tập dữ liệu không được gắn nhãn thành các cụm khác nhau. Ở đây K xác định số lượng cụm được xác định trước cần được tạo trong quá trình này, như nếu K = 2, sẽ có hai cụm và đối với K = 3, sẽ có ba cụm, v.v. . Thuật toán được dựa trên ý tưởng về việc tính toán khoảng cách của các đối tượng dữ liệu trong một cụm. Nó là một thuật toán dựa trên tâm cụm, trong đó mỗi cụm được liên kết với một tâm. Mục đích chính của thuật toán này là giảm thiểu tổng khoảng cách giữa điểm dữ liệu và các cụm tương ứng của chúng.

Ý tưởng của thuật toán k-means:

- **Bước 1**: Chọn số K để quyết định số lượng cụm mong muốn.
- **Bước 2**: Chọn K điểm hoặc trọng tâm ngẫu nhiên. (Nó có thể khác với tập dữ liệu đầu vào).
- **Bước 3**: Gán mỗi điểm dữ liệu cho tâm gần nhất của chúng, sẽ tạo thành các cụm K được xác định trước.
- **Bước 4**: Tính khoảng cách và đặt trọng tâm mới ở mỗi cụm.
- **Bước 5**: Lặp lại Bước 3 và Bước 4 cho tới khi vị trí của tâm cụm không thay đổi hoặc tâm của tất cả các điểm dữ liệu không thay đổi.
- **Bước 6**: Thu được mô hình phân cụm.



Hình 2.8: Phân cụm k-means với k = 3

Vì K-means là học tập không giám sát, nên chúng ta không biết nhãn (label) của từng điểm dữ liệu. Mục đích là phân dữ liệu thành các cụm (cluster) khác nhau sao cho dữ liệu trong cùng một cụm có tính chất giống nhau.

Mục đích của thuật toán K-means là sinh k cụm dữ liệu {C1, C2, ..., Ck} từ một tập dữ liệu chứa n đối tượng trong không gian d chiều $Xi = \{x_{i1}, x_{i2}, \ldots, x_{id}\}$, i = 1 ÷ n, sao cho hàm tiêu chuẩn:

$$E = \sum_{i=1}^{k} \sum_{x \in C_i} D^2 x - m_i$$

đạt giá trị tối thiểu, trong đó m_i là trọng tâm của cụm C_i, D là khoảng cách giữa hai đối tượng.

2.5.2 Hạn chế của K-Means

Thuật toán k-Means có một số hạn chế đó là:

- Chúng ta cần phải xác định trước số cụm cho thuật toán: Vì bộ dữ liệu của chúng ta 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 (try and error) và xác định số cụm thông qua một phương pháp chẳng hạn như Elbow.

- 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.

2.6 Thuật toán K-Means++

Một nhược điểm cố hữu của thuật toán K-means là nó nhạy cảm với việc khởi tạo các trọng tâm hoặc các điểm trung bình. Vì vậy, nếu một tâm cụm được khởi tạo là một điểm "xa", nó có thể chỉ kết thúc không có điểm nào được liên kết với nó và đồng thời, nhiều hơn một cụm có thể kết thúc với một tâm duy nhất. Tương tự, nhiều hơn một trung tâm có thể được khởi tạo vào cùng một cụm dẫn đến phân cụm kém.

Để giải quyết nhược điểm trên của thuật toán K-means thì một thuật toán xác định tâm cụm K-means++ được đề xuất. Nó được đề xuất vào năm 2007 bởi David Arthur và Sergei Vassilvitskii.

Thuật toán này đảm bảo khởi tạo tâm cụm thông minh hơn và cải thiện chất lượng của phân cụm. Ngoài phần khởi tạo, phần còn lại của thuật toán giống như thuật toán K-means tiêu chuẩn. Đó là K-means++ là thuật toán K-means tiêu chuẩn kết hợp với việc khởi tạo centroid thông minh hơn.

Các bước thực hiện là:

Bước 1: Chọn ngẫu nhiên tâm đầu tiên từ các điểm dữ liệu.

Bước 2: Đối với mỗi điểm dữ liệu, tính toán khoảng cách của nó từ trung tâm gần nhất, đã chọn trước đó.

Bước 3: Chọn centroid tiếp theo từ các điểm dữ liệu sao cho xác suất chọn một điểm làm tâm tỷ lệ thuận với khoảng cách của nó từ tâm gần nhất, đã chọn trước đó.

(tức là điểm có khoảng cách tối đa từ tâm gần nhất có nhiều khả năng được chọn tiếp theo làm tâm).

Bước 4: Lặp lại các bước 2 và 3 cho đến khi k cụm được lấy mẫu

2.7 Các thuật toán xác định số cụm tối ưu

2.7.1 Phương pháp khủy tay(Elbow method)

Elbow method được minh họa dưới dạng đồ thị đường cong với trục hoành là số k các cụm, trục tung sẽ là tiêu chí đánh giá bao gồm SSE, Silhouette. Ở phần này chúng ta tìm hiểu trước về SSE – (Sum Squared Error) – đo lường sự khác biệt giữa các điểm trong cluster[10][11]. Trong k-means clustering, SSE được tính là tổng các khoảng cách tính từ các điểm trong cluster đến điểm trung tâm Centroid của cluster, tính tất cả các cluster, dựa theo công thức Euclidean. Khi các điểm dữ liệu hay các đối tượng, các quan sát càng gần nhau thì sẽ có đặc điểm gần giống nhau, được phân trong một cụm, thì cụm đó chứng tỏ "chất lượng" và ngược lại.

$$\mathrm{SSE} = \sum_{i=1}^{k} \sum_{x_{ij} \in C_i} \mathrm{Distance}^2(x_{ij}, m_i)$$

Trong đó:

k: là số cụm tối đa cần phải tính SSE, thông thường k sẽ chạy từ 1-20

n_i: là số lượng phần tử của cụm C_i

x_{ij}: là các điểm trong cụm C_i

m_i: là tâm của cụm C_i

$\mathrm{Distance}^2(x_{ij}, m_i)$: là bình phương khoảng cách của điểm x_{ij} tới tâm cụm m_i

Sẽ có k cluster cần tính giá trị SSE thường k sẽ chạy từ 1 đến 10 hay 20. Như vậy với mỗi k chúng ta sẽ có 1 SSE. Minh họa các cặp k và SSE lên đồ thị. Số k tối ưu chính là điểm mà ở đó SSE bắt đầu giảm đều, nhìn trên đồ thị nó là điểm "turning point", điểm nằm ở vị trí "cùi chỏ" sẽ là số k cần tìm.

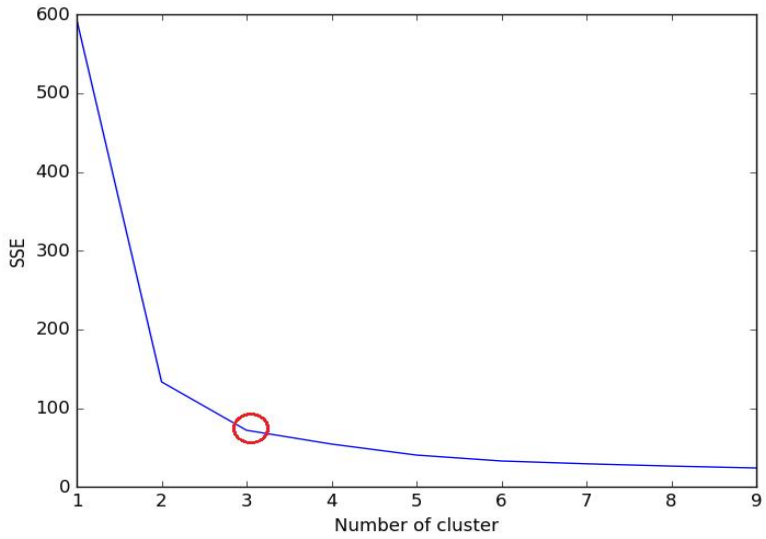
Các bước thực hiện của thuật toán được minh họa như sau:

Bước 1: Tính toán thuật toán phân cụm (ví dụ: phân cụm k-means) cho các giá trị khác nhau của k. Ví dụ: bằng cách thay đổi k từ 1 đến 10 cụm.

Bước 2: Với mỗi k, hãy tính tổng của bình phương khoảng cách trong một cụm (SSE)

Bước 3: Vẽ đồ thị đường cong của SSE theo số cụm k.

Bước 4: Vị trí của một khúc quanh (khủy tay) trong đồ thị được coi là số lượng cụm thích hợp để thực hiện phân cụm.



Hình 2.9: Xác định số cụm tối ưu là 3 bằng phương pháp Elbow method

2.7.2 Phương pháp điểm hình bóng trung bình(Average silhouette method)

Điểm hình bóng(Silhouette) được sử dụng để đánh giá chất lượng của các cụm được tạo bằng thuật toán phân cụm như K-Means về mức độ tốt của các mẫu được nhóm với các mẫu khác tương tự nhau. Điểm hình bóng được tính cho từng mẫu của các cụm khác nhau. Để tính điểm hình bóng cho mỗi điểm quan sát / dữ liệu, cần tìm ra các khoảng cách sau cho mỗi lần quan sát thuộc tất cả các cụm:

Khoảng cách trung bình là khoảng cách giữa quan sát và tất cả các điểm dữ liệu khác trong cùng một cụm. Khoảng cách này cũng có thể được gọi là khoảng cách trung bình trong cụm. Khoảng cách trung bình được ký hiệu bằng **a(i).**

Khoảng cách trung bình giữa quan sát và tất cả các điểm dữ liệu khác của cụm gần nhất tiếp theo. Khoảng cách này cũng có thể được gọi là khoảng cách cụm gần nhất trung bình. Khoảng cách trung bình được ký hiệu là **b(i).**

Điểm hình bóng trung bình - **S**, cho mỗi mẫu được tính theo công thức sau:

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

Trong đó:

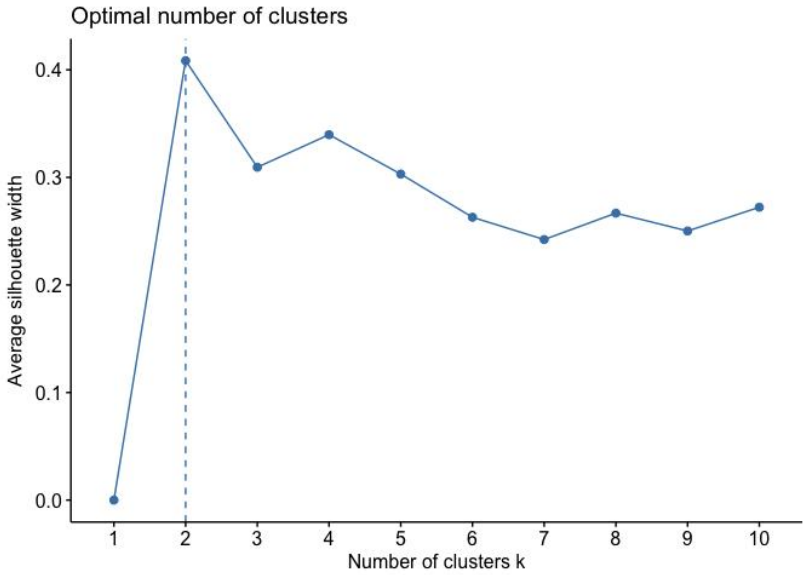
a(i) là khoảng cách từ điểm i trong cụm A đến điểm trung tâm của cụm A.

b(i) là khoảng cách từ điểm i trong cụm A đến điểm trung tâm của cụm B.

Cụm tương ứng với **b(i)** này được gọi là cụm hàng xóm của i.

Giá trị của điểm hình bóng thay đổi từ -1 đến 1. Nếu điểm là 1, cụm này dày đặc và tách biệt tốt hơn các cụm khác. Giá trị gần 0 đại diện cho các cụm chồng chéo với các mẫu rất gần với ranh giới quyết định của các cụm lân cận. Điểm âm [-1, 0] chỉ ra rằng các mẫu có thể đã được gán cho các cụm sai.

Một cách ngắn gọn, phương pháp tính độ hình bóng trung bình của một cụm là xác định mức độ phù hợp của mỗi đối tượng nằm trong cụm của nó. Điểm hình bóng trung bình cao cho thấy sự phân nhóm tốt. Phương pháp tính độ hình bóng trung bình quan sát cho các giá trị khác nhau của k. Số cụm tối ưu k là cụm tối đa hóa hình bóng trung bình trên một phạm vi các giá trị có thể cho k.



Hình 2.10: Xác định số cụm tối ưu là 2 bằng phương pháp Average silhouette

Sau đây là một số kết luận rút ra về phương pháp tính điểm hình bóng trung bình:

- Điểm hình bóng cho một tập hợp các điểm dữ liệu mẫu được sử dụng để đo lường mức độ dày đặc và được phân tách rõ ràng của các cụm.
- Điểm hình bóng xem xét khoảng cách trong cụm giữa mẫu và các điểm dữ liệu khác trong cùng một cụm (a) và khoảng cách liên cụm giữa mẫu và cụm gần nhất tiếp theo (b).
- Điểm hình bóng nằm trong khoảng [-1, 1].
- Điểm hình bóng bằng 1 có nghĩa là các cụm rất dày đặc và được phân tách độc đáo. Điểm 0 có nghĩa là các cụm trùng nhau. Điểm nhỏ hơn 0 có nghĩa là dữ liệu thuộc các cụm có thể bị sai / không chính xác.
- Các đồ thị hình bóng có thể được sử dụng để chọn giá trị tối ưu nhất của K (số của cụm) trong phân cụm K-mean.
- Các khía cạnh cần chú ý trong biểu đồ Hình bóng là điểm số cụm thấp hơn điểm số hình bóng trung bình, sự dao động lớn về kích thước của các cụm và độ dày của ô hình bóng.

2.8 Các phương pháp đánh giá kết quả phân tích phân cụm

2.8.1 Tại sao phải đánh giá kết quả phân tích phân cụm

Để có kết quả phù hợp, chính xác, đáng tin cậy thì cần có phương pháp cụ thể để đánh giá kết quả đạt được sau khi tiến hành phân cụm dữ liệu. Tuy nhiên quyết định bao nhiêu cụm cần phân như thế nào để tối ưu nhất và các cụm có được khi kết thúc thuật toán clustering được đánh giá là phù hợp, chính xác, đáng tin cậy thì cực kỳ quan trọng và cần có phương pháp cụ thể. Nếu quá trình chọn k được thực hiện chỉ dựa trên kinh nghiệm phân tích, kiến thức chuyên môn, và mục đích kinh doanh mà không dựa trên chính đặc tính của dữ liệu thì khả năng cao việc ứng dụng clustering sẽ không mang lại giá trị như mong đợi khi các cluster có thể không phản ánh tốt các quy luật, các mối quan hệ những đối tượng quan sát trong tự nhiên đang tiềm ẩn trong tập dữ liệu. Vì thế việc đánh giá kết quả đạt được là vô cùng quan trọng.

2.8.2 Các phương pháp đánh giá kết quả phân cụm

Có một số phương pháp đánh giá phân cụm như sau:

- Đánh giá trong (internal evaluation): là phương pháp đánh giá kết quả dựa trên chính dữ liệu được phân cụm bằng cách sử dụng các đại lượng đánh giá sự gắn kết cụm như mật độ (density), khoảng cách giữa các phần tử trong cụm hay khoảng cách giữa các cụm với nhau.

- Đánh giá ngoài (external evaluation): là phương pháp đánh giá kết quả dựa vào tập dữ liệu chuẩn(dữ liệu mẫu) đã được phân cụm từ trước đó, còn được gọi là tập benchmark.

- Ngoài ra ta có thể đánh giá việc phân cụm bằng cách so sánh với các kết quả phân cụm khác được sinh ra bởi cùng một thuật toán nhưng với các giá trị tham số đầu vào khác nhau.

2.8.3 Các độ đo đánh giá trong kết quả phân cụm

a. Độ đo Silhouette

Chỉ số Silhouette Index là chỉ số đánh giá các kết quả phân cụm phổ biến và được sử dụng nhiều nhất. Phân tích chỉ số Silhouette mục đích để đo lường mức độ tối ưu khi một quan sát, một điểm dữ liệu được phân vào các cluster bất kỳ. Cụ thể, phương pháp Silhouette, với tên nghĩa tiếng Việt "hình bóng", sẽ cho chúng ta biết những điểm dữ liệu hay những quan sát nào nằm gọn bên trong cụm (tốt) hay nằm gần ngoài rìa cụm (không tốt) để đánh giá hiệu quả phân cụm.

Silhouette đo lường khoảng cách của một điểm dữ liệu trong cụm đến Centroid, điểm trung tâm của cụm, và khoảng cách của chính điểm đó đến điểm trung tâm của cụm gần nhất (hoặc đến các điểm trung tâm của các cụm còn lại, và chọn ra khoảng cách ngắn nhất). Đó là trường hợp đo lường cho K-means clustering.

Nếu các cluster được tìm không phải dựa trên clustering, thì Silhouette cũng sẽ đo lường theo cách tương tự nhưng thay vì tính khoảng cách giữa điểm đó với điểm trung tâm, thì chúng ta sẽ tính khoảng cách trung bình với tất cả các điểm còn lại trong cluster của điểm đó, và khoảng cách trung bình với tất cả các điểm còn lại của các cluster khác (lấy khoảng cách trung bình ngắn nhất).

Độ đo hình bóng được tính với công thức như sau:

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

Trong đó:

a(i) là khoảng cách từ điểm i trong cụm A đến điểm trung tâm của cụm A.

b(i) là khoảng cách từ điểm i trong cụm A đến điểm trung tâm của cụm B.

Cụm tương ứng với **b(i)** này được gọi là cụm hàng xóm của i.

Khi đó:

=> s(i) nằm trong đoạn **[-1,1]. s(i)** càng gần 1 thì node i càng phù hợp với cụm mà nó được phân vào. **s(i) = 0** thì không thể xác định được i nên thuộc về cụm nào giữa cụm hiện tại và cụm hàng xóm của nó. **s(i)** càng gần -1 thì chứng tỏ i bị phân sai cụm, nó nên thuộc về cụm hàng xóm chứ không phải cụm hiện tại.

b. Độ đo Davies – Bouldin

Để tính chỉ số DB, chúng ta phải đo lường mức độ phân tán và tương đồng của các cụm.

Độ đo Davies-Bouldin được tính theo công thức:

$$\mathbf{DB} = \frac{\mathbf{1}}{\mathbf{n}}\sum_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{Max_{i\neq j}}\left(\frac{\boldsymbol{\sigma_i} + \boldsymbol{\sigma_j}}{\mathbf{d(c_i, c_j)}}\right)$$

n: là số cụm.

$\mathbf{c_x}$**:** là trọng tâm của cụm x.

$\boldsymbol{\sigma_x}$**:** là trung bình khoảng cách của tất cả các phần tử trong cụm x tới trọng tâm $\mathbf{c_x}$

$\mathbf{d(c_i, c_j)}$ là khoảng cách giữa hai trọng tâm của cụm i và j.

=> Giá trị DB càng nhỏ thì chất lượng phân cụm càng tốt.

c. Độ đo Dunn

Độ đo Dunn, đây là phương pháp đánh giá cluster phổ biến khác sử dụng thông tin bên trong tập dữ liệu. Độ đo Dunn được là phương pháp đơn giản nhất. Cách đánh giá dựa trên việc so sánh giữa kích thước (size) của các cluster với khoảng cách giữa các cluster với nhau. Các cụm càng xa nhau, so với kích thước của chúng, chỉ số càng lớn thì kết quả phân cụm sẽ càng chính xác.

Độ đo Dunn được tính theo công thức:

$$\mathbf{D} = \mathbf{min_{1\leq i\leq n}}\left\{\mathbf{min_{i\leq j\leq n, i\neq j}}\left\{\frac{\mathbf{d(i,j)}}{\mathbf{max_{1\leq k\leq n} d'(k)}}\right\}\right\}$$

d(i, j) là khoảng cách giữa hai cụm i và j, thường được tính là khoảng cách giữa hai tâm cụm i và j.

d'(k) là khoảng cách trung bình bên trong cụm k.

n là số cụm.

=> D càng lớn thì phép chia cụm càng tốt.

Chương 3: ÁP DỤNG CÁC THUẬT TOÁN XÁC ĐỊNH SỐ CỤM TỐI ƯU VÀO BÀI TOÁN PHÂN KHÚC KHÁCH HÀNG SỬ DỤNG DỊCH VỤ DI ĐỘNG TẠI VNPT TÂY NINH

Trong chương này sẽ tiến hành thu thập dữ liệu và mô tả dữ liệu liên quan đến tình hình sử dụng dịch vụ di động của khách hàng. Tiến hành áp dụng các thuật toán phân cụm, các thuật toán lựa chọn số cụm tối ưu vào tập dữ liệu khách hàng sử dụng dịch vụ di động đã thu thập được. Sau đó đánh giá kết quả đạt được bằng các phương pháp đánh giá số cụm tối ưu. Trình bày kết quả đạt được. Phác thảo sơ đồ giải quyết bài toán như sau:

- Tiến hành thu thập dữ liệu về hành vi sử dụng dịch vụ di động của từng khách hàng trong tháng gần nhất.
- Mô tả dữ liệu thu thập được.
- Làm sạch dữ liệu nếu cần.
- Chọn các trường dữ liệu mục tiêu.
- Triển khai phân cụm đối tượng khách hàng dựa bằng thuật toán K-means, K-medoids. Áp dụng các thuật toán xác định số cụm tối ưu là: Elbow method và Average silhouette method.
- Đánh giá các kết quả đạt được trên từng thuật toán bằng các độ đo đánh giá số cụm tối ưu.
- Chọn ra được số cụm tối ưu nhất cho bài toán.

3.1. Giới thiệu

Để giải quyết bài toán yêu cầu, chúng tôi xây dựng một mã nguồn code với các đặc điểm như sau:

- CSDL: file csv trích xuất từ dữ liệu khách hàng.
- Ngôn ngữ lập trình: Python

Các phần mềm hỗ trợ:

- Phần mềm Weka phiên bản 3.8.6

3.2. Các thử nghiệm

- Thực hiện phân cụm khách hàng thành các cụm khác nhau cho 19 chu kỳ bằng 2 giải thuật xác định số cụm tối ưu cùng với phương pháp phân cụm k-means.
- Dùng phương pháp thực nghiệm để đánh giá tìm số cụm tối ưu.
- So sánh 2 giải thuật xác định cụm tối ưu trên các tiêu chí khác nhau.

3.3. Thu thập dữ liệu về hành vi sử dụng dịch vụ di động của khách hàng trong tháng gần nhất

Giai đoạn thu thập và xử lý dữ liệu ban đầu luôn là một giai đoạn quan trọng trong quy trình khai phá dữ liệu. Dữ liệu là một trong hai thành phần của phân lớp dữ liệu. Truy cập dữ liệu thực hiện việc trích xuất và thu thập dữ liệu cần thiết cho mô hình phân cụm khách hàng. Thông tin khách hàng cần thiết để phân cụm khách hàng gồm: quản lý dữ liệu khách hàng thuê bao, chi tiết dữ liệu sử dụng dịch vụ của thuê bao, thanh toán và khuyến mại của thuê bao. Từ các dữ liệu khác nhau, ta tiến hành phân cụm lần lượt dựa trên các tiêu chí đã chọn.

Dữ liệu thu thập được sau khi lọc và loại bỏ các thông tin không chính xác, không cần thiết thì gồm các thông tin:

- Dữ liệu quản lý khách hàng: loại thuê bao, bưu cục thu, thời gian hoạt động.

- Dữ liệu thanh toán: tiền phát sinh cho cuộc gọi, tiền phát sinh SMS, tiền phát sinh dữ liệu di động, tiền phát sinh các dịch vụ gia tăng khác, tổng tiền phát sinh từ tài khoản chính, tổng số tiền phát sinh, số tiền được khuyến mãi, tổng tiền còn lại trong tài khoản chính.

Sau đây là dữ liệu thu thập và tiền xử lý của tập dữ liệu khách hàng sử dụng di động trả trước do Vinaphone Tây Ninh quản lý.

SUBSCRIBER_ID	PROVI	BTS_NAME	TOTAL_CALL	TOTAL_SMS	TOTAL_DATA	TOTAL_VAS	TOTAL_OTHER	TOTAL_TKC	TOTAL_CORE_	TKC_CALL	TKC_SMS	TKC_DATA	TKC_VAS	TKC_OTHER	COS_GROUP	MA_CSHT
813634419	TNH	3G_CTH071M_TNH	0	0	0	0	0	0	35721.57	0	0	0	0	0	VINA690	CSHT_TNH_00547
914479709	TNH	3G_GDA026M_TNH	117288.98	700	158.4	0	0	118147.2	32437.26	117288.8	700	158.4	0	0	VINACARD_201	CSHT_TNH_00059
918864464	TNH	2G_CTH026M_TNH	107073.68	0	35.2	0	0	103107.2	19722.03	103072	0	35.2	0	0	MYZONE_KM_2	CSHT_TNH_00233
907671477	TNH	4G-TBA047M-TNH	136086.06	15849	140.8	0	0	128501.5	71466.86	112860.7	15500	140.8	0	0	VINA690	CSHT_TNH_00377
948976578	TNH	4G-TCH054M-TNH	0	0	0	0	0	0	115.23	0	0	0	0	0	VINA690	CSHT_TNH_00418
835517859		3G_HTH010M_TNH	23490.06	19950	0	1000	0	39728	59499.47	19585.6	19142.4	0	1000	0	VINACARD_KM	CSHT_TNH_00009
817543901	BTE	3G_GDA010M_TNH	0	0	0	0	0	0	11.65	0	0	0	0	0	VINACARD_KM	CSHT_TNH_00106
833343467	TNH	4G-HTH036M-TNH	8950.57	350	1925.2	0	0	11225.6	59403.78	8950.4	350	1925.2	0	0	MYZONE_KM_2	CSHT_TNH_00385
888475204	TNH	2G_GDA011M_TNH	3703	0	0	0	0	3703	26812.34	3703	0	0	0	0	VINACARD_KM	CSHT_TNH_00297
918561152	TNH	2G_TBA029M_TNH	1002.74	0	0	0	0	0	41150.17	0	0	0	0	0	VINACARD	CSHT_TNH_00093
918743508	TNH	3G_TBI039M_TNH	851	0	0	0	0	851	665.12	851	0	0	0	0	VINACARD	CSHT_TNH_00248
859343943	DLC	2G_HTH040M_TNH	31222.5	3495	0	1500	15440	51657.5	27	31222.5	3495	0	1500	15440	VINA690	CSHT_TNH_00053
946346657	TNH	2G_HTH016M_TNH	29153.19	0	0	0	0	29152.7	28269.65	29152.7	0	0	0	0	VINA690	CSHT_TNH_00226
855131280		4G-TNI062M-TNH	0	0	211.02	0	0	211	9789.04	0	0	211	0	0	VINA690	CSHT_TNH_00402
946586120	TNH	3G_CTH082M_TNH	0	0	0	0	0	0	3.96	0	0	0	0	0	VINAXTRA_201	CSHT_TNH_00489
949954482	TNH	2G_TBA013M_TNH	100896.7	0	0	0	0	96894.8	12476.59	96894.8	0	0	0	0	VINA690	CSHT_TNH_00224
854614958	TNH	2G_BCA008M_TNH	25075.12	0	0	9000	0	34075.1	32092.03	25075.1	0	0	9000	0	VINACARD_KM	CSHT_TNH_00312
843845497	HCM	2G_TNI041M_TNH	271324.22	0	0	9000	0	250323.8	26627.22	241323.8	0	0	9000	0	STUDENT_MIG	CSHT_TNH_00149
834667664	LAN	3G_TBA007M_TNH	59292.84	500	17050	190000	0	264024.6	1122.41	56474.6	500	17050	190000	0	VINAXTRA_201	CSHT_TNH_00108
823623598	TNH	2G_TBI044M_TNH	96372.67	6410	0	1500	0	96281.4	1.54	88721.4	6060	0	1500	0	MYZONE_KM	CSHT_TNH_00373
915925863	TNH	4G-TBA018M-TNH	2935.5	0	26.4	19000	0	21961.9	5.7	2935.5	0	26.4	19000	0	VINAXTRA_KM	CSHT_TNH_00191
943992397	HCM	2G_TBA027M_TNH	47537.06	0	0	71000	0	116565.3	2035.79	45565.3	0	0	71000	0	VINAXTRA_KM	CSHT_TNH_00185
919813287	HCM	2G_TNI022M_TNH	0	0	0	0	0	0	13.59	0	0	0	0	0	VINAXTRA_KM	CSHT_TNH_00142
819473662	TNH	2G_TNI005M_TNH	10035.64	0	0	300	0	10335.5	3096.65	10035.5	0	0	300	0	VINACARD_KM	CSHT_TNH_00029
941379936	TNH	4G-TBI034M-TNH	4753.27	0	8564.74	52200	0	63245.1	168.61	2480.4	0	8564.7	52200	0	VINA690	CSHT_TNH_00042
947756487	TNH	2G_HTH003M_TNH	39803.02	500	0	8000	0	48303	46152.23	39803	500	0	8000	0	VINAXTRA_KM	CSHT_TNH_00147

Hình 3.1: Dữ liệu thực tế vào tháng 11/2021

3.4. Mô tả dữ liệu thu thập được

- Dữ liệu được các tổng hợp từ các hệ thống thông tin của doanh nghiệp, trong đó dữ liệu cước phát sinh đã được xử lý nhiễu bằng một số biện pháp nghiệp vụ đặc thù như: băm, trùng, chờm, chẻ, áp giá, tính toán khuyến mãi, …

- Từ dữ liệu trên tiến hành làm sạch dữ liệu bằng cách loại bỏ các dòng dữ liệu có trường trống hoặc null, các trường dữ liệu xuất hiện nhiều giá trị 0 có ảnh hưởng đến quá trình chạy mô hình. Loại bỏ một số trường mang tính bảo mật người dùng: họ tên, địa chỉ, số điện thoại... Tiến hành chuyển đổi kiểu dữ liệu từ dạng chữ (chuỗi) sang dạng số bằng cách mã hóa các kí tự bằng số.

- Số dòng dữ liệu: 709,820 dòng

- Số trường dữ liệu: 14

Bảng 3.1: Mô tả từng trường dữ liệu

STT	Tên trường	Mô tả	Kiểu dữ liệu
1	SUBSCRIBER_ID	Số điện thoại	Ký tự
2	PROVINCE_CODE_DD	Mã tỉnh	Ký tự
3	TOTAL_CALL	Tổng chi phí thực hiện cuộc gọi	Số thực
4	TOTAL_SMS	Tổng chi phí thực hiện gửi tin nhắn SMS	Số thực

5	TOTAL_DATA	Tổng chi phí sử dụng dữ liệu di động	Số thực
6	TOTAL_VAS	Tổng chi phí sử dụng các DV Giá trị gia tăng	Số thực
7	TOTAL_OTHER	Tổng chi phí khác	Số thực
8	TOTAL_TKC	Tổng chi phí đã sử dụng trong tài khoản chính	Số thực
9	TOTAL_CORE_BALANCE	Tổng tiền còn lại	Số thực
10	TKC_CALL	Tổng tiền đã sử dụng TKC để thực hiện cuộc gọi	Số thực
11	TKC_SMS	Tổng tiền đã sử dụng TKC để thực hiện gửi tin nhắn SMS	Số thực
12	TKC_DATA	Tổng tiền đã sử dụng TKC cho dữ liệu di động	Số thực
13	TKC_VAS	Tổng tiền đã sử dụng TKC các DV Giá trị gia tăng	Số thực
14	MA_CSHT	Mã khách hàng tại tỉnh	Ký tự

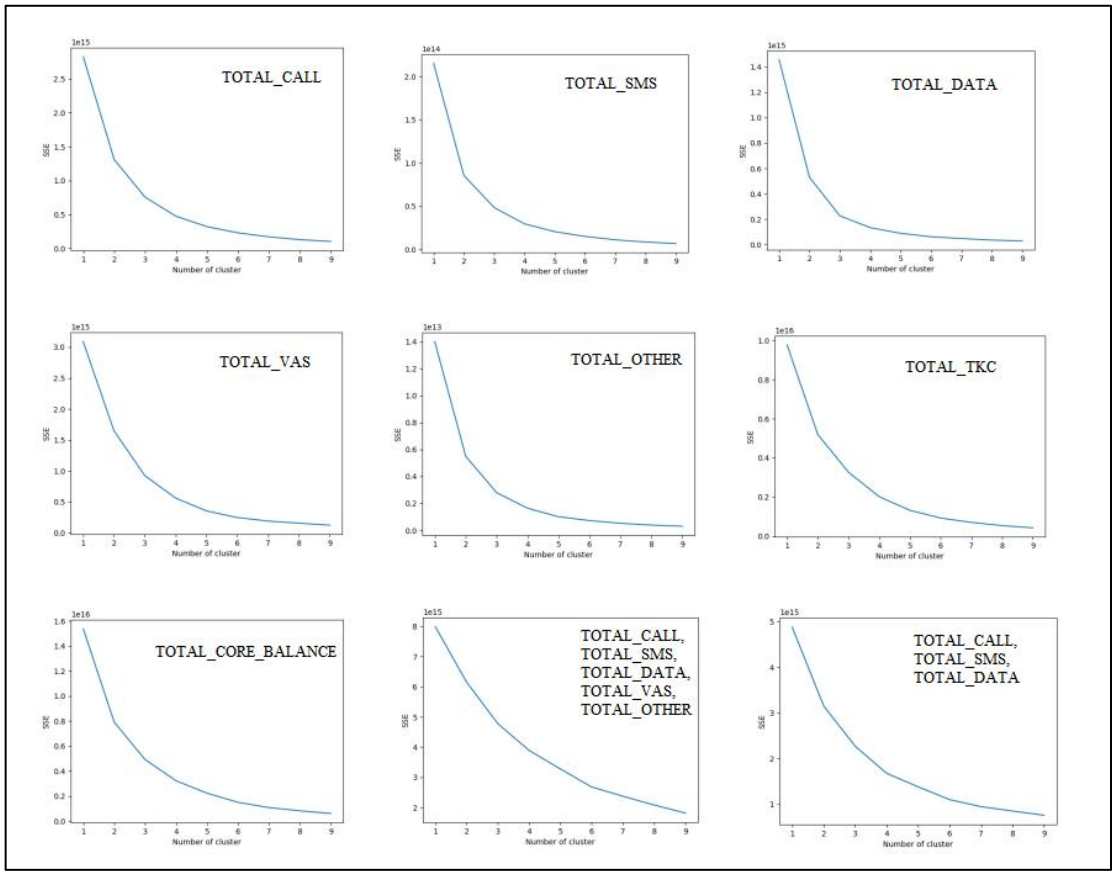
Bảng 3.2: Giá trị min - max, và trung bình của từng trường

STT	Tên trường	Min (VNĐ)	Max (VNĐ)	Trung bình (VNĐ)
1	SUBSCRIBER_ID			
2	PROVINCE_CODE_DD			
3	TOTAL_CALL	0	3,358,522.64	35,965.15
4	TOTAL_SMS	0	1,456,550.00	3,039.27
5	TOTAL_DATA	0	1,320,000.07	23,282.08
6	TOTAL_VAS	0	7,736,500.00	13,701.85
7	TOTAL_OTHER	0	672,370.00	451.32
8	TOTAL_TKC	0	9,050,749.20	73,638.12

9	TOTAL_CORE_BALANCE	0	20,891,749.61	35,073.39
10	TKC_CALL	0	3,298,520.50	33,425.74
11	TKC_SMS	0	1,390,144.10	2,796.51
12	TKC_DATA	0	1,319,999.90	23,264.93
13	TKC_VAS	0	7,736,500.00	13,699.62
14	TKC_OTHER	0	672,370.00	451.32

3.5. Tiến hành phân cụm bằng k-means và tìm kiếm số cụm tối ưu bằng Elbow method và Silhouette Score method

3.5.1 Kết quả xác định số cụm tối ưu khi sử dụng Phương pháp khủy tay(Elbow method) trên tập dữ liệu



Hình 3.2: Biểu đồ hiển thị kết quả xác định số cụm tối ưu bằng phương pháp khủy tay

Bảng 3.3: Kết quả xác định số cụm tối ưu khi sử dụng Phương pháp khủy tay

STT	Tên trường	Số dòng dữ liệu	Số cụm tối ưu với phương thức Elbow
Thực hiện phân cụm từng trường dữ liệu			
1	TOTAL_CALL	709,820	**3**
2	TOTAL_SMS	709,820	**3**
3	TOTAL_DATA	709,820	**3**
4	TOTAL_VAS	709,820	**3**
5	TOTAL_OTHER	709,820	**3**
6	TOTAL_TKC	709,820	**4**
7	TOTAL_CORE_BALANCE	709,820	**3**
Thực hiện phân cụm kết hợp các trường dữ liệu			
8	'TOTAL_CALL', 'TOTAL_SMS', 'TOTAL_DATA', 'TOTAL_VAS', 'TOTAL_OTHER'	709,820	**6**
9	'TOTAL_CALL', 'TOTAL_SMS', 'TOTAL_DATA'	709,820	**4**

3.5.2 Kết quả xác định số cụm tối ưu khi sử dụng phương pháp điểm hình bóng(Silhouette Score) trên tập dữ liệu

Bảng 3.4: Kết quả xác định số cụm tối ưu khi sử dụng Phương pháp điểm hình bóng(Silhouette Score)

STT	Tên trường	Số dòng dữ liệu	Số cụm tối ưu với phương thức điểm hình bóng(Silhouette Score)
Thực hiện phân cụm từng trường dữ liệu			
1	TOTAL_CALL	709,820	**3**
2	TOTAL_SMS	709,820	**3**
3	TOTAL_DATA	709,820	**19**
4	TOTAL_VAS	709,820	**3**
5	TOTAL_OTHER	709,820	**5**
6	TOTAL_TKC	709,820	**6**
7	TOTAL_CORE_BALANCE	709,820	**3**
Thực hiện phân cụm kết hợp các trường dữ liệu			
8	'TOTAL_CALL', 'TOTAL_SMS', 'TOTAL_DATA', 'TOTAL_VAS', 'TOTAL_OTHER'	709,820	**3**
9	'TOTAL_CALL', 'TOTAL_SMS', 'TOTAL_DATA'	709,820	**6**

3.5.3 So sánh kết quả lựa chọn cụm tối ưu giữa hai phương pháp Khủy tay và phương pháp tính điểm Silhouette

Bảng 3.5: So sánh kết quả của hai phương pháp

STT	Tên trường	Số cụm tối ưu với phương thức Khủy tay	Số cụm tối ưu với cách tính điểm Silhouette(độ đo Euclidean)
Thực hiện phân cụm từng trường dữ liệu			
3	TOTAL_CALL	**3**	**3**
4	TOTAL_SMS	**3**	**3**
5	TOTAL_DATA	**3**	**19**
6	TOTAL_VAS	**3**	**3**
7	TOTAL_OTHER	**3**	**5**
8	TOTAL_TKC	**4**	**6**
9	TOTAL_CORE_BALANCE	**3**	**3**
Thực hiện phân cụm kết hợp các trường dữ liệu			
11	'TOTAL_CALL', 'TOTAL_SMS', 'TOTAL_DATA', 'TOTAL_VAS', 'TOTAL_OTHER'	**6**	**3**
12	'TOTAL_CALL', 'TOTAL_SMS', 'TOTAL_DATA'	**4**	**6**

3.5.4 Tiến hành phân cụm với số lượng cụm tối ưu thu thập được cùng với đó áp dụng thuật toán K-Means++ để khởi tạo tâm cụm và phân cụm

Các trường chủ yếu được sử dụng trong phân khúc khách hàng tập trung vào các trường quan trọng sau: TOTAL_CALL (Tổng chi phí gọi), TOTAL_SMS(Tổng chi phí gửi tin nhắn SMS), TOTAL_DATA(Tổng chi phí sử dụng dữ liệu di động). 3 trường này được sử dụng vì đây là 3 trường dữ liệu chính của một tài khoản di động, nó phản ánh mức độ sử dụng dịch vụ của khách hàng.

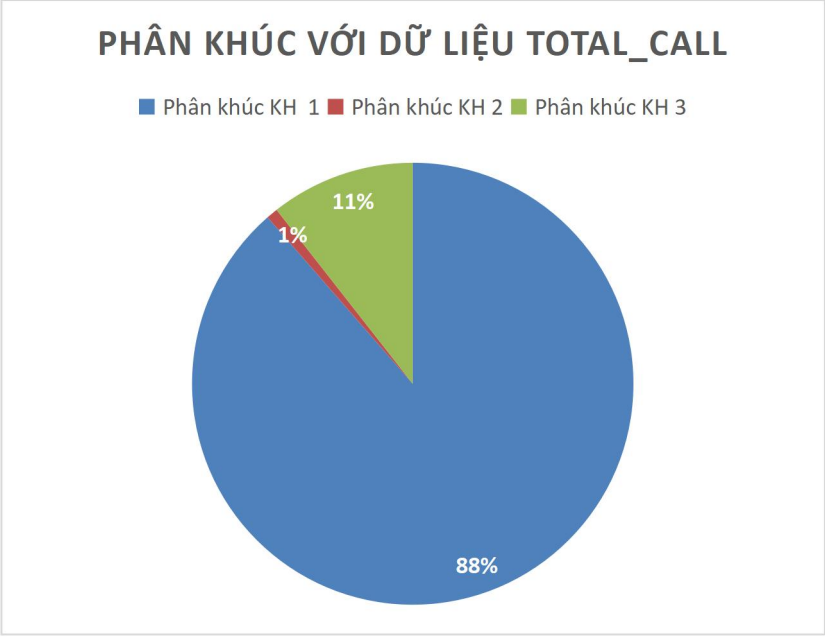
a. Phân cụm với thuộc tính TOTAL_CALL

Số lượng cụm: 3

Thuật toán: K-Means++

Bảng 3.6: Phân khúc với thuộc tính TOTAL_CALL(đơn vị: ngìn đồng)

Thuộc tính	**Phân khúc 1** (628,717 thuê bao)	**Phân khúc 2** (5,972 thuê bao)	**Phân khúc 3** (75,132 thuê bao)
TOTAL_CALL	16,338 +/-20,127	509,455 +/-247,249	137,184 +/-56,841



Hình 3.3: Tỉ lệ phân khúc khách hàng theo tổng chi phí cuộc gọi

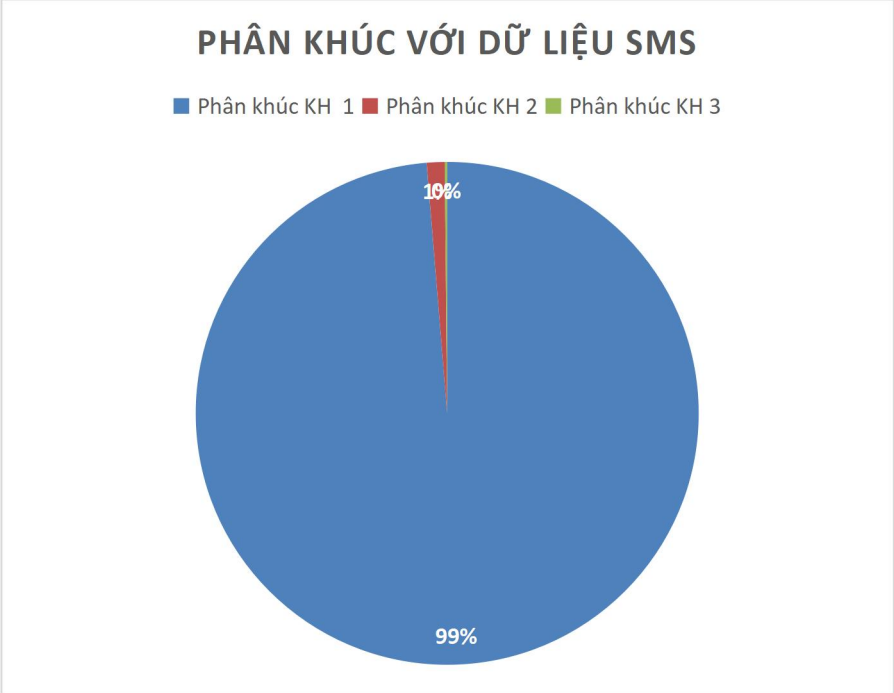
b. Phân cụm với thuộc tính TOTAL_SMS

Số lượng cụm: 3

Thuật toán: K-Means++

Bảng 3.7: Phân khúc với thuộc tính TOTAL_SMS(đơn vị tính: VNĐ)

Thuộc tính	Phân khúc 1 (700,563 thuê bao)	Phân khúc 2 (8,110 thuê bao)	Phân khúc 3 (1,148 thuê bao)
TOTAL_SMS	1,370 +/-40,96	82,328 +/-38,833	318,333 +/-146,213



Hình 3.4: Tỉ lệ phân khúc khách hàng theo tổng chi phí sms

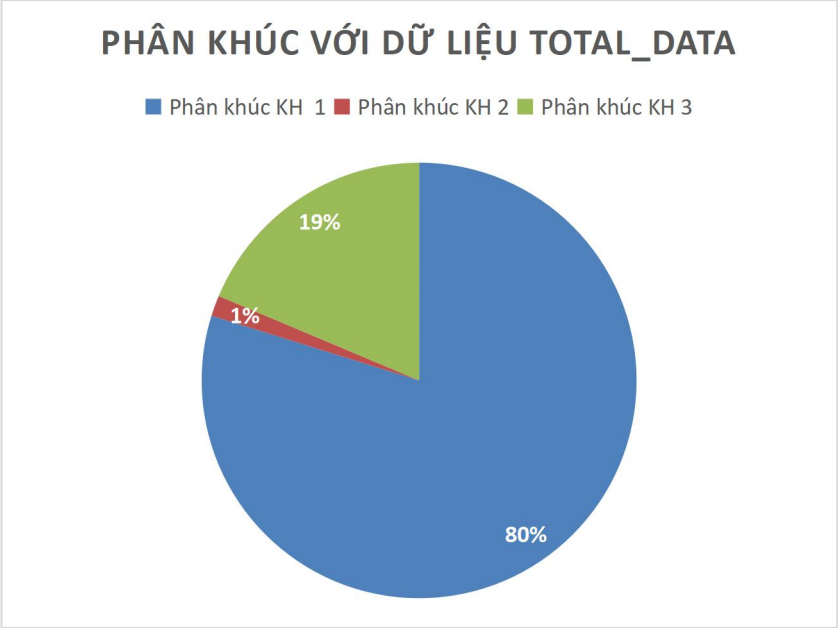
c. Phân cụm với thuộc tính TOTAL_DATA

Số lượng cụm: 3

Thuật toán: K-Means++

Bảng 3.8: Phân khúc với thuộc tính TOTAL_DATA(đơn vị tính: VNĐ)

Thuộc tính	Phân khúc 1 (566,482 thuê bao)	Phân khúc 2 (10,864 thuê bao)	Phân khúc 3 (132,475 thuê bao)
TOTAL_DATA	2,186 +/-6,873	255,210 +/-97,347	78,312 +/-26,979



Hình 3.5: Tỉ lệ phân khúc khách hàng theo tổng chi phí gọi

d. Phân cụm với thuộc tính TOTAL_CALL, TOTAL_SMS, TOTAL_DATA

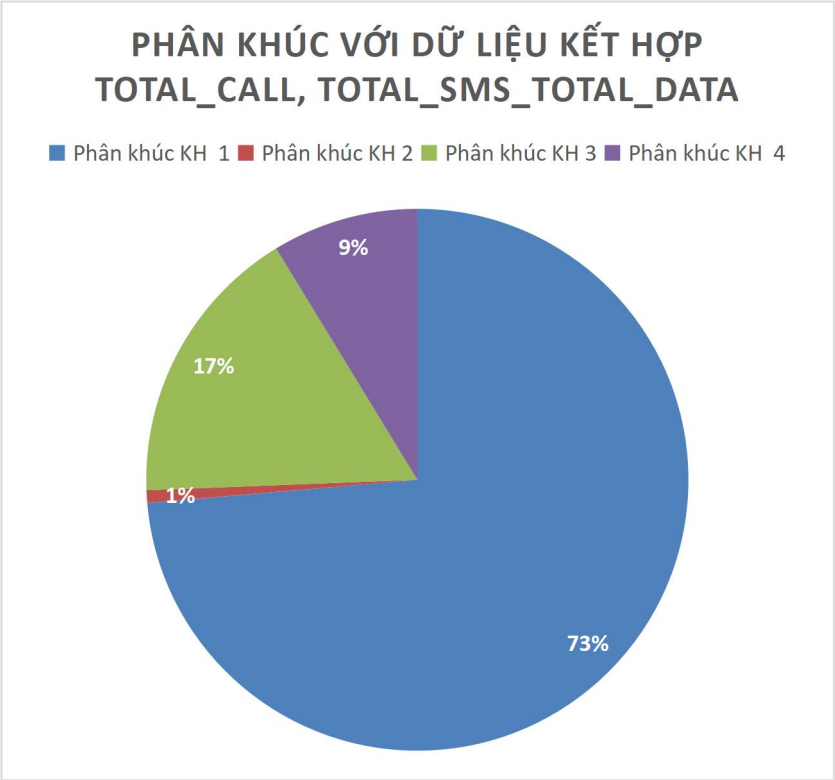
Số lượng cụm: 4

Thuật toán: K-Means++

Bảng 3.9: Phân khúc với thuộc tính TOTAL_CALL, TOTAL_SMS, TOTAL_DATA (đơn vị tính: VNĐ)

Thuộc tính	Phân khúc 1 (522,646 thuê bao)	Phân khúc 2 (5,342 thuê bao)	Phân khúc 3 (119,915 thuê bao)	Phân khúc 4 (61,918 thuê bao)
TOTAL_CALL	13,502	527,403	43,151	138,455

	+/-19,165	+/-255,232	+/-44,015	+/-61,024
TOTAL_SMS	999 +/-5,472	29,186 +/-86,251	4,438 +/-16,817	12,635 +/-43,003
TOTAL_DATA	3,757 +/-10,649	36,150 +/-67,776	99,466 +/-61,396	4,864 +/-13,237



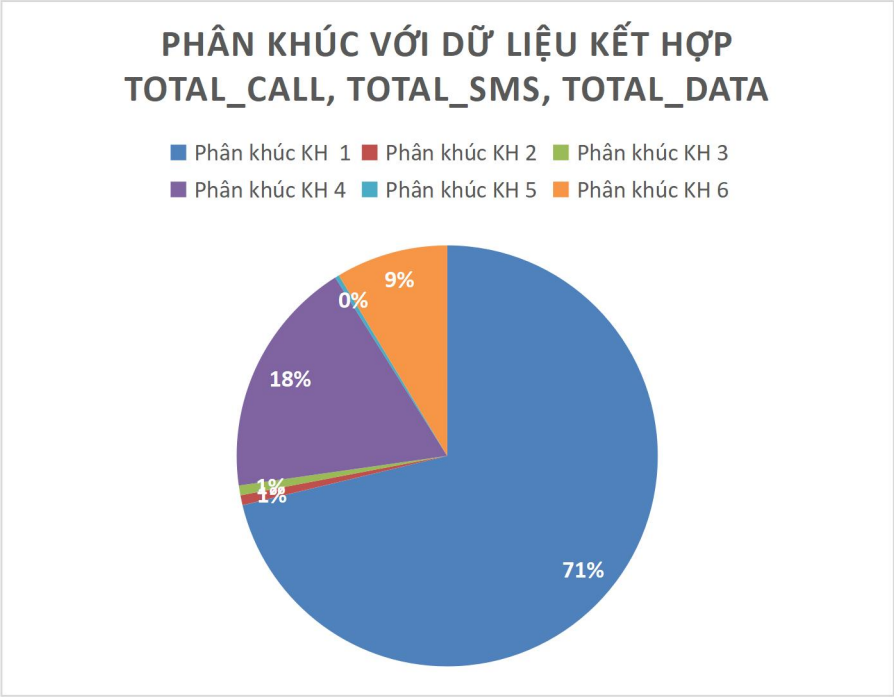
Hình 3.6: Tỉ lệ phân khúc khách hàng theo tổng chi phí dữ liệu di động

Số lượng cụm: 6

Thuật toán: K-Means++

Bảng 3.10: Phân khúc với thuộc tính TOTAL_CALL, TOTAL_SMS, TOTAL_DATA(đơn vị tính: VNĐ)

Thuộc tính	Phân khúc 1 (501,450 thuê bao)	Phân khúc 2 (5,342 thuê bao)	Phân khúc 3 (5,310 thuê bao)	Phân khúc 4 (128,981 thuê bao)	Phân khúc 5 (2,251 thuê bao)	Phân khúc 6 (60,591 thuê bao)
TOTAL_CALL	13,134 +/-18,710	520,751 +/-256,049	68,430 +/-91,888	37,995 +/-38,002	112,624 +/-126,998	137,763 +/-58,788
TOTAL_SMS	1,040 +/-5693	15,511 +/-32203	6,832 +/-21316	3,264 +/-8698	236,618 +/-133043	5,919 +/-14533
TOTAL_DATA	2,148 +/-6,802	28,905 +/-49,872	251,469 +/-97,285	77,965 +/-26,480	24,502 +/-47,550	3,593 +/-10,771



Hình 3.7: Tỉ lệ phân khúc khách hàng theo tổng chi phí

3.6 Đánh giá kết quả phân khúc khách hàng

Theo như kết quả đạt được ở chương này ta rút ra được một số kết quả sau:

+ Ở tất cả các trường hợp phân khúc trên, lượng khách hàng không sử dụng hoặc ít sử dụng các dịch vụ có phát sinh chi phí đang chiếm tỉ lệ cao nhất(>70%). Chi phí phát sinh hàng tháng đối với nhóm này là <45,000VNĐ/tháng. Trong đó Cuộc gọi <30,000; SMS<5,000 VNĐ/tháng; Data < 8,000 VNĐ/tháng. Đây là nhóm khách hàng còn rất nhiều tiềm năng để phát triển và là nhóm khách hàng trọng tâm khi triển khai các chương trình khuyến mãi, các gói tiện ích để nâng cao doanh thu cho đơn vị. Tuy nhiên đây cũng là nhóm khách hàng chúng ta cần xem xét về vấn đề có thể rời mạng, nhất là các thuê bao không phát sinh chi phí ở nhiều kỳ liên tiếp.

+ Với những phân khúc khách hàng thuộc nhóm có chi phí sử dụng dịch vụ hàng tháng rơi vào khoảng > 500,000VNĐ/tháng. Đây là những đối tượng khách hàng VIP của đơn vị. Đối với những đối tượng này chúng ta cần tập trung nâng cao chất lượng dịch vụ, thường xuyên chăm sóc khách hàng để khách hàng cảm thấy hài lòng nhất với chi phí mà họ đã bỏ ra.

Về việc lựa chọn số lượng phân khúc khách hàng cho từng tiêu chí sử dụng dịch vụ:

+ Đối với riêng từng dịch vụ Gọi di động, gửi SMS, và dữ liệu di động ta có thể chọn số cụm tối ưu là 3 và xây dựng các model riêng cho từng trường dữ liệu.

+ Đối với việc kết hợp cả 3 trường Gọi di động, gửi tin nhắn SMS, và dữ liệu di động ta có 2 phương án chọn phân khúc là 4 và 6.

Chương 4. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Nội dung chính trong chương này tập trung đánh giả tổng quan về các kết quả đã đạt được trong luận văn, trình bày các hạn chế mà luận văn gặp phải để tìm hướng khắc phụ. Và qua đó tìm ra hướng phát triển tốt hơn trong trong tương lai.

4.1 Kết luận

Đề tài trình bày một cách khái quát về KPDL, các phương pháp KPDL và ứng dụng kỹ thuật phân cụm và một số kỹ thuật KPDL khác để hỗ trợ cho doanh nghiệp viễn thông.

Luận văn đã trình bày tổng quan về bài toán phân khúc khách hàng sử dụng dịch vụ di động nói chung dựa trên chi phí sử dụng hàng tháng ở các nhóm dịch vụ khác nhau. Từ đó dựa trên các tiêu chí mà chọn ra số lượng phân khúc tối ưu đối với từng đối tượng khách hàng sử dụng dịch vụ. Doanh nghiệp có thể dễ dàng triển khai các chính sách, ưu đãi, khuyến mãi trên từng đối tượng khách hàng đem lại tính hợp lý và thỏa mãn từng nhóm khách hàng.

Luận văn đã giải quyết được bài toán xác định số cụm tối ưu trong phân khúc khách hàng đối với tường loại dịch vụ riêng biệt. Dữ liệu sử dụng trong luận văn được thu thập thực tế trên dữ liệu về chi phí của thuê bao di động trả trước của Vinaphone trên địa bàn tỉnh Tây Ninh.

Từ việc nghiên cứu những yêu cầu đặt ra trong công tác phân khúc khách hàng của nhà mạng di động Vinaphone, luận văn đã đạt được một số kết quả chính sau đây:

- Xây dựng được mô hình phân khúc khách hàng, có thể áp dụng theo khoảng thời gian mong muốn.
- Triển khai mô hình đề xuất, áp dụng trên dữ liệu thực tế, so sánh với các giải pháp đã sử dụng được áp dụng. Các kết quả đạt được đã cho thấy được tiềm năng áp dụng phương pháp đề xuất vào thực tiễn.

Trong thời gian tới chúng tôi sẽ nghiên cứu tích hợp các kỹ thuật này vào các chương trình hỗ trợ kinh doanh của Vinaphone Tây Ninh bên cạnh đó cũng đồng thời cải tiến thuật toán phân khúc khách hàng để cho kết quả nhanh và chính xác hơn.

4.2 Hạn chế của đề tài và hướng phát triển trong tương lai

Phạm vi của nghiên cứu này chỉ tập trung tại Tỉnh Tây Ninh, và trong một khoảng thời gian nhất định nên kết quả chỉ mang tính chất tương đối, tính khái quát chưa cao, có thể đúng trong thời gian nghiên cứu, nhưng về lâu dài thì có thể không còn đáp ứng được. Để có những quyết định chính xác hơn cần có những nghiên cứu thêm trong tương lai.

Với rất nhiều ứng dụng thực tế của khai phá dữ liệu trong ngành viễn thông, đặc biệt là về phân khúc khách hàng. Với thời gian có hạn luận văn mới chỉ nghiên cứu và thực nghiệm trên 2 thuật toán xác định số cụm tối ưu để phân khúc khách hàng, vì vậy yêu cầu với bài toán trong tương lai là áp dụng các thuật toán khác như hồi quy dự báo, áp dụng mạng nơron xây dựng các mô hình dự báo…

Với sự ứng dụng rộng rãi của khai phá dữ liệu trong ngành viễn thông như đã trình bày thì còn rất nhiều bài toán có thể tìm hiểu và nghiên cứu thêm trong tương lai như: Xác định thuê bao rời mạng, Phân khúc khách hàng với nhiều tiêu chí hơn(thời gian thực hiện cuộc gọi, chi phí từng cuộc gọi, vị trí thực hiện cuộc gọi,…) .

DANH MỤC TÀI LIỆU THAM KHẢO

[1] Bộ Thông tin và Truyền thông. *Sách Trắng về Công nghệ thông tin và Truyền thông (CNTT-TT) Việt Nam 2021*. Hà Nội: Nhà xuất bản Thông tin và Truyền thông, 2021, tr.21.

[2] Bruce Cooil, Lerzan Aksoy, Timothy L. Keiningham (2006), "Approaches to Customer Segmentation", *Journal of Relationship Marketing, pp. 3-4.*

[3] A.K. Jain, "*Data clustering: 50 years beyond k-means*", Pattern recognition letters, vol.31, no.8, pp.651–666, 2010.

[4] H.S. Park and C.H. Jun, "*A simple and fast algorithm for k-medoids clustering*" , Expert systems with applications, vol.36, no.2, pp.3336–3341, 2009.

[5] Edy Umargono, Jatmiko Endro Suseno, Vincensius Gunawan S.K2 (2019) "*K-Means Clustering Optimization Using the Elbow Method and Early Centroid Determination Based on Mean and Median Formula*", Department of Information System, Post Graduated School, Diponegoro University University of Science and Technology, China.

[6] Daniel T. Larose (2005), "*Discovering Knowledge in Data - An Introduction to Data Mining"*, Wiley, Hoboken, New Jersey.

[7] R.Agrawal and R.Srikant (1995), "Mining sequential patterns", Proceedings of the Eleventh International Conference on Data Engineering.

[8] R Ragavi, B Srinithi (2018) "*Data Mining Issues and Challenges: A Review*", Sri Krishna Arts and Science College, Coimbatore (Vol. 7, Issue 11, November 2018), *pp. 118-121.*

[9] Parul Agarwal (2011) "*Issues, Challenges and Tools of Clustering Algorithms*"- Department of Computer Science, Jamia Hamdard, *pp. 4-5.*

[10] D. Xu and Y. Tian (2015), "A comprehensive survey of clustering algorithms", Annals of Data Science, vol.2, no.2, pp.165–193.

[11] Nguyễn Đức Thắng, Lê Văn Chiến, Nguyễn Văn Thưởng, Phạm Kiên Trung (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, Kỳ 5 2020), tr.145-150.

[12] Nguyễn Văn Chức, Đào Thị Giang (2015), "*Ứng dụng kỹ thuật phân cụm và luật kết hợp khai phá dữ liệu khách hàng sử dụng dịch vụ khách sạn*" - Tạp chí KHCN ĐHĐN (Số 12(97).2015), tr. 1-4.

PHỤ LỤC

MÃ NGUỒN

Báo cáo sử dụng Ngôn ngữ Python để tính toán các giá trị và cài đặt chương trình. Phần này trình bày các đoạn mã nguồn được thực hiện trong báo cáo

Code cài đặt thuật toán Kmean cho từng trường dữ liệu với elbow method

```
import pandas as pd
import os
from sklearn.cluster import KMeans
import matplotlib.pyplot as plt
import numpy as np
from sklearn.metrics import silhouette_score
def clean_dataset(df):
    assert isinstance(df, pd.DataFrame), "df needs to be a pd.DataFrame"
    df.dropna(inplace=True)
    indices_to_keep = ~df.isin([np.nan, np.inf, -np.inf]).any(1)
    return df[indices_to_keep].astype(np.float64)
fileDir = os.path.dirname(os.path.realpath('__file__'))
datafile = '.\\data\\dulieu_result.csv'
datafile = os.path.join(fileDir, datafile)
# load the CSV file as a numpy matrix
dataset = pd.read_csv(datafile, delimiter=",")
print(dataset.shape)
# separate the data from the target attributes
df = pd.DataFrame(dataset, columns= ['TOTAL_CALL', 'TOTAL_SMS',
```

```
'TOTAL_DATA', 'TOTAL_VAS', 'TOTAL_OTHER', 'TOTAL_TKC', 'TOTAL_CORE_BALANCE', 'TKC_CALL', 'TKC_SMS', 'TKC_DATA', 'TKC_VAS', 'TKC_OTHER'])
#df = pd.DataFrame(dataset, columns= ['TOTAL_SMS'])
clean_dataset(df)
#data = df[['TOTAL_CALL','TOTAL_SMS','TOTAL_DATA','TOTAL_VAS','TOTAL_OTHER','TOTAL_TKC','TOTAL_CORE_BALANCE','TKC_CALL','TKC_SMS','TKC_DATA','TKC_VAS','TKC_OTHER']]
data = df[['TOTAL_CALL','TOTAL_DATA','TOTAL_SMS']].head(120000)
print(data[:0-10])
sse = {}
for k in range(1, 10):
    kmeans = KMeans(n_clusters=k, max_iter=1000).fit(data)
    data["clusters"] = kmeans.labels_
    #print(data["clusters"])
    label = kmeans.labels_
      sse[k] = kmeans.inertia_ # Inertia: Sum of distances of samples to their closest cluster center
    #sil_coeff = silhouette_score(data, label, metric='euclidean')
    #print("For n_clusters={}, The Silhouette Coefficient is {}".format(k, sil_coeff))
plt.figure()
plt.plot(list(sse.keys()), list(sse.values()))
plt.xlabel("Number of cluster")
plt.ylabel("SSE")
plt.show()
```

Code cài đặt thuật toán Kmean cho từng trường dữ liệu với Silhouette Score

```
import pandas as pd
import os
from sklearn.cluster import KMeans
import matplotlib.pyplot as plt
import numpy as np
from sklearn.metrics import silhouette_score


def clean_dataset(df):
    assert isinstance(df, pd.DataFrame), "df needs to be a pd.DataFrame"
    df.dropna(inplace=True)
    indices_to_keep = ~df.isin([np.nan, np.inf, -np.inf]).any(1)
    return df[indices_to_keep].astype(np.float64)
fileDir = os.path.dirname(os.path.realpath('__file__'))
datafile = '.\\data\\dulieu_result.csv'
datafile = os.path.join(fileDir, datafile)
# load the CSV file as a numpy matrix
dataset = pd.read_csv(datafile, delimiter=",")
print(dataset.shape)
# separate the data from the target attributes
df = pd.DataFrame(dataset, columns = ['TOTAL_CALL', 'TOTAL_SMS', 'TOTAL_DATA', 'TOTAL_VAS', 'TOTAL_OTHER', 'TOTAL_TKC', 'TOTAL_CORE_BALANCE', 'TKC_CALL', 'TKC_SMS', 'TKC_DATA', 'TKC_VAS', 'TKC_OTHER'])
#df = pd.DataFrame(dataset, columns= ['TOTAL_SMS'])
clean_dataset(df)
```

```
#data =
df[['TOTAL_CALL','TOTAL_SMS','TOTAL_DATA','TOTAL_VAS','TOTAL_OTHE
R','TOTAL_TKC','TOTAL_CORE_BALANCE','TKC_CALL','TKC_SMS','TKC_DA
TA','TKC_VAS','TKC_OTHER']]
data = df[['TOTAL_CALL']].head(50000)


for n_cluster in range(2, 11):
    kmeans = KMeans(n_clusters=n_cluster).fit(data)
    label = kmeans.labels_
    sil_coeff = silhouette_score(data, label, metric='euclidean')
        print("For n_clusters={}, The Silhouette Coefficient is {}".format(n_cluster,
sil_coeff))
```