HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

Trần Thành Nguyên

XÁC ĐỊNH SỐ CỤM TỐI ƯU VÀO BÀI TOÁN PHÂN KHÚC KHÁCH HÀNG SỬ DỤNG DỊCH VỤ DI ĐỘNG TẠI VNPT TÂY NINH

Chuyên ngành: Hệ Thống thông tin
Mã số: 8.48.01.04

TÓM TẮT LUẬN VĂN THẠC SĨ
(Theo định hướng ứng dụng)

THÀNH PHỐ HỒ CHÍ MINH - NĂM 2022

Luận văn được hoàn thành tại:
HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

Người hướng dẫn khoa học: PGS. TS Nguyễn Đình Thuân

Phản biện 1: ……..

Phản biện 2: ……..

Luận văn sẽ được bảo vệ trước Hội đồng chấm luận văn thạc sĩ tại Học viện Công nghệ Bưu chính Viễn thông
Vào lúc: ..... giờ ..... ngày ..... tháng ..... năm .....

Có thể tìm hiểu luận văn tại:
- Thư viện của Học viện Công nghệ Bưu chính Viễn thông

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.

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**"

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 VNPT 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.

Luận văn được trình bày thành 4 chương:

Chương 1. Tổng quan: Giới thiệu Bài toán phân khúc khách hàng dựa trên hành vi sử dụng dịch vụ di động. Xác định mục tiêu, nội dung và phương pháp nghiên cứu của đề tài.

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.

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.

Chương 4: Kết luận và khuyến nghị: Đánh giá về các kết quả đạt được và hướng phát triển tiếp theo của đề tài.

Chương 1. TỔNG QUAN

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 có tên cho mảng dịch vụ di động là Vinaphone

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. Kết quả của việc cạnh tranh 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.

Để giữ chân 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

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

Để có kết quả phân khúc khách hàng nhanh và chính xác thì việc ứng dụng các công nghệ mới, đặc biệt là khai phá dữ liệu là phương án tối ưu nhất. Trong đó phân cụm dữ liệu là phương pháp để phân khúc khách hàng nhanh và mạnh mẽ. Vì thế, trong luận văn này sẽ sử dụng các thuật toán phân cụm để tiến hành phân khúc khách hàng.

Song song với đó, thì việc tìm ra số lượng cụm tối ưu nhất cũng là một vấn đề quan trọng. Vì vậy, luận văn này cũng sẽ trình bày và ứng dụng các thuật toán xác định số cụm tối ưu vào bài toán phân khúc khách hàng.

1.2 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 Vinaphone Tây Ninh
- Các giải thuật phân cụm trong khai phá dữ liệu
- Các thuật toán về xác định số cụm tối ưu

1.3 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

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.

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.

Những 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[6]

Quá trình khám phá tri thức, gồm các bước:

Bước 1. Phát triển và hiểu về ứng dụng

Bước 2. Lựa chọn dữ liệu mục tiêu

Bước 3. Làm sạch và tiền xử lý dữ liệu

Bước 4. Giảm và chiếu dữ liệu

Bước 6. Lựa chọn thuật toán khai thác dữ liệu.

Bước 7. Khai phá dữ liệu(Data mining).

Bước 8. Đánh giá và biểu diễn tri thức.

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…

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. 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
- Bước 2: Xác định các dữ liệu, dữ kiện liên quan
- Bước 3: Thu thập và tiền xử lý dữ liệu
- Bước 4: Tiến hành khai phá bằng thuật toán KPDL

2.2 Quá trình khám phá tri thức, 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).

- 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.

- 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.

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. 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.

- 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. 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. 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.

- 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

- 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.

- 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.

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.

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, …

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.

- Chọn độ đo: Ứng với từng phương pháp phân cụm khác nhau mà ta lựa chọn 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.

- 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.

- Giải thích kết quả: Dựa vào kinh nghiệm thực tế và kết quả phân cụm vừa đạt được, 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)
- …

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

- 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.

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

- Khoảng cách giữa hai cụm được xác định là khoảng cách ngắn nhất giữa hai điểm trong mỗi cụm.(Single linkage);

- Khoảng cách giữa hai cụm được xác định là khoảng cách xa nhất giữa hai điểm trong mỗi cụm. (Complete 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. (Average linkage).

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

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

K-Means là một trong những thuật toán phân cụm phân hoạch được sử dụng phổ biến nhất là thuật toán phân cụm dễ cài đặt, và cho ra kết quả dễ hiểu. Nhược điểm là dễ bị ảnh hưởng bởi các phần tử nhiễu, ngoại lệ.

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.

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.

Các phương pháp phân cụm dựa trên mật độ là rất tốt vì chúng không chỉ định trước số lượng các cụm.

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.

Thuật toán phân cụm dựa trên lưới bao gồm năm bước cơ bản sau (Grabusts và Borisov, 2002):

B1: Tạo cấu trúc lưới.

B2: Tính mật độ ô cho mỗi ô.

B3: Sắp xếp các ô theo mật độ của chúng.

B4: Xác định các trung tâm cụm.

B5: Truyền qua các ô lân cận.

e. Phương pháp phân cụm có dữ liệu ràng buộc

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 [9]:

- Dữ liệu bị nhiễu và không đầy đủ.
- Làm sạch và tiền xử lý dữ liệu.

- Quá khớp (Overfitting)
- Dữ liệu đa dạng và không đồng nhất.
- Thông tin hạn chế.
- Qui mô dữ liệu.
- Việc kết hợp các kiến thức nền.
- Trực quan hóa dữ liệu.
- Tốc độ/tính chuyển động liên tục.
- Ngôn ngữ truy vấn khai phá dữ liệu.
- Bảo mật dữ liệu riêng tư.
- Giải thích kết quả.

b. Các thách thức trong phân cụm dữ liệu[10]:

- Xác định cách tính khoảng cách.
- Xác định số lượng cụm.
- Thiếu nhãn.
- Cấu trúc của cơ sở dữ liệu.
- Các loại thuộc tính trong cơ sở dữ liệu
- Chọn tâm cụm khởi tạo ban đầu.

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.

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

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 , i = 1 ÷ n, sao cho hàm tiêu chuẩn:

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

đạt giá trị tối thiểu, trong đó là trọng tâm của cụm , 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.

- Vị trí tâm của cụm sẽ bị phụ thuộc vào điểm khởi tạo ban đầu của chúng.

2.6 Thuật toán K-Means++

Để giải quyết nhược điểm trên của thuật toán K-mean là nhạy cảm với việc khởi tạo các trọng tâm thì thuật toán xác định tâm cụm K-mean++ được đề xuất.

Thuật toán này đảm bảo khởi tạo centroid thông minh hơn và cải thiện chất lượng của phân cụm. K-mean ++ là thuật toán K-mean tiêu chuẩn kết hợp với việc khởi tạo centroid thông minh hơn.

Các bước thực hiện là:

B1: Chọn ngẫu nhiên tâm đầu tiên từ các điểm dữ liệu.

B2: Đối với mỗi điểm dữ liệu, tính toán khoảng cách của nó từ trung tâm gần nhất, đã chọn trước đó.

B3: 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.

B4: 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)

Tư tưởng chính của phương pháp phân cụm phân hoạch (như k-means) là định nghĩa 1 cụm sao cho tổng biến thiên bình phương khoảng cách trong cụm là nhỏ nhất, tham số này là WSS (Within-cluster Sum of Square). Elbow method chọn số sụm k sao cho khi thêm vào một cụm khác thì không làm cho WSS thay đổi nhiều..[10][11]

SSE được tính như sau:

$$\text{SSE} = \sum_{i=1}^{k} \sum_{j=1}^{n_j} \text{Distance}^2(x_{ij}, m_i)$$

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-mean) 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

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 Silhouette, S, cho mỗi mẫu được tính theo công thức sau:

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

a: Khoảng cách trung bình giữa một mẫu và tất cả các điểm khác trong cùng một nhóm.

b: Khoảng cách trung bình giữa một mẫu và tất cả các điểm khác trong cụm gần nhất tiếp theo.

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.

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.

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
- Đánh giá ngoài (external evaluation)

- 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.

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:

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

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

$\mathbf{d}(\mathbf{c_i}, \mathbf{c_j})$**:** 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.

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

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.

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

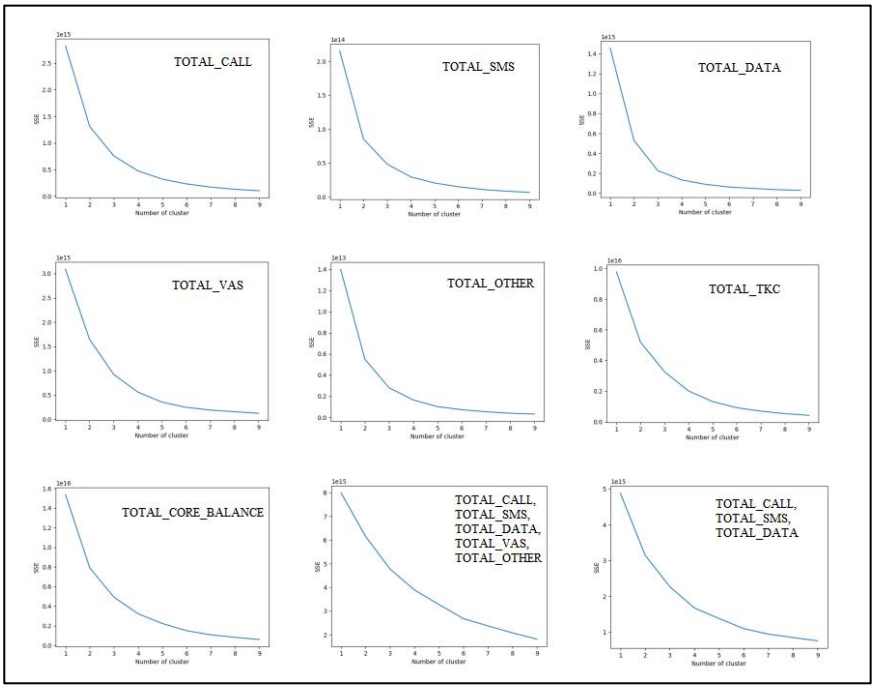
Số trường dữ liệu: **14**

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

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',	709,820	**3**

	'TOTAL_DATA', 'TOTAL_VAS', 'TOTAL_OTHER'		
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

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

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 tính: VNĐ)

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

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)
TKC_SMS	1,370 +/-40,96	82,328 +/-38,833	318,333 +/-146,213

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)
TKC_DATA	2,186 +/-6,873	255,210 +/-97,347	78,312 +/-26,979

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)
TKC_CALL	13,502 +/-19,165	527,403 +/- 255,232	43,151 +/-44,015	138,455 +/-61,024
TKC_SMS	999 +/-5,472	29,186 +/-86,251	4,438 +/-16,817	12,635 +/-43,003
TKC_DATA	3,757 +/-10,649	36,150 +/-67,776	99,466 +/-61,396	4,864 +/-13,237

Số lượng cụm: 6

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 (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)
TKC_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
TKC_SMS	1,040 +/-5693	15,511 +/- 32,203	6,832 +/- 21,316	3,264 +/- 8,698	236,618 +/- 133,043	5,919 +/- 14533

TKC_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

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 ở 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 nhưng 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, chăm sóc khách hàng để khách hàng cảm thấy hài lòng nhất với chi phí mà họ đã bỏ ra.

Chương 4. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

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*", Departement 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.