MỤC LỤC

DANH SÁCH HÌNH VẼ

LỜI CẢM ƠN

Trước hết em xin gửi lời cảm ơn đến Ths. Nguyễn Trịnh Đông, người thầy đã hướng dẫn em rất nhiều trong suốt quá trình tìm hiểu nghiên cứu và hoàn thành đồ án tốt nghiệp từ lý thuyết đến ứng dụng. Sự hướng dẫn của thầy đã giúp em có thêm được những hiểu biết khai phá dữ liệu và ứng dụng của nó trong phân tích cuộc gọi điện thoại.

Đồng thời em cũng xin chân thành cảm ơn các thầy cô trong bộ môn cũng như các thầy cô trong trường đã trang bị cho em những kiến thức cơ bản cần thiết để em có thể hoàn thành tốt đồ án.

Em xin gửi lời cảm ơn đến gia đình, bạn bè đã tạo mọi điều kiện thuận lợi để em có thể xây dựng thành công đồ án này.

Hải Phòng, Ngày 10 tháng 7 năm 2010

Sinh viên thực hiện

Nguyễn Thu Hà

DANH MỤC TỪ VIẾT TẮT

Ký hiệu viết tắt	Giải thích
CDR	Call Detail Records
CSDL	Cơ sở dữ liệu
KDD	Khai phá tri thức trong cơ sở dữ liệu
KPDL	Khai phá dữ liệu

LỜI MỞ ĐẦU

Cuộc cách mạng của kỹ thuật số cho phép số hóa thông tin dễ dàng và chi phí lưu trữ thấp.Với sự phát triển của phần mềm, phần cứng và trang bị nhanh hệ thống máy tính trong kinh doanh. Số lượng dữ liệu khổng lồ được tập trung và lưu trữ trong cơ sở dữ liệu trên các thiết bị điện tử như: đĩa cứng, băng từ, đĩa quang, CD-ROM,… Tốc độ tăng dữ liệu quá lớn [4].

Dữ liệu sau khi phục vụ cho một mục đích nào đó được lưu lại trong kho dữ liệu và theo ngày tháng khối lượng dữ liệu được lưu trữ ngày càng lớn. Trong khối lượng dữ liệu to lớn này có rất nhiều thông tin có ích mang tính tổng quát, thông tin có tính quy luật vẫn còn đang tiềm ẩn mà chúng ta chưa biết. Từ khối lượng dữ liệu rất lớn cần có những công cụ tự động rút các thông tin và kiến thức có ích. Một hướng tiếp cận có khả năng giúp các công ty khai thác các thông tin có nhiều ý nghĩa từ các tập dữ liệu lớn đó là khai phá dữ liệu (Data Mining).

Viễn thông là một ngành đã có những bước phát triển ngoạn mục, trong những năm gần đây. Số lượng các thuê bao và các dịch vụ viễn thông kèm theo đang tăng một cách chóng mặt. Các công nghệ mới cũng phát triển một cách mạnh mẽ. Đây là ngành có tỷ lệ tin học hóa cao, hầu hết các giao dịch, thao tác hoạt động đều được lưu lại trong cơ sở dữ liệu. Từ đó lượng dữ liệu thu thập và lưu trữ được về các hoạt động sản xuất kinh doanh cũng trở nên ngày càng khổng lồ. Tiềm ẩn bên trong lượng dữ liệu này là những tri thức hết sức quý báu về thị trường, khách hàng, sản phẩm…

Đối với ngành viễn thông, thị phần và khách hàng là hai yếu tố hết sức quan trọng, quyết định sự thành công của doanh nghiệp. Chính vì vậy việc nắm được các nhu cầu sở thích của khách hàng cũng như những xu hướng biến động của thị trường là một lợi thế to lớn cho các doanh nghiệp cạnh tranh và mở rộng thị trường của mình. Khai phá dữ liệu chính là một trong những kỹ thuật hữu ích nhất để giải quyết những vấn đề này.

Ngày nay, các công ty viễn thông không ngừng nâng cao, cải tiến các dịch vụ của mình và tìm kiếm dich vụ mới để đáp ứng nhu cầu ngày càng lớn của khách hàng.

Các công ty viễn thông có một nguồn dữ liệu rất quý giá là các bản ghi chi tiết cuộc gọi (Call Detail Records - CDR). Hàng ngày hàng triệu cuộc gọi được ghi nhận tại các tổng đài với mục đich trước tiên là để tính cước cho khách hàng và quản lý mạng. Nguồn dữ liệu này chứa đựng thông tin của khách hàng, cách mà khách hàng sử dụng mạng, các sản phẩm và dịch vụ viễn thông. CDR không chỉ cho biết khi nào một dịch vụ được sử dụng mà còn cho biết dịch vụ đó sử dụng như thế nào. Với các thông tin đó sẽ giúp cho các công ty viễn thông lập kế hoạch phát triển dịch vụ chăm sóc khách hàng để khách hàng yên tâm với dịch vụ, gắn bó lâu dài với công ty. Đồng thời thu hút được nhiều khách hàng mới. Tạo điều kiện phát triển và mở rộng thị trường... Đó là lý do vì sao nhiều công ty viễn thông đã tiến hành xử lý lấy các thông tin này phục vụ cho việc kinh doanh của mình [2].

Vấn đề đặt ra: Làm thế nào có thể trích rút được thông tin có ích từ kho dữ liệu là các bản ghi chi tiết cuộc gọi điện thoại? Trong đồ án tốt nghiệp này em trình bày ứng dụng khai phá dữ liệu trong phân tích dữ liệu cuộc gọi điện thoại. Từ đó tìm ra quy luật sử dụng dịch vụ của khách hàng. Làm cơ sở để hỗ trợ ra quyết định cho các công ty viễn thông.

CHƯƠNG 1: TỔNG QUAN VỀ KHAI PHÁ DỮ LIỆU

1.1 Định nghĩa khai phá dữ liệu

Khai phá dữ liệu (Data Mining) là quá trình tìm kiếm các mẫu mới, những thông tin tiềm ẩn mang tính dự đoán trong các khối dữ liệu lớn cho các đơn vị, tổ chức, doanh nghiệp,… Từ đó làm thúc đẩy khả năng sản xuất, kinh doanh, cạnh tranh cho các đơn vị, tổ chức này. Các tri thức mà khai thác dữ liệu mang lại giúp cho các công ty kinh doanh ra các quyết định kịp thời và có thể trả lời những câu hỏi trong lĩnh vực kinh doanh mà trước đây tốn nhiều thời gian để xử lý. Sự phân tích một cách tự động và mang tính dự báo của các dữ liệu có ưu thế hơn hẳn so với phân tích thông thường dựa trên sự kiện trong quá khứ của các hệ hỗ trợ quyết định trước đây.

Giáo sư Tom Mitchell đã đưa ra định nghĩa của khai phá dữ liệu (KPDL) như sau: “KPDL là việc sử dụng dữ liệu lịch sử để khám phá những qui tắc và cải thiện những quyết định trong tương lai” [10]. Với một cách tiếp cận ứng dụng hơn, Tiến sĩ Fayyad đã phát biểu: “KPDL thường được xem là việc khám phá tri thức trong các cơ sở dữ liệu, là một quá trình trích xuất những thông tin ẩn, trước đây chưa biết và có khả năng hữu ích, dưới dạng các qui luật, ràng buộc, qui tắc trong cơ sở dữ liệu.” [8] Nói tóm lại, KPDL là một quá trình học tri thức mới từ những dữ liệu đã thu thập được.

Khai phá dữ liệu là sự kết hợp của nhiều ngành như: Cơ sở dữ liệu, hiển thị dữ liệu, máy học, trí tuệ nhân tạo, lý thuyết thông tin, xác suất thống kê, tính toán hiệu năng cao, và các phương pháp tính toán mềm,… Khai phá dữ liệu được định nghĩa là quá trình tìm kiếm thông tin (tri thức) có ích, tiềm ẩn và mang tính dự đoán trong các khối CSDL lớn. Một số nhà khoa học xem khai phá dữ liệu như là một cách gọi khác của một thuật ngữ rất thông dụng là khám phá tri thức trong CSDL (Knowlwdge Discovery in Data bases - KDD), vì cho rằng mục đích của quá trình khám phá tri thức là thông tin là tri thức có ích, những đối tượng mà chúng ta phải xử lý rất nhiều trong suốt quá trình khám phá tri thức lại chính là dữ liệu. Một số nhà khoa học khác thì xem khai thác dữ liệu như một bước chính trong quá trình khám phá tri thức.

1.2 Quá trình khai phá tri thức trong cơ sở dữ liệu

Khám phá tri thức trong CSDL (Knowledge Discovery in Databases - KDD) là lĩnh vực liên quan đến các ngành như: thống kê, học máy, CSDL, thuật toán, trực quan hóa dữ liệu, tính toán song song và hiệu năng cao,…

Quá trình KDD có thể phân thành các giai đoạn sau [5][9]:

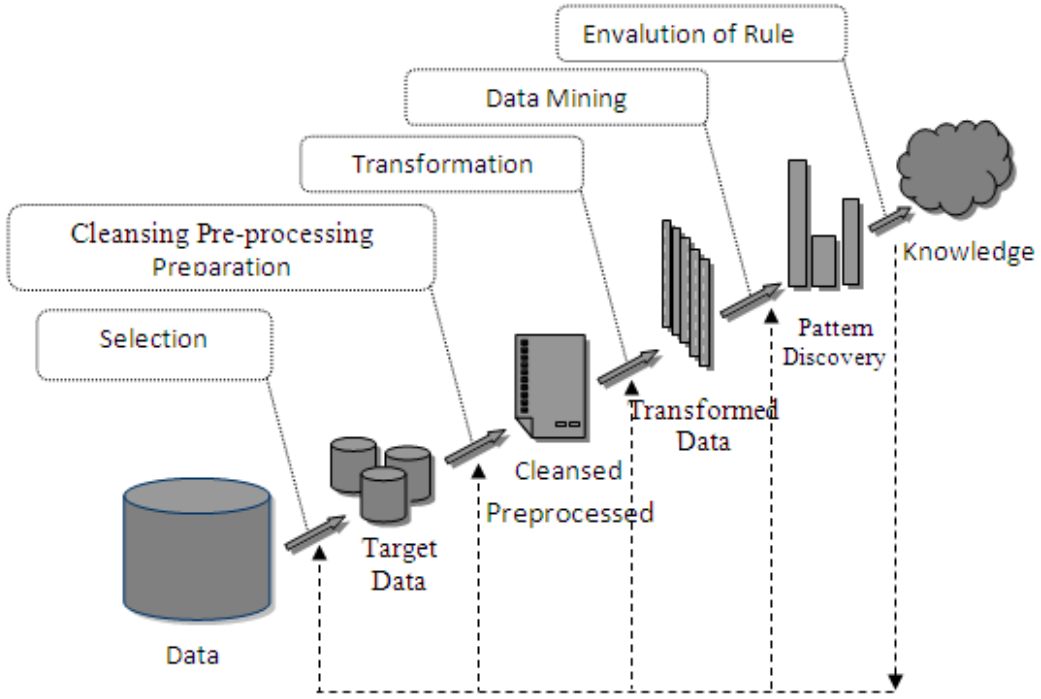
Trích chọn dữ liệu (Data selection): Là bước trích chọn những tập dữ liệu cần được khai phá từ các tập dữ liệu lớn (databases, data warehouses, data repositories) ban đầu theo một số tiêu chí nhất định.

Tiền xử lý dữ liệu (Data preprocessing): Là bước làm sạch dữ liệu (xử lý với dữ liệu không đầy đủ, dữ liệu nhiễu, dữ liệu không nhất quán,.v.v.), rút gọn dữ liệu (sử dụng hàm nhóm và tính tổng, các phương pháp nén dữ liệu, sử dụng histograms, lấy mẫu,.v.v.), rời rạc hóa dữ liệu (rời rạc hóa dựa vào histograms, dựa vào entropy, dựa vào phân khoảng,.v.v.). Sau bước này, dữ liệu sẽ nhất quán, đầy đủ, được rút gọn, và được rời rạc hóa.

Biến đổi dữ liệu (Data transformation): Là bước chuẩn hóa và làm mịn dữ liệu để đưa dữ liệu về dạng thuận lợi nhất nhằm phục vụ cho các kỹ thuật khai phá ở bước sau.

Khai phá dữ liệu (Data mining): Là bước áp dụng những kỹ thuật phân tích (phần nhiều là các kỹ thuật của học máy) nhằm để khai thác dữ liệu, trích chọn được những mẫu thông tin, những mối liên hệ đặc biệt trong dữ liệu. Đây được xem là bước quan trọng và tốn nhiều thời gian nhất của toàn quá trình KDD.

Đánh giá và biểu diễn tri thức (Knowlwdge representation and evaluation): Dùng các kỹ thuật hiển thị dữ liệu để trình bày những mẫu thông tin (tri thức) và mối liên hệ trong dữ liệu đã được khám phá ở bước trên được chuyển dạng và biểu diễn ở một dạng gần gũi với người sử dụng như đồ thị, cây, bảng biểu, luật... Đồng thời bước này cũng đánh giá những tri thức khám phá được theo những tiêu chí nhất định.



Hình 1: Các giai đoạn khai phá tri thức trong cơ sở dữ liệu

1.3 Các kỹ thuật tiếp cận trong khai phá dữ liệu

Nếu đứng trên quan điểm của học máy (Machine Learning), thì các kỹ thuật trong Data Mining, bao gồm [5][9]:

Học có giám sát (Supervised learning): Là quá trình gán nhãn lớp cho các phần tử trong CSDL dựa trên một tập các ví dụ huấn luyện và các thông tin về nhãn lớp đã biết.

Học không có giám sát (Unsupervised learning): Là quá trình phân chia một tập dữ liệu thành các lớp hay là cụm (clustering) dữ liệu tương tự nhau mà chưa biết trước các thông tin về lớp hay tập các ví dụ huấn luyện.

Học nửa giám sát (Semi - Supervised learning): Là quá trình phân chia một tập dữ liệu thành các lớp dựa trên một tập nhỏ các ví dụ huấn luyện và một số các thông tin về một số nhãn lớp đã biết trước.

Nếu căn cứ vào lớp các bài toán cần giải quyết, thì Data Mining bao gồm các kỹ thuật sau [5][9]:

Phân lớp và dự đoán (Classification & prediction): xếp đối tượng vào một trong các lớp đã biết trước. Ví dụ: phân lớp loại cước hoặc loại dịch vụ dựa trên số máy bị gọi của cuộc gọi, phân lớp khu vực dựa trên số máy chủ gọi, phân lớp giờ cao điểm, thấp điểm dựa trên giờ bắt đầu đàm thoại… Phân lớp là một lĩnh vực rất quan trọng trong khai thác dữ liệu. Phân lớp còn được gọi là học có giám sát, hướng tiếp cận này thường được sử dụng một số kỹ thuật của học máy như cây quyết định (decision tree), mạng nơ ron nhân tạo (neural network)…

Luật kết hợp (Association rules): Là dạng luật biểu diễn tri thức ở dạng tương đối đơn giản. Ví dụ: "70% khách hàng gọi liên tỉnh thì có 99% trong số khách hàng đó gọi nội tỉnh". Luật kết hợp có khả năng ứng dụng trong rất nhiều lĩnh vực.

Khai thác mẫu tuần tự (Sequential/temporal patterns): Tương tự như khai thác luật kết hợp nhưng có theo tính thứ tự và tính thời gian. Một luật mô tả mẫu tuần tự có dạng biểu diễn X→Y phản ánh sự xuất hiện của biến cố X sẽ dẫn đến việc xuất hiện kế tiếp biến cố Y. Hướng tiếp cận này có tính dự báo cao.

Phân cụm (Clustering/segmentation): Sắp xếp các đối tượng theo từng cụm. Các đối tượng được gom cụm sao cho mức độ tương tự giữa các đối tượng trong cùng một cụm là lớn nhất và mức độ tương tự giữa các đối tượng nằm trong các cụm khác nhau là nhỏ nhất. Phân cụm còn được gọi là học không giám sát (unsupervised learning).

1.4 Ứng dụng của khai phá dữ liệu

Khai phá dữ liệu có nhiều ứng dụng trong thực tế. Một trong số ứng dụng điển hình như:

Tài chính và thị trường chứng khoán: phân tích tình hình tài chính và dự báo giá của các loại cổ phiếu trong thị trường chứng khoán. Danh mục vốn và giá, lãi suất, dữ liệu thẻ tín dụng, phát hiện gian lận…

Phân tích dữ liệu và hỗ trợ ra quyết định.

Điều trị và chăm sóc y tế: Một số thông tin về chuẩn đoán lưu bệnh trong các hệ thống quản lý bệnh viện. Phân tích mối liên hệ giữa triệu chứng bệnh, chuẩn đoán và phương pháp điều trị (chế độ dinh dưỡng, thuốc..).

Text mining & Web mining: Phân lớp văn bản và các trang web, tóm tắt văn bản…

Lĩnh vực khoa học: Quan sát thiên văn, dữ liệu gene, dữ liệu sinh vật học, tìm kiếm, so sánh các hệ gene và thông tin di truyền, mối liên hệ gene và một số bệnh di truyền.

Mạng viễn thông: Phân tích các cuộc gọi điện thoại và hệ thống giám sát lỗi, phát hiện gian lận, các ứng dụng quản lý và chăm sóc khách hàng, phát hiện sự cố để đưa ra biện pháp phát triển chất lượng dịch vụ…

1.5 Cấu trúc của Call Detail Records (CDR)

Ngành viễn thông lưu trữ một khối dữ liệu khổng lồ bản ghi chi tiết cuộc gọi (Call Detail Records). Những thông tin này có thể cho ta nhận diện được những đặc tính của khách hàng và thông qua đó có thể đưa ra các chính sách chăm sóc khách hàng thích hợp dựa trên dự đoán hoặc có một chiến lược tiếp thị hiệu quả**.**

1.5.1 Giới thiệu CDR

Hàng ngày tại các tổng đài điện thoại, có một số lượng rất lớn các cuộc gọi điện thoại được ghi nhận đó bản ghi chi tiết cuộc gọi và thường được viết tắt là CDR [1]. Các thông số liên quan tới cuộc gọi được ghi lại tại các tổng đài có thể cho chúng ta biết chất lượng của dịch vụ, cách sử dụng dịch vụ của khách hàng. CDR là một khối dữ liệu lớn và rất quan trọng.

Khi một khách hàng nhấc máy quay số thì tổng đài sẽ thiết lập một đường nối giữa hai số điện thoại. Cuộc gọi được bắt đầu khi việc kết nối được thực hiện xong và kết thúc khi một trong hai khách hàng kết thúc cuộc gọi [12].

Sau khi một cuộc gọi điện thoại kết thúc thì các số liệu liên quan tới chi tiết cuộc gọi đó như: số điện thoại gọi, số điện thoại bị gọi, thời gian bắt đầu gọi, thời gian

kết thúc cuộc gọi,… Được lưu xuống bộ nhớ của tổng đài. Chi tiết các cuộc gọi của khách hàng được tổng đài lưu lại dưới dạng tập tin theo cấu trúc quy định trước. Chúng được gọi là CDR.

1.5.2 Cấu trúc của CDR

CDR có hàng triệu bản tin, mỗi bản tin có 39 thuộc tính [6].

ID	Field_name	Type	Width
1	REC_TYPE	Character	2
2	CAUSE_OUTP	Character	1
3	REC_NUMBER	Character	2
4	CALL_ID_NO	Character	8
5	REC_SEQ_NO	Character	8
6	A_SUBS	Character	18
7	A_CATEGORY	Character	2
8	TYPE_A_SUB	Character	2
9	B_SUBS	Character	26
10	B_CATEGORY	Character	2
11	FAULT_CODE	Character	5
12	CALL_STATS	Character	1
13	ABNORM_RLS	Character	1
14	DATE	Character	6
15	START_TIME	Character	6
16	STOP_TIME	Character	6
17	TIME_REGIS	Character	6
18	DURATION	Character	6
19	INTER_TIME	Character	6
20	CHARG_PART	Character	1

ID	Field_name	Type	Width
21	TYPE_SIGN	Character	1
22	EXCHANG_ID	Character	3
23	OUT_ROUTE	Character	7
24	INC_ROUTE	Character	7
25	REROUTE	Character	1
26	DEST_CODE	Character	1
27	FORCE_DISC	Character	1
28	TYPE_A_NO	Character	3
29	TYPE_B_NO	Character	3
30	REDIRECT	Character	18
31	ORI_CALLED	Character	18
32	TAR_SWITCH	Character	1
33	CAUSE_CODE	Character	3
34	LOCATION	Character	2
35	CALLED_SUB	Character	1
36	TELEC_SERV	Character	3
37	NO_MESSAGE	Character	1
38	SEIZ_EOS	Character	8
39	NETWORK_NO	Character	1

Hình 2: Cấu trúc các thuộc tính của CDR

Trong đó một số thuộc tính liên quan tới thông số kỹ thuật của cuộc gọi như:

- **Call_stats:** Cuộc gọi thành công hay không thành công.
- **Redirect:** Cuộc gọi đi hoặc đến theo hướng nào.
- **Fault_code:** Mã lỗi cuộc gọi bao gồm các thông số báo lỗi trùng, chập chờn…
- **Telec_serv:** Các loại dịch vụ được ghi nhận gồm có gọi tự động IDD, điện thoại IP 177,178,177…

Một số thuộc tính để xử lý tính cước cho khách hàng:

- **A_subs:** Số điện thoại của khách hàng gọi đi.
- **B_subs:** Số điện thoại khách hàng gọi đến.
- **A_category:** Phân loại khách hàng gọi đi.
- **B_category:** Phân loại khách hàng gọi đến.
- **Type_a_subs:** Loại dich vụ của khách hàng gọi đến.
- **Date:** Ngày thực hiện giao tác các cuộc gọi điện thoại. định dạng thuộc tính date.
- **Start_time:** Thời điểm lúc bắt đầu thực hiện giao tác (chính xác đến từng giây)
- **Stop_time:** Thời điểm lúc kết thúc thực hiện giao tác (chính xác đến từng giây).
- **Inter_time:** Độ dài cuộc gọi được định dạng là [hhmmss] với h,m,s lần lượt là giờ, phút, giây, (chính xác đến từng giây)
- **Duation:** Độ dài cuộc gọi được làm tròn theo phút.

CHƯƠNG 2: LÝ THUYẾT THỐNG KÊ VÀ MỘT SỐ THUẬT TOÁN ỨNG DỤNG TRONG KHAI PHÁ DỮ LIỆU

2.1 Lý thuyết thống kê

2.1.1 Tổng quan về thống kê

Thống kê là một nhánh của toán học liên quan đến việc thu thập, phân tích, diễn giải hay giải thích và trình bày các dữ liệu. Thống kê được vận dụng trong nhiều lĩnh vực khoa học xã hội và nhân văn. Thống kê cũng được sử dụng để ra quyết định trong tất cả mọi lĩnh vực kinh doanh và quản trị nhà nước [3].

Thống kê là hệ thống các phương pháp dùng để thu thập xử lý và phân tích các con số (mặt lượng) của những hiện tượng số lớn để tìm hiểu bản chất và tính quy luật vốn có của chúng (mặt chất) trong điều kiện thời gian và không gian cụ thể.

Mọi sự vật hiện tượng đều có hai mặt chất và lượng không thể tách rời nhau và khi chúng ta nghiêm cứu hiện tượng, điều chúng ta muốn biết đó là bản chất của hiện tượng. Nhưng mặt chất đều ẩn bên trong còn mặt lượng biểu hiện ra bên ngoài dưới dạng các đại lượng ngẫu nhiên. Do đó phải thông qua các phương pháp xử lý thích hợp trên mặt lượng của số lớn đơn vị cấu thành hiện tượng, tác động của các yếu tố ngẫu nhiên mới được bù trừ và triệt tiêu, bản chất của hiện tượng mới bộc lộ ra và ta có thể nhận thức đúng dắn bản chất, quy luật vận động của nó.

2.1.2 Chức năng của thống kê

Thống kê mô tả: là phương pháp sử dụng để tóm tắt hoặc mô tả một tập hợp dữ liệu.

Thống kê suy diễn: là phương pháp mô hình hóa trên các dữ liệu quan sát để giải thích được những biến thiên "dường như " có tính ngẫu nhiên và không chắc chắn của các quan sát và dùng để rút ra các suy diễn về quá trình hay về tập hợp các đơn vị được nghiêm cứu.

Thống kê mô tả và thống kê suy diễn tạo thành thống kê trong ứng dụng. Còn thống kê toán là lĩnh vực nghiêm cứu cơ sở lý thuyết của khoa học thống kê.

2.1.3 Các khái niệm căn bản

2.1.3.1 Tổng thể và đơn vị đo tổng thể

Tổng thể thống kê (còn gọi là tổng thể chung) là tập hợp các đơn vị (hay phần tử) thuộc hiện tượng nghiêm cứu, cần quan sát, thu thập và phân tích mặt lượng của chúng theo một hoặc một số tiêu thức nào đó.

Các đơn vị (hay phần tử) cấu thành tổng thể thống kê gọi là đơn vị tổng thể.

Ví dụ: muốn tìm độ dài trung bình của các cuộc gọi điện thoại trong khoảng 21giờ – 22 giờ ngày 14/2/1010 tại tổng đài của VNPT thì tổng thể sẽ là toàn bộ các cuộc gọi điện thoại trong khoảng 21 giờ – 22 giờ ngày 8/3/2010 tại tổng đài của VNPT.

Vậy thực chất của việc xác định tổng thể là xác định các đơn vị tổng thể. Đơn vị tổng thể là xuất phát điểm của quá trình nghiêm cứu thống kê vì nó chứa đựng những thông tin ban đầu cần cho quá trình nghiêm cứu [3].

2.1.3.2 Mẫu và đơn vị mẫu

Mẫu là tổng thể bao gồm một số đơn vị được chọn ra từ tổng thể chung theo một phương pháp lấy mẫu nào đó. Các đặc trưng mẫu được sử dụng để suy rộng ra các đặc trưng của tổng thể chung [3].

Quan sát là cơ sở thu thập số liệu và thông tin cần nghiêm cứu. Chẳng hạn trong điều tra chọn mẫu, mỗi đơn vị mẫu sẽ được tiến hành ghi chép, thu thập thông tin được gọi là một quan sát.

2.1.3.3 Dữ liệu định tính và dữ liệu định lượng

Dữ liệu định tính phản ánh tính chất, sự hơn kém của đối tượng của các đối tượng nghiêm cứu, là các dữ liệu ban đầu không được thể hiện dưới dạng số.

Dữ liệu định lượng phản ánh mức độ hơn kém, là các dữ liệu có thể cân, đo, đong, đếm được.. Ví dụ độ dài cuộc gọi điện thoại có thể đếm chính xác tới từng giây.

Dữ liệu định tính dễ thu thập hơn dữ liệu định lượng, nhưng dữ liệu định lượng thường cung cấp nhiều thông tin và dễ áp dụng nhiều phương pháp phân tích hơn. Khi

thực hiện nghiêm cứu, trong giai đoạn lập kế hoạch nghiêm cứu và thu thập dữ liệu, người nghiêm cứu cần xác định được các phương pháp phân tích cần sử dụng để phục vụ cho mục tiêu của mình, từ đó xác định loại dữ liệu cần thu thập để thu nhận được dữ liệu mong muốn.

2.1.3.4 Tiêu thức thống kê

Tiêu thức thống kê là khái niệm dùng để chỉ các đặc điểm của đơn vị tổng thể. Ví dụ khi phân tích chi tiết cuộc gọi điện thoại có các tiêu thức như: số điện thoại gọi, số điện thoại bị gọi, ngày thực hiện cuộc gọi điện thoại, thời gian bắt đầu thực hiện cuộc, thời gian đàm thoại, thời gian kết thúc cuộc gọi….

Tiêu thức thống kê được chia thành 2 loại [3]:

- Tiêu thức thuộc tính: là tiêu thức phản ánh tính chất hay loại hình của đơn vị tổng thể, không có biểu hiện trực tiếp bằng các con số. Vi dụ: tiêu thức loại khách hàng, loại dịch vụ cuộc gọi, lỗi cuộc gọi… là các tiêu thức thuộc tính.
- Tiêu thức số lượng: là tiêu thức có thể biểu hiện trực tiếp bằng con số. Ví dụ: số điện thoại khách hàng gọi đi, số điện thoại khách hành gọi đến, thời gian bắt đầu, thời gian đàm thoại, thời gian kết thúc…

2.1.4 Cấp bậc đo lường và các thang đo dữ liệu

2.1.4.1 Thang đo định danh

Là loại thang đo dùng cho các tiêu thức thuộc tính. Người ta sử dụng các mã số để phân loại các đối tượng, chúng không mang ý nghĩa nào khác.
Thước đo độ tập trung duy nhất là mode, độ phân tán thống kê có thể đo bằng các tỷ lệ, không tính được độ lệch chuẩn.

2.1.4.2 Thang đo thứ bậc

Là loại thang đo dùng cho các tiêu thức thuộc tính và các tiêu thức số lượng. Trong thang đo này, giữa các biểu hiện của tiêu thức có liên quan thứ bậc hơn kém. Sự chênh lệch giữa các biểu hiện không nhất thiết phải bằng nhau. Thước đo độ tập trung là mode hay trung vị, trung vị cung cấp nhiêu thông tin hơn mode.

2.1.4.3 Thang đo khoảng

Là loại thanh đo dùng cho các tiêu thức số lượng và các thang đo thuộc tính. Thang đo khoảng là thang đo thứ bậc có các khoảng cách đều nhau. Khuynh hướng trung tâm của dữ liệu thu thập từ thang đo khoảng có thể là mode, trung vị và trung bình cộng. Trong đó trung bình cộng chứa nhiêu thang đo nhất.

2.1.4.4 Thang đo tỷ lệ

Là loại thanh đo dùng cho dữ liệu số lượng. Thang đo tỷ lệ có đầy đủ các đặc tính của thang đo khoảng, tức là có thể áp dụng các phép tính cộng trừ. Ngoài ra, thang đo này có một giá trị 0 "thật", cho phép lấy tỷ lệ so sánh giữa hai giá tri thu thập cho nên gọi là thang đo tỷ lệ. Đây là thang đo cao nhất trong các loại thang đo. Khuynh hướng trung tâm của dữ liệu thu thập là mode, trung vị và trung bình cộng, trong đó trung bình cộng chứa nhiều thông tin nhất.

2.2 Một số thuật toán trong khai phá dữ liệu

Thống kê là hệ thống các phương pháp dùng để thu thập xử lý và phân tích các con số để tìm hiểu bản chất và tính quy luật vốn có của chúng. Một trong các phương pháp dể xử lý, phân tích, khai phá dữ liệu đó là sử dụng thuật toán.

Ta tìm hiểu một số thuật toán khai phá dữ liệu.

2.2.1 Thuật toán phân hoạch K-MEANS

Thuật toán phân hoạch K-Means do MacQeen đề xuất trong lĩnh vực thống kê năm 1967.

Tư tưởng của thuật toán K-Means là sinh ra k cụm dữ liệu $\{C_1, C_2, \ldots, C_k\}$ từ một tập dữ liệu chứa n đối tượng trong không gian d chiều $X_i = (x_{i1}, x_{i2}, \ldots, x_{id})$ $(i = \overline{1,n})$, 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 tâm của cụm C_i, D là khoảng cách giữa hai đối tượng.

Tâm của một cụm là một véc tơ, trong đó giá trị của mỗi phần tử của nó là trung bình cộng của các thành phần tương ứng của các đối tượng vectơ dữ liệu trong cụm

đang xét. Độ đo khoảng cách D giữa các đối tượng dữ liệu thường được sử dụng dụng là khoảng cách Euclide, bởi vì đây là mô hình khoảng cách dễ để lấy đạo hàm và xác định các cực trị tối thiểu. Hàm tiêu chuẩn và độ đo khoảng cách có thể được xác định cụ thể hơn tuỳ vào ứng dụng hoặc các quan điểm của người dùng.

Các bước tiến hành thuật toán K-Means [5][9]:

Input: Tập dữ liệu chứa n đối tượng, số cụm k.

Output: Tâm các cụm C_i ($i = \overline{1,k}$) và hàm tiêu chuẩn E đạt giá trị tối thiểu.

Thuật toán K-Means bao gồm các bước cơ bản sau:

Bước 1: Chọn k tâm $\{m_j\}_{j=1}^{k}$ ban đầu trong không gian R^d (d là số chiều của dữ liệu). Việc lựa chọn này có thể là ngẫu nhiên hoặc theo kinh nghiệm.

Bước 2: Đối với mỗi điểm X_i (1<=i<=n), tính toán khoảng cách của nó tới mỗi tâm m_j j=1,k. Sau đó tìm tâm gần nhất đối với mỗi điểm.

Bước 3: Đối với mỗi j=1,k, cập nhật tâm cụm m_j bằng cách xác định trung bình cộng của các vectơ đối tượng dữ liệu.

Bước 4: Lặp các bước 2 và 3 đến khi các tâm của cụm không thay đổi.

Thuật toán K-Means tuần tự trên có độ phức tạp tính toán là: $O((3nkd)\tau T^{flop})$ Trong đó: n là số đối tượng dữ liệu, k là số cụm dữ liệu, d là số chiều, τ là số vòng lặp, T^{flop} là thời gian để thực hiện một phép tính cơ sở như phép tính nhân, chia, … Như vậy, do K-Means phân tích phân cụm đơn giản nên có thể áp dụng đối với tập dữ liệu lớn. Tuy nhiên, nhược điểm của K-means là chỉ áp dụng với dữ liệu có thuộc tính số và khám ra các cụm có dạng hình cầu, K-means còn rất nhạy cảm với nhiễu và các phần tử ngoại lai trong dữ liệu.

Chất lượng phân cụm dữ liệu của thuật toán K-means phụ thuộc nhiều vào các tham số đầu vào như: số cụm k và k tâm khởi tạo ban đầu. Trong trường hợp, các tâm khởi tạo ban đầu mà quá lệch so với các tâm cụm tự nhiên thì kết quả phân cụm của K-Means là rất thấp, nghĩa là các cụm dữ liệu được khám phá rất lệch so với các cụm trong thực tế. Trên thực tế người ta chưa có một giải pháp tối ưu nào để chọn các tham

số đầu vào, giải pháp thường được sử dụng nhất là thử nghiệm với các giá trị đầu vào k khác nhau rồi sau đó chọn giải pháp tốt nhất.

2.2.2 Thuật toán PAM

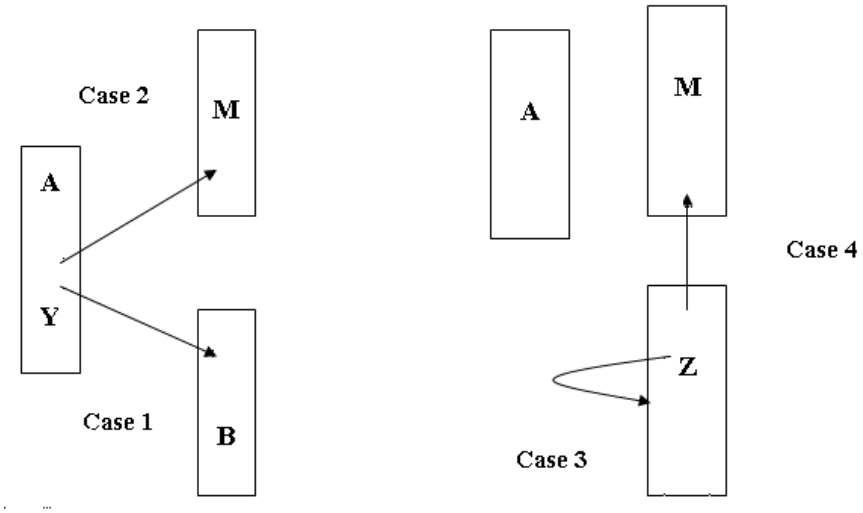
Thuật toán PAM được đề xuất bởi Kaufman và Rousseeuw. PAM (Partitioning Around Medoids) là thuật toán mở rộng của thuật toán K-means, nhằm có khả năng xử lý hiệu quả đối với dữ liệu nhiễu hoặc các phần tử ngoại lai.

Tư tưởng: Thay vì sử dụng các tâm như K-Means, PAM sử dụng các đối tượng medoid để biểu diễn cho các cụm dữ liệu, một đối tượng medoid là đối tượng đặt tại vị trí trung tâm nhất bên trong của mỗi cụm. Vì vậy, các đối tượng medoid ít bị ảnh hưởng của các đối tượng ở rất xa trung tâm, trong khi đó các tâm của thuật toán K-means lại bị tác động bởi các điểm xa trung tâm này. Ban đầu, PAM khởi tạo k đối tượng medoid và phân phối các đối tượng còn lại vào các cụm với các đối tượng medoid đại diện tương ứng sao cho chúng tương tự với đối tượng medoid trong cụm nhất [5][9][10].

Thí dụ: Nếu O_j là đối tượng không phải là medoid và O_m là một đối tượng medoid, khi đó ta nói O_j thuộc về cụm có đối tượng medoid là O_m làm đại diện nếu: $d(O_j, O_m) = \min_{Oe} d(O_j, O_e)$. Trong đó: $d(O_j, O_e)$ là độ phi tương tự giữa O_j và O_e, $\min_{Oe}$ là giá trị nhỏ nhất của độ phi tương tự giữa O_j và tất cả các đối tượng medoid của các cụm dữ liệu. Chất lượng của mỗi cụm được khám phá được đánh giá thông qua độ phi tương tự trung bình giữa một đối tượng và đối tượng medoid tương ứng với cụm của nó, nghĩa là chất lượng phân cụm được đánh giá thông qua chất lượng của tất cả các đối tượng medoid. Độ phi tương tự ở đây thông thường được xác định bằng độ đo khoảng cách, thuật toán PAM thường được áp dụng cho dữ liệu không gian.

Để xác định các medoid, PAM bắt đầu bằng cách lựa chọn k đối tượng medoid bất kỳ. Sau mỗi bước thực hiện, PAM cố gắng hoán chuyển giữa đối tượng medoid O_m và một đối tượng O_p không phải là medoid, miễn là sự hoán chuyển này nhằm cải tiến chất lượng của phân cụm, quá trình này kết thúc khi chất lượng phân cụm không thay đổi. Chất lượng phân cụm được đánh giá thông qua hàm tiêu chuẩn, chất lượng phân cụm tốt nhất khi hàm tiêu chuẩn đạt giá trị tối thiểu.

Xét ví dụ: Cho hai đối tượng medoid A và B. Đối với tất cả các đối tượng Y thuộc cụm với đối tượng medoid đại diện A, chúng ta tìm medoid của cụm gần nhất để thay thế. Có hai trường hợp có thể xẩy ra, hoặc Y được chuyển tới cụm dữ liệu có đại diện là B hoặc được chuyển tới cụm dữ liệu có đại diện là M. Tiếp đến, chúng ta xét lần lượt cho tất cả các đối tượng trong cụm có đại diện là A. Tương tự như vậy, đối với tất các các đối tượng trong cụm có đối tượng đại diện là B, chúng ta có thể di chuyển chúng tới cụm có đại diện là M hoặc là chúng ở lại B. Thí dụ này có thể biểu diễn như hình dưới đây:



Hình 3: Biểu diễn ví dụ cho huật toán PAM

Một số biến được sử dụng trong thuật toán PAM:

- O_m: Là đối tượng medoid hiện thời cần được thay thế.
- O_p: Là đối tượng medoid mới thay thế cho O_m.
- O_j: Là đối tượng dữ liệu (không phải là medoid) có thể được di chuyển sang cụm khác.
- $O_{j,2}$: Là đối tượng medoid hiện thời gần đối tượng O_j nhất mà không phải là các đối tượng A và M như trong ví dụ trên.

PAM tính giá trị C_{jmp} cho tất cả các đối tượng O_j. C_{jmp} ở đây nhằm để làm căn cứ cho việc hoán chuyển giữa O_m và O_p. Trong mỗi trường hợp C_{jmp} được tính với 4 cách khác nhau như sau:

Trường hợp 1: Giả sử O_j hiện thời thuộc về cụm có đại diện là O_m và O_j tương tự với $O_{j,2}$ hơn O_p ($d(O_j, O_p) \geq d(O_j, O_{j,2})$). Trong khi đó, $O_{j,2}$ là đối tượng medoid tương tự xếp thứ 2 tới O_j trong số các medoid. Trong trường hợp này, chúng ta thay thế O_m bởi đối tượng medoid mới O_p và O_j sẽ thuộc về cụm có đối tượng đại diện là $O_{j,2}$. Vì vậy, giá trị hoán chuyển C_{jmp} được xác định như sau:

$C_{jmp} = d(O_j, O_{j,2}) – d(O_j, O_m)$.

Giá trị C_{jmp} là không âm.

Trường hợp 2: O_j hiện thời thuộc về cụm có đại diện là O_m, nhưng O_j ít tương tự với $O_{j,2}$ so với O_p (Nghĩa là, $d(O_j, O_p) < d(O_j, O_{j,2})$). Nếu O_m, được thay thế bởi O_p thì O_j sẽ thuộc về cụm có đại diện là O_p. Vì vậy, giá trị C_{jmp} được xác định như sau:

$C_{jmp} = (O_j, O_p) - d(O_j, O_m)$.

C_{jmp} ở đây có thể là âm hoặc dương.

Trường hợp 3: Giả sử O_j hiện thời không thuộc về cụm có đối tượng đại diện là O_m mà thuộc về cụm có đại diện là $O_{j,2}$. Mặt khác, giả sử O_j tương tự với $O_{j,2}$ hơn so với O_p, khi đó, nếu O_m được thay thế bởi O_p thì O_j vẫn sẽ ở lại trong cụm có đại diện là $O_{j,2}$. Do đó:

$C_{jmp} = 0$.

Trường hợp 4: O_j hiện thời thuộc về cụm có đại diện là $O_{j,2}$ nhưng O_j ít tương tự tới $O_{j,2}$ hơn so với O_p. Vì vậy, nếu chúng ta thay thế O_m bởi O_p thì O_j sẽ chuyển từ cụm $O_{j,2}$ sang cụm O_p. Do đó, giá trị hoán chuyển C_{jmp} được xác định là:

$C_{jmp} = (O_j, O_p) - d(O_j, O_{j,2})$.

C_{jmp} ở đây luôn âm.

Kết hợp cả bốn trường hợp trên, tổng giá trị hoán chuyển O_m bằng O_p được xác định như sau: $TC_{mp} = \sum_j C_{jmp}$.

Input: Tập dữ liệu có n phần tử, số cụm k.

Output: k cụm dữ liệu sao cho chất lượng phân hoạch là tốt nhất.

Sử dụng các khái niệm trên, thuật toán PAM có các bước thực hiện sau [5][9][10]:

Bước 1: Chọn k đối tượng medoid bất kỳ.

Bước 2: Tính TC_{mp} cho tất cả các cặp đối tượng O_m, O_p. Trong đó O_m là đối tượng medoid và O_p là đối tượng không phải là modoid.

Bước 3: Chọn cặp đối tượng O_m và O_p. Tính $min_{Om,}$ min_{Op}, TC_{mp}.

Nếu TC_{mp} là âm, thay thế O_m bởi O_p và quay lại bước 2. Nếu TC_{mp} dương, chuyển sang bước 4.

Bước 4: Với mỗi đối tượng không phải là medoid, xác định đối tượng medoid tương tự với nó nhất đồng thời gán nhãn cụm cho chúng.

Độ phức tạp tính toán của PAM là $O(Ik\ (n-k)^2)$, trong đó I là số vòng lặp. Như vậy, thuật toán PAM kém hiệu quả về thời gian tính toán khi giá trị của k và n là lớn.

2.2.3 Thuật toán CLARA

CLARA (Clustering LARge Application) được Kaufman đề xuất năm 1990 [5], thuật toán này nhằm khắc phục nhược điểm của thuật toán PAM trong trường hợp giá trị của k và n là lớn.

Tư tưởng: CLARA tiến hành trích mẫu cho tập dữ liệu có n phần tử, nó áp dụng thuật toán PAM cho mẫu này và tìm ra các các đối tượng tâm medoid cho mẫu được trích từ dữ liệu này [5][10].

Người ta thấy rằng, nếu mẫu dữ liệu được trích theo cách ngẫu nhiên, thì các medoid của nó xấp xỉ với các medoid của toàn bộ tập dữ liệu ban đầu. Để tiến tới một xấp xỉ tốt hơn, CLARA đưa ra nhiều cách lấy mẫu và thực hiện phân cụm cho mỗi trường hợp và tiến hành chọn kết quả phân cụm tốt nhất khi thực hiện phân cụm trên các mẫu này. Để cho chính xác, chất lượng của các cụm được đánh giá thông qua độ phi tương tự trung bình của toàn bộ các đối tượng dữ liệu trong tập đối tượng ban đầu. Kết quả thực nghiệm chỉ ra rằng, 5 mẫu dữ liệu có kích thước 40 + 2k cho các kết quả tốt.

Input: Tập dữ liệu n phần tử, các mẫu của tập dữ liệu..

Output: k cụm dữ liệu sao cho chất lượng phân hạch tốt nhất.

Các bước thực hiện của thuật toán CLARA [5][10]:

Bước 1: Lấy 5 mẫu dữ liệu có kích thước 40 + 2k.

Bước 2: Lấy một mẫu có 40 + 2k đối tượng dữ liệu ngẫu nhiên từ tập dữ liệu và áp dụng thuật toán PAM cho mẫu dữ liệu này nhằm để tìm các đối tượng medoid đại diện cho các cụm.

Bước 3: Đối với mỗi đối tượng O_j trong tập dữ liệu ban đầu, xác định đối tượng medoid tương tự nhất trong số k đối tượng medoid.

Bước 4: Tính độ phi tương tự trung bình cho phân hoạch các đối tượng ở bước trước, nếu giá trị này bé hơn giá trị tối thiểu hiện thời thì sử dụng giá trị này thay cho giá trị tối thiểu ở trạng thái trước, như vậy, tập k đối tượng medoid xác định ở bước này là tốt nhất cho đến thời điểm này.

Bước 5: Quay trở về bước 2 rồi tiếp tục thực hiện tới khi duyệt hết 5 mẫu dữ liệu có kích thước 40 + 2k.

Độ phức tạp tính toán của nó là $O(k(40+k)^2 + k(n-k))$, và CLARA có thể thực hiện đối với tập dữ liệu lớn.

Chú ý đối với kỹ thuật tạo mẫu trong PCDL: kết quả phân cụm có thể không phụ thuộc vào tập dữ liệu khởi tạo nhưng nó chỉ đạt tối ưu cục bộ. Thí dụ: Nếu các đối tượng medoid của dữ liệu khởi tạo không nằm trong mẫu, khi đó kết quả thu được không đảm bảo là tốt nhất được.

2.2.4 Thuật toán CLARAS

Thuật toán CLARANS được Ng & Han đề xuất năm 1994 [5], nhằm để cải tiến cho chất lượng cũng như mở rộng áp dụng cho tập dữ liệu lớn. CLARANS cũng sử dụng các đối tượng trung tâm medoids làm đại diện cho các cụm dữ liệu.

Như đã biết, PAM là thuật toán phân hoạch có kiểu K-medoid. Nó bắt đầu khởi tạo k tâm đại diện medoid và liên tục thay thế mỗi tâm bởi một đối tượng khác trong

cụm cho đến khi là tổng khoảng cách của các đối tượng đến tâm cụm không giảm. CLARAS là thuật toán PCDL kết hợp thuật toán PAM với chiến lược tìm kiếm kinh nghiệm mới.

Tư tưởng: CLARAS không xem xét tất cả các khả năng có thể thay thể các đối tượng tâm medoids bởi một đối tượng khác, nó ngay lập tức thay thế các đối tượng tâm này nếu việc thay thế này có tác động tốt đến chất lượng phân cụm chứ không cần xác định cách thay thể tối ưu nhất. Một phân hoạch cụm phát hiện được sau khi thay thế đối tượng trung tâm được gọi là một láng giềng (Neighbor) của phân hoạch cụm trước đó. Số các láng giềng được hạn chế bởi tham số do người dùng đưa vào là Maxneighbor, quá trình lựa chọn các láng giềng này là hoàn toàn ngẫu nhiên. Tham số Numlocal cho phép người dùng xác định số vòng lặp tối ưu cục bộ được tìm kiếm. Không phải tất các các láng giềng được duyệt mà chỉ có Maxneighbor số láng giềng được duyệt [5][9].

Giả sử O là một tập có n đối tượng và $M \subseteq O$ là tập các đối tượng tâm medoid, NM = O - M là tập các đối tượng không phải tâm. Các đối tượng dữ liệu sử dụng trong thuật toán CLARANS là các khối đa diện. Mối đối tượng được diễn tả bằng một tập các cạch, mỗi cạnh được xác định bằng 2 điểm. Giả sử $P \subseteq R^3$ là một tập tất cả các điểm. Nói chung, các đối tượng ở đây là các đối tượng dữ liệu không gian và chúng ta định nghĩa tâm của một đối tượng chính là trung bình cộng toán học của tất cả các đỉnh hay còn gọi là trọng tâm:

$$\text{Center} : O \longrightarrow P$$

Giả sử *dist* là một hàm khoảng cách, khoảng cách thường được chọn ở đây là khoảng cách Euclidean : *dist:* $P \times P \rightarrow R_0^+$

Hàm khoảng cách *dist* có thể mở rộng cho các điểm của khối đa diện thông qua hàm tâm: dist: $O \times O \rightarrow R_0^+$ sao cho dist (o_i, o_j) = dist (center(o_i), center(o_j))

Mỗi đối tượng được được gán cho một tâm medoid của cụm nếu khoảng cách từ trọng tâm của đối tượng đó tới tâm medoid của nó là nhỏ nhất. Vì vậy, chúng ta

định nghĩa một tâm medoid như sau: medoid: $O \rightarrow M$ sao cho medoid (o) = m_i, $m_i \in M$, $\forall m_j \in M$: dist (o, m_i) $\leq$ dist (o, m_j), $o \in O$.

Cuối cùng, chúng ta định nghĩa một cụm với tâm medoid m_i tương ứng là một tập con các đối tượng trong O với medoid(o) = m_i.

Giả sử C_0 là tập tất cả các phân hoạch của O. Hàm tổng để đánh giá chất lượng một phân hoạch được định nghĩa như sau: total_distance: $C_0 \rightarrow R_0^+$ sao cho total_distance(c) = $\Sigma \Sigma$ dist (o, mi) với $m_i \in M$, $o \in$ cluster(mi).

Input: O, k, dist, numlocal, maxneighbor.

Output: k cụm dữ liệu.

Các bước thực hiện:

Bước 1: Chọn ngẫu nhiên k đối tượng medoil từ n đối tượng dữ liệu.

Bước 2: Thay thế một đối tượng tâm cụm medoil cũ bởi đối tượng khác trong khi số phân hoạch nhỏ hơn maxneighbor.

Bước 3: Tính toán sự khác nhau về tổng khoảng cách giữa phân hoạch hiện thời (Neighbor) và phân hoạch cụm trước đó.

Bước 4: Hoán đổi giữa đối tượng tâm cụm medoil với đối tượng không phải medoil nếu không có sự khác nhau giữa phân hoạch hiện thời (Neighbor) và phân hoạch cụm trước đó.

Bước 5: Tính tổng để đánh giá chất lượng của mỗi phân hoạch. Nếu tổng đó nhỏ hơn chất lượng phân hoạch cho trước thì cập nhật chất lượng phân hoạch này.

Bước 6: Lặp lại bước 1 numlocal lần.

Quá trình hoạt động của CLARANS tương tự với quá trình hoạt động của thuật toán CLARA. Tuy nhiên, ở giai đoạn lựa chọn các trung tâm medoid của cụm dữ liệu, CLARANS lựa chọn một giải pháp tốt hơn bằng cách lấy ngẫu nhiên một đối tượng của k đối tượng trung tâm medoid của cụm và cố gắng thay thế nó với một đối tượng được chọn ngẫu nhiên trong (n-k) đối tượng còn lại, nếu không có giải pháp nào tốt

hơn sau một số cố gắng lựa chọn ngẫu nhiên xác định, thuật toán dừng và cho kết quả phân cụm tối ưu cục bộ.

Trong trường hợp tệ nhất, CLARANS so sánh một đối tượng với tất các đối tượng Medoid. Vì vậy, độ phức tạp tính toán của CLARANS là $O(kn^2)$, do vậy CLARANS không thích hợp với tập dữ liệu lớn (khi trường hợp xấu nhất xẩy ra). CLARANS có ưu điểm là không gian tìm kiếm không bị giới hạn như đối với CLARA, và trong cùng một lượng thời gian thì chất lượng của các cụm phân được là lớn hơn so với CLARA [5].

2.2.5 Thuật toán K - PROTOTYPE

Thuật toán K-Prototypes là thuật toán mở rộng của thuật toán K-Means để làm việc với tập dữ liệu hỗn hợp giữa thuộc tính số và thuộc tính hạng mục. Thuật toán K-Prototypes sử dụng các đối tượng mẫu (prototype) để biểu diễn cho các cụm thay vì sử dụng các đối tượng tâm như trong thuật toán K-Means. Các đối tượng dữ liệu lần lượt được phân phối cho các cụm dữ liệu sao cho chúng tương tự nhất với đối tượng mẫu tương ứng với cụm dữ liệu mà chúng được phân phối.

Tư tưởng: Ban đầu, chọn k đối tượng mẫu theo ngẫu nhiên hoặc theo kinh nghiệm, giai đoạn tiếp theo chúng ta phân phối lần lượt từng đối tượng dữ liệu cho các cụm ứng với đối tượng mẫu mà chúng tương tự nhất, sau mỗi lần phân phối đối tượng dữ liệu cho các cụm, chúng ta cập nhật giá trị cho các đối tượng mẫu. Sau khi tất các các đối tượng đã được phân về cho các cụm dữ liệu, chúng ta lần lượt kiểm tra lại từng đối tượng dữ liệu cho các cụm, nếu đối tượng dữ liệu nào phân phối chưa phù hợp thì ta tiến hành di chuyển đối tượng đó sang cụm thích hợp và tiến hành cập nhật lại các đối tượng mẫu đại diện cho hai cụm này. Quá trình kiểm tra này được lặp cho đến khi chúng ta chuyển đến trạng thái tất cả các đối tượng đã được phân về đúng cụm của mình. Các đối tượng mẫu có mô hình giống như mô hình của các đối tượng dữ liệu, nghĩa là chúng được biểu diễn bằng vectơ và được xác định như sau: Mỗi giá trị của các thuộc tính số được tính bằng trung bình cộng của các giá trị các thuộc tính số tương ứng của các đối tượng trong cụm. Trong khi đó, mỗi giá trị của các thuộc tính

hạng mục được tính bằng tần suất giá trị lớn nhất của giá trị thuộc tính hạng mục tương ứng của các đối tượng trong cụm.

Thí dụ về xác định các giá trị thuộc tính cho đối tượng mẫu như sau:

Xét một cụm dữ liệu có các đối tượng dữ liệu là bản ghi chi tiết cuộc gọi có các giá trị thuộc tính lần lượt là: độ dài cuộc gọi, thời gian gọi, loại dịch vụ cuộc gọi:

X_1 = (5, 8, 171); X_2 = (3, 10, 178); X_3 = (6, 13, 171); X_4 = (7, 14, 171).

Lúc này, đối tượng mẫu được xác định như sau: Prototypes = (21/4, 45/4, 171).

Thuật toán K-Prototypes là thuật toán phân cụm phân hoạch sử dụng hàm tiêu chuẩn E và cách thức biểu diễn cụm bằng đối tượng mẫu.

Input: Tập dữ liệu ban đầu X và số cụm k.

Output: k đối tượng mẫu sao cho hàm tiêu chuẩn đạt giá trị tối thiểu.

Các bước thực hiện [5][9]:

Bước 1: Khởi tạo k đối tượng mẫu ban đầu cho X, mỗi đối tượng mẫu đóng vai trò là tâm đại diện của mỗi cụm.

Bước 2: Phân phối mỗi đối tượng trong X cho mỗi cụm sao cho chúng gần nhất với đối tượng mẫu trong cụm, đồng thời cập nhật lại đối tượng mẫu cho mỗi cụm.

Bước 3: Sau khi tất cả các đối tượng đã được phân phối hết cho các cụm, kiểm tra lại độ tương tự của các đối tượng trong mỗi cụm với các đối tượng mẫu, nếu có một đối tượng mẫu tương tự nhất với nó mà khác với đối tượng mẫu của cụm hiện thời thì di chuyển đối tượng đang xét này sang cụm tương ứng với đối tượng mẫu mà nó gần nhất và đồng thời cập nhật các đối tượng mẫu cho hai cụm này.

Bước 4: Lặp bước 3 cho đến khi không có đối tượng nào thay đổi sau khi đã kiểm tra toàn bộ các đối tượng.

Thuật toán K-Prototypes là thuật toán dựa trên lược đồ của thuật toán K-Means nhằm áp dụng cho tập dữ liệu lớn có kiểu hỗn hợp. Vì vậy, K-Prototypes rất có ý nghĩa trong ngữ cảnh hầu hết các hệ quản trị CSDL hiện nay đều chứa dữ liệu có kiểu hỗn hợp. Giống như K-Means, các nhược điểm của K-Prototypes là rất nhạy cảm với các

giá trị khởi tạo của các prototypes và không tìm ra các cụm với hình dạng bất kỳ. Ngoài ra, trong một số trường hợp, K-Prototypes khá nhạy cảm với nhiễu và phần tử ngoại lai trong dữ liệu, để khắc phục nhược điểm này ta có thể cải tiến hàm tính độ tương tự của cho thuật toán hoặc là cải tiến cách cập nhật lại đối tượng mẫu cho thuật toán.

CHƯƠNG 3: CHƯƠNG TRÌNH THỬ NGHIỆM VÀ ĐÁNH GIÁ

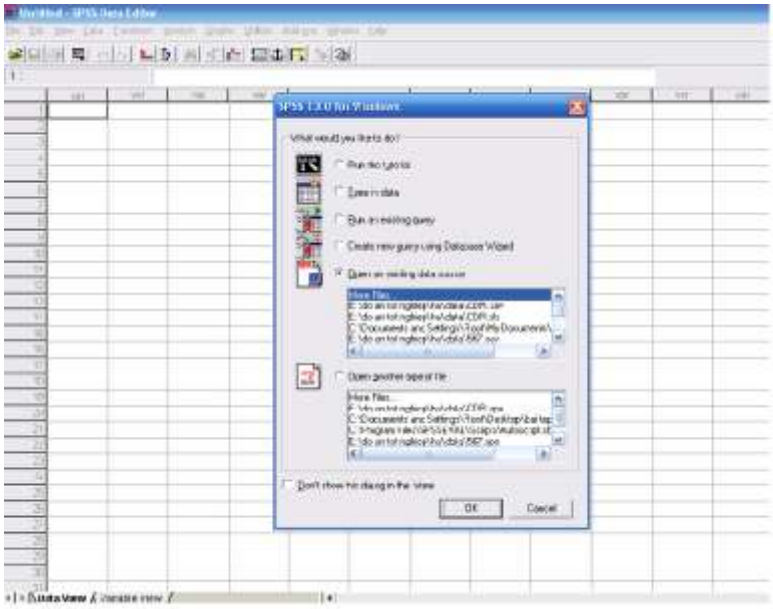
Hiện nay có rất nhiều thuật toán ứng dụng khai phá, phân tích dữ liệu. Cũng có rất nhiều phần mềm hỗ trợ cho việc phân tích dữ liệu. Vì thời gian nghiêm cứu có hạn nên em chưa thể xây dựng được phần mềm mới khai phá dữ liệu. Trong đồ án này em sử dụng phần mềm SPSS để phân tích dữ liệu cuộc gọi điện thoại. Từ đó hỗ trợ cho việc ra quyết định của công ty viễn thông.

3.1 Giới thiệu khái quát về phần mềm SPSS

SPSS là tên viết tắt của cụm từ Staticscal Package for the Social Sciences. Đây là một phần mềm được sử dụng rộng rãi nhất trong nghiêm cứu khoa học tự nhiên và khoa học xã hội nói chung [11].

SPSS là một hệ thống phần mềm thống kê toàn diện được thiết kế để thực hiện tất cả các bước trong phân tích thống kê từ thống kê mô tả (liệt kê dữ liệu, lập đồ thị) đến thống kê suy luận (tương quan, hồi quy…).

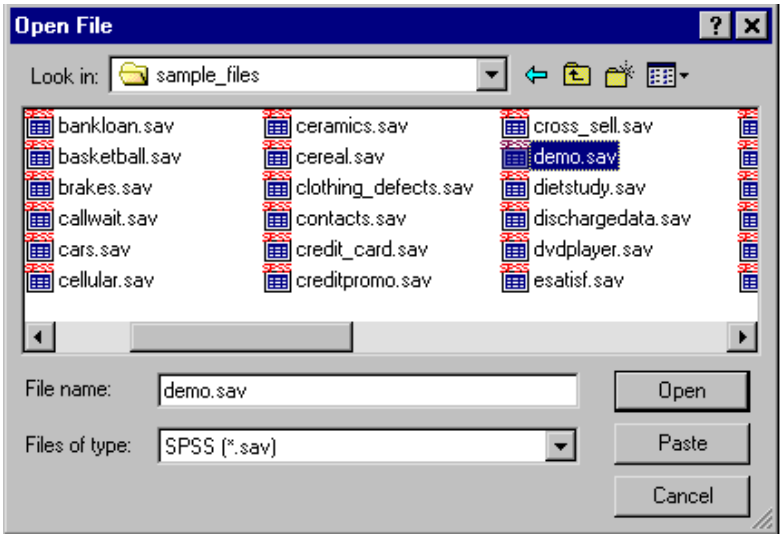
Giao diện của SPSS khi khởi động.



Hình 4: Giao diện của SPSS khi khởi động

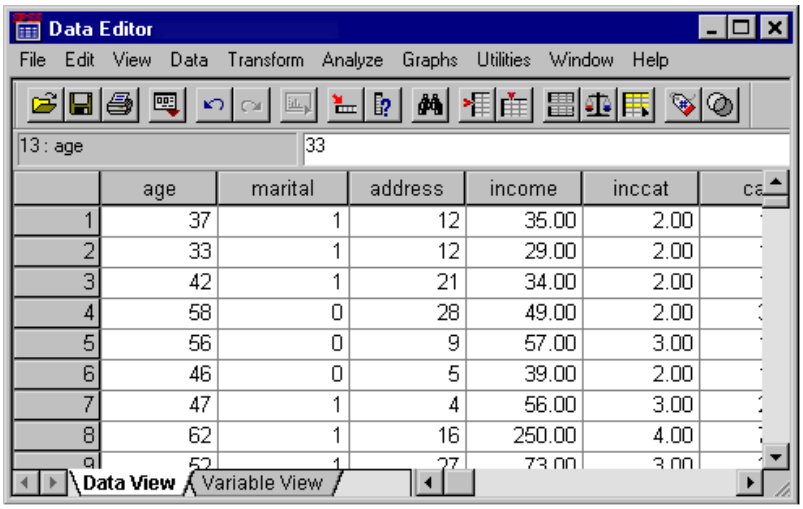
Các tệp tin có thể mở:

- Các bảng tính worksheet được thiết lập trong Execl hoặc Lotus.
- Cơ sở dữ liệu được lập dưới dạng dBASE và SQL.
- Các file dạng text ASCII với kiểu Tab-dliminated.
- Các file trong dạng SPSS được lập trong các hệ điều hành khác.
- Các file dữ liệu SYTAT.



Hình 5: Mở file dữ liệu

Dữ liệu được lưu dưới dạng các bản ghi.



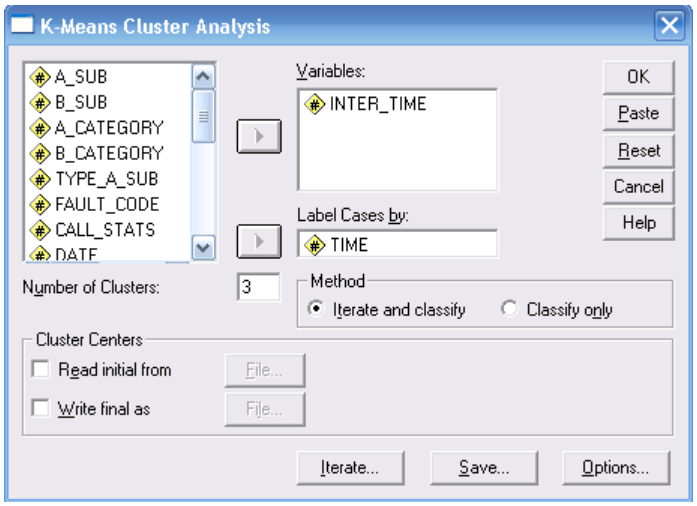
	age	marital	address	income	inccat
1	37	1	12	35.00	2.00
2	33	1	12	29.00	2.00
3	42	1	21	34.00	2.00
4	58	0	28	49.00	2.00
5	56	0	9	57.00	3.00
6	46	0	5	39.00	2.00
7	47	1	4	56.00	3.00
8	62	1	16	250.00	4.00

Hình 6: Dữ liệu trong SPSS

3.2 Kết quả thực nghiệm

Tập CDR bao gồm hơn 10 nghìn bản ghi mỗi bản ghi có 30 thuộc tính, nhưng ta chỉ dùng một vài thuộc tính quan trọng như độ dài cuộc gọi, ngày và giờ gọi… đó là những thông tin rất quan trọng mà chúng ta cần khai phá để đưa ra quy luật.

Ta có thể phân cụm giá trị độ dài cuộc gọi dựa trên thời gian gọi.



Hình 7: Phân cụm K-Means

Thuật toán phân cụm phân hoạch K–Means được đưa vào trong SPSS để phân cụm dữ liệu. Ta phân làm 3 cụm dựa vào giá trị độ dài cuộc gọi có các tâm khởi tạo lần lượt là: 2, 16, 31.

Initial Cluster Centers

	Cluster		
	1	2	3
INTER_TIME	2	16	31

Hình 8: Tâm khởi tạo của cụm

Với 4 bước lặp để thay đổi tâm cụm.

Iteration History[a]

Iteration	Change in Cluster Centers		
	1	2	3
1	2.377	3.184	2.407
2	.000	.240	1.814
3	.000	.378	2.290
4	.000	.000	.000

Hình 9: Quá trình thay đổi tâm cụm

Các tâm cuối cùng của cụm lần lượt là: 4, 12, 24.

Final Cluster Centers

	Cluster		
	1	2	3
INTER_TIME	4	12	24

Hình 10: Tâm cuối cùng của cụm

Mỗi trường hợp tương ứng với một bản ghi trong kho dữ liệu. Sau khi kết thúc thuật toán các bản ghi được đưa về các cụm.

Cluster Membership

Case Number	TIME	Cluster	Distance
1	1	1	.377
2	1	1	3.623
3	1	2	.198
4	1	1	.377
5	1	1	3.623
6	1	1	3.623
7	1	2	2.198
8	1	2	.198
9	1	1	2.377
10	1	1	3.623
11	1	1	3.623
12	1	2	.198
13	1	2	1.802
14	1	2	3.802
15	1	1	2.377
16	1	1	.377
17	2	1	1.623
18	2	2	2.198
19	2	3	4.489
20	2	3	4.489
21	3	1	2.377
22	3	1	.377
23	3	1	1.623
24	4	1	2.377
25	4	1	2.377

Hình 11: Các bản ghi thuộc các cụm

Trong Hình 3.8 có 25 bản ghi được phân về các cụm: bản ghi số 1, 2, 4, 5, 6, 9, 10, 11, 12, 13, 14, 18, 19, 20, 21, 22 được phân về cụm 1. Cụm số 1 có tâm cụm là 4 phân theo độ dài cuộc gọi tương ứng với độ dài cuộc gọi ở mức trung bình. Các bản ghi số 3, 7, 8, 12, 13, 14, 18 được phân về cụm 2. Cụm 2 có tâm cụm là 12 phân theo độ dài cuộc gọi tương ứng với độ dài cuộc gọi điện thoại ở mức độ cao. Các bản ghi số 19, 20 được phân về cụm 3 có số tâm cụm là 24 phân theo độ dài cuộc gọi tương ứng với độ dài cuộc gọi ở mức độ rất cao.

Thống kê số bản ghi trong mỗi cụm.

Number of Cases in each Cluster

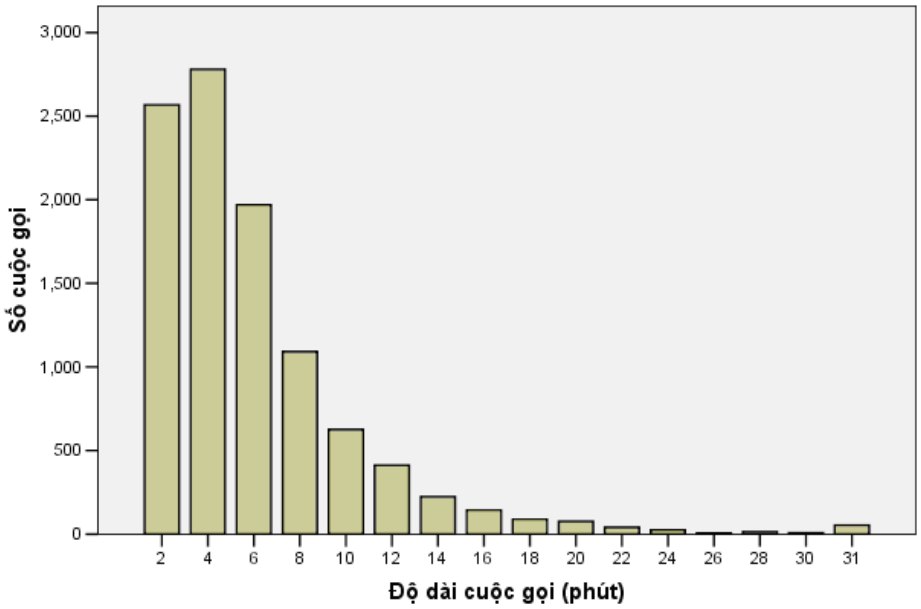
Cluster	1	8409.000
	2	1493.000
	3	225.000
Valid		10127.000
Missing		.000

Hình 12: Số bản ghi thuộc các cụm

Kết quả sau khi phân cụm dữ liệu sẽ cho thấy tại từng thời điểm thói quen sử dụng điện thoại của khách hàng như thế nào. Với việc phân cụm độ dài cuộc gọi điện thoại theo thời gian gọi sẽ thấy được tại khoảng thời gian nào khách hàng gọi điện với độ dài cuộc gọi lớn, khoảng thời gian nào khách hàng gọi điện với độ dài cuộc gọi nhỏ. Kết quả thử nghiệm cho thấy:

- Độ dài cuộc gọi thuộc mức trung bình có 8409 cuộc gọi, chiếm 83,04% tổng số các cuộc gọi, phân bố trong tất cả các giờ trong ngày nhưng tập trung chủ yếu vào khoảng 7 giờ 30 phút đến 10 giờ và khoảng 14 giờ tới 16 giờ 30 phút. Khoảng thời gian đó thuộc giờ hành chính. Các cơ quan, công ty và khách hàng khác có nhu cầu sử dụng điện thoại cao.
- Độ dài cuộc gọi thuộc mức cao có 1493 cuộc gọi, chiếm 14,75% tổng số các cuộc gọi, phân bố đồng đều trong tất cả các giờ trong ngày.
- Độ dài cuộc gọi thuộc mức độ rất cao có 225 cuộc gọi, chiếm 2,21% tổng các cuộc gọi, phân bố chủ yếu vào thời gian ngoài giờ hành chính. Khoảng 21 giờ tới 1 giờ sang ngày hôm sau. Khi đó khách hàng có nhiều thời gian rảnh nên họ có thể gọi điện với khoảng thời gian lớn.

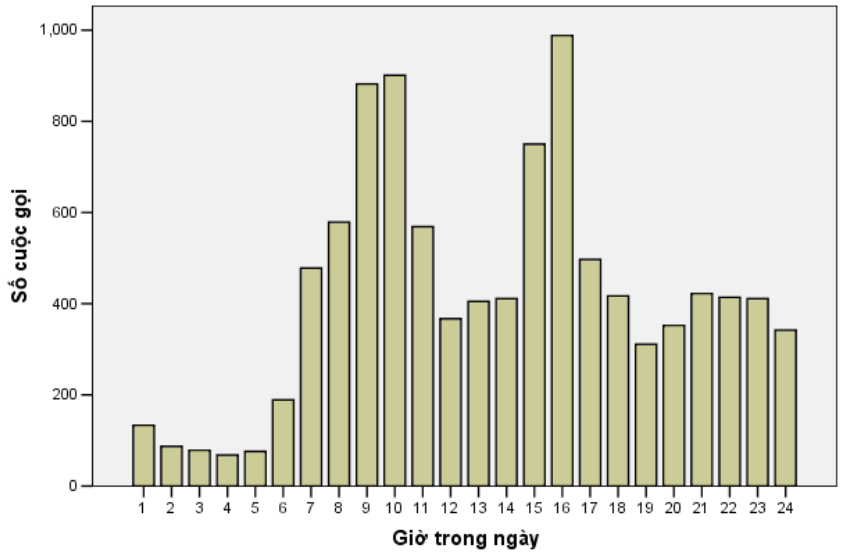
Độ dài cuộc gọi điện thoại là đặc trưng cơ bản thể hiện việc sử dụng điện thoại của khách hàng. Các kết quả thống kê độ dài cuộc gọi điện thoại cho thấy được các đặc điểm của độ dài các cuộc gọi điện thoại.



Hình 13: Thống kê số cuộc gọi theo độ dài cuộc gọi

Với đồ thị trong hình 13 cho chúng ta thấy việc sử dụng điện thoại của khách hàng. Các cuộc gọi điện thoại có độ dài dưới 4 phút rất nhiều sau đó giảm dần trong đoạn từ 6 phút tới 8 phút. Trong khoảng thời gian lớn hơn 10 phút, độ dài cuộc gọi giảm nhanh chóng.

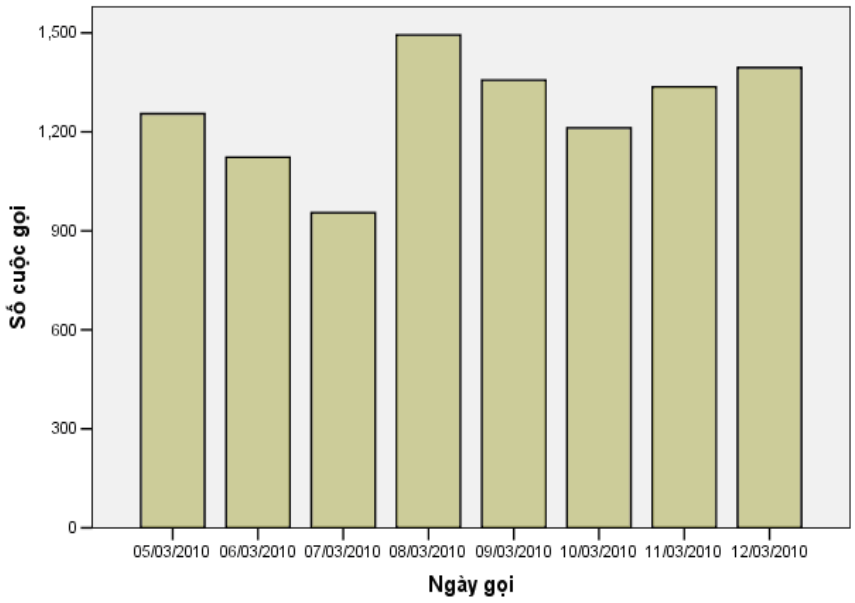
Tổng số các cuộc gọi theo giờ bắt đầu trong ngày thể hiện thói quen sử dụng điện thoại của khách hàng.



Hình 14: Thống kê số cuộc gọi theo giờ trong ngày

Theo đồ thị trong hình 14 thì khách hàng gọi nhiều nhất khoảng 8 giờ tới 10 giờ và khoảng 14 giờ tới 16 giờ. Đây là khoảng thời gian làm việc hành chính nhu cầu sử dụng điện thoại rất lớn tại các văn phòng, cơ quan. Trong khoảng từ 20 giờ tới 22 giờ số cuộc gọi tương đối lớn, đó là khoảng thời gian khách hàng có thể gọi điện nói chuyện hỏi thăm nhau. Trong khoảng 0 giờ tới 5 giờ nhu cầu sử dụng điện thoại rất thấp.

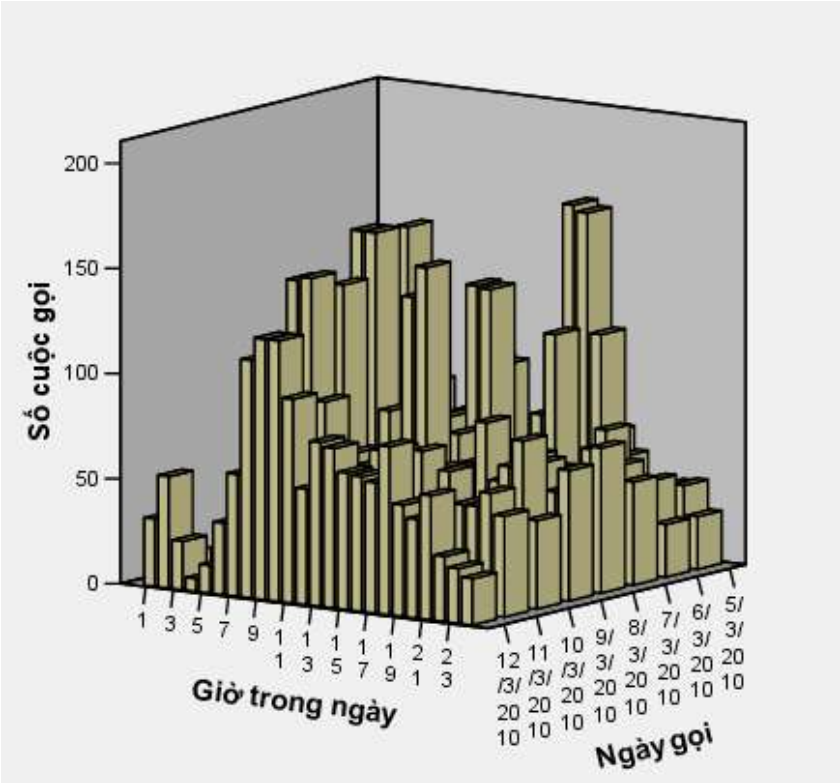
Tổng số các cuộc gọi theo ngày.



Hình 15: Thống kê số cuộc gọi theo ngày

Trong hình 3.4 ta thấy được khách hàng gọi nhiều các ngày trong tuần. Riêng ngày 3/8 nhu cầu sử dụng điện thoại của khách hàng lớn. Ngày 6/3 đó là ngày thứ 7, số các cuộc gọi giảm. Ngày 7/3 thuộc ngày chủ nhật, số các cuộc gọi giảm rất nhiều. Các ngày khác trong tuần số cuộc gọi tương đối đồng đều.

Hình ảnh sử dụng điện thoại của khách hàng.



Hình 16: Hình ảnh sử dụng điện thoại của khách hàng theo ngày gọi và giờ gọi

Với hình trên chúng ta có thể thấy được tổng quan về thời gian khách hàng sử dụng điện thoại.

Thống kê số cuộc gọi của từng khách hàng theo thời gian gọi trong ngày. Khi đó ta có thể xác định được thói quen gọi điện của từng khách hàng để có thể đưa ra nhiều dịch vụ chăm sóc khách hàng ngày càng tốt.

Thống kê số cuộc gọi của khách hàng theo thời gian gọi và độ dài cuộc gọi để tìm ra quy luật sử dụng điện thoại của khách hàng: họ thường gọi điện vào thời gian nào? Thời gian đàm thoại là bao lâu?...

Thống kê số cuộc gọi của mỗi khách hàng tới các thuê bao di động và các thuê bao thuộc các tỉnh khác nhau.

	Thue bao Hai Phong	Thue bao Quang Ninh	Thue bao Hai Duong	Thue bao Thai Binh	Thue bao Thanh Hoa	Thue bao di dong
313117151	1			1		
313118111	2					
313118485	3					
313118585	1					1
313118851	2					
313118885	3					
313121019						1
313121991	2					
313123092	1					
313123099	1		1			
313124568						1
313125025		1				5
313126026	5					
313126200	1	1	3			4
313126210	1	1	3		1	4
313126262	2		3			4
313126815		1	2			3
313126819	1		1		1	1
313129029			2			2
313129200	2		1			
313129210	2		1			
313129226	3		1			
313129292	3		1			
313130700	4		1			4

Hình 17: Số cuộc gọi của mỗi khách hàng tới các thuê bao

Thống kê số cuộc gọi của khách hàng tới các thuê bao khác theo cách sử dụng dich vụ điện thoại IP. Số điện thoại gọi đến thuộc các thuê bao di động, thuê bao thuộc tỉnh Hải Phòng, Quảng Ninh, Hải Dương, Thái Bình và Thanh Hóa. Phần lớn khách hàng sử dụng dịch vụ điện thoại IP 171, 178 để gọi đến các thuê bao thuộc tỉnh Quảng Ninh, Hải Dương, Thái Bình và Thanh Hóa.

	0	171	178
Thue bao Hai Phong	3894		
Thue bao Quang Ninh	11	244	167
Thue bao Hai Duong		362	226
Thue bao Thai Binh	9	310	204
Thue bao Thanh Hoa	16	261	173
Thue bao di dong	3235		

Hình 18: Khách hàng sử dụng dịch vụ điện thoại IP

3.3 Đánh giá kết quả

Sau khi phân tích và thống kê trên tập dữ liệu nhận được thói quen sử dụng điện thoại của khách hàng:

Khách hàng thường gọi điện với độ dài cuộc gọi ở mức trung bình (dưới 4 phút).

Thời gian sử dụng điện thoại chủ yếu của khách hàng là khoảng 8 giờ tới 10 giờ 30 phút và khoảng 14 giờ tới 16 giờ 30 phút.

Ngày 5/7, 9/7, 10/7, 11/7, 12/7 nhu cầu sử dụng điện thoại của khách hàng tương đối đồng đều. Ngày 6/3 thuộc ngày thứ 7 nhu cầu sử dụng điện thoại giảm. Ngày 7/3 thuộc ngày chủ nhật nhu cầu sử dụng điện thoại thấp. Riêng ngày 8/3 nhu cầu sử dụng tăng đột biến.

Với kết quả đạt được sẽ hỗ trợ cho việc ra quyết định của công ty. Công ty nên mở rộng nâng cấp băng thông như thế nào để có thể đáp ứng tốt nhu cầu sử dụng của khách hàng. Với những khoảng thời gian khách hàng có nhu cầu sử dụng dịch vụ lớn như khoảng 8 giờ tới 10 giờ 30 phút và khoảng 14 giờ tới 16 giờ 30 phút công ty cần có biện pháp xử lý thật tốt để đảm bảo chất lượng mạng khi đàm thoại khi đó khách hàng có thể yên tâm với dịch vụ mình đã lựa chọn. trong những ngày lễ lớn nhu cầu sử dụng.

Trong bảng thống kê số cuộc gọi của khách hàng theo ngày gọi, thời gian gọi và độ dài cuộc gọi. nhận được một số khách hàng thường xuyên sử dụng dịch vụ, khách hàng thường xuyên gọi điện với độ dài cuộc gọi lớn và thói quen gọi điện của họ.

Thống kê một số khách hàng thường xuyên gọi điện tại các thời điểm:

- Số thuê bao 313226623 gọi 3 cuộc vào khoảng 6 giờ, 4 cuộc vào khoảng 13 giờ, 1 cuộc vào khoảng 18 giờ, 1 cuộc vào khoảng 21 giờ, 3 cuộc vào khoảng 23 giờ và 1 cuộc vào khoảng 24 giờ.
- Số thuê bao 313313700 gọi vào hầu hết các giờ trong ngày có 2 cuộc vào khoảng 5 giờ, 2 cuộc vào khoảng 6 giờ, 6 cuộc vào khoảng 8 giờ, 4 cuộc vào khoảng 9 giờ, 2 cuộc vào khoảng 10 giờ, 2 cuộc vào khoảng 11 giờ, 1 cuộc vào

khoảng 12 giờ, 3 cuộc vào khoảng 13 giờ, 2 cuộc vào khoảng 15 giờ, 1 cuộc vào khoảng 16 giờ, 2 cuộc vào khoảng 18 giờ, 2 cuộc vào khoảng 19 giờ, 2 cuộc vào khoảng 20 giờ, 6 cuộc vào khoảng 21 giờ.

- Số thuê bao 3137170 có số lượng cuộc gọi rất lớn. có 2 cuộc gọi vào 2 giờ, 1 cuộc vào 3 giờ, 3 cuộc vào khoảng 6 giờ, 2 cuộc vào khoảng 7 giờ, 14 cuộc vào khoảng 8 giờ, 3 cuộc vào khoảng 9 giờ, 4 cuộc vào khoảng 11 giờ, 4 cuộc vào khoảng 13 giờ, 6 cuộc vào khoảng 15 giờ, 3 cuộc vào khoảng 17 giờ, 5 cuộc vào khoảng 19 giờ, 4 cuộc vào khoảng 20 giờ, 4 cuộc vào khoảng 21 giờ.
- Số thuê bao 313710177 gọi 1 cuộc vào khoảng 1 giờ, có 1 cuộc vào khoảng 4 giờ, 2 cuộc vào khoảng 7 giờ, 3 cuộc vào khoảng 8 giờ, 3 cuộc vào khoảng 15 giờ, 4 cuộc vào khoảng 16 giờ, 1 cuộc vào khoảng 18 giờ, 2 cuộc vào khoảng 19 giờ, 2 cuộc vào khoảng 20 giờ, 3 cuộc vào khoảng 21 giờ, 2 cuộc vào khoảng 22 giờ và 1 cuộc vào khoảng 23 giờ.

Thống kê một số khách hàng thường xuyên gọi điện và có thời gian gọi lớn:

- Số thuê bao 313313741 gọi 4 cuộc có độ dài khoảng 2 phút, có 2 cuộc độ dài khoảng 4 phút, có 3 cuộc độ dài 6 phút, 1 cuộc độ dài 8 phút và 2 cuộc có độ dài lớn hơn 30 phút.
- Số thuê bao 313726210 gọi 10 cuộc có độ dài khoảng 2 phút, 7 cuộc có độ dài khoảng 4 phút, 1 cuộc có độ dài khoảng 6 phút, 1 cuộc có độ dài khoảng 10 phút, 1 cuộc gọi có độ dài khoảng 12 phút và 2 cuộc gọi có độ dài lớn hơn 30 phút.
- Số thuê bao 313073107 gọi 4 cuộc có độ dài khoảng 2 phút, 3 cuộc có độ dài khoảng 4 phút, 5 cuộc có độ dài khoảng 6 phút, 6 cuộc gọi có độ dài khoảng 8 phút, có 1 cuộc có độ dài khoảng 18 phút và 1 cuộc có độ dài khoảng 20 phút.

Với bảng thống kê số cuộc gọi của khách hàng theo ngày gọi, thời gian gọi và độ dài cuộc gọi. Có thể tìm được những khách hành thường xuyên sử dụng dịch vụ và khách hàng thường gọi điện với độ dài cuộc gọi lớn. Từ đó có thể đưa khách hàng vào nhóm khách hàng đặc biệt, dựa vào báo cáo về nhóm khách hàng này bộ phận chăm

sóc khách hàng có thể đưa ra bản kế hoạch chiến lược cung cấp dich vụ và các chương trình khuyễn mãi riêng nhằm đảm bảo sự bền vững kết nối giữa khách hàng và công ty.

Phần lớn khách hàng gọi tới các thuê bao thuộc ngoại tỉnh sử dụng dịch vụ điện thoại IP 171, 187. Kết quả này có thể hỗ trợ cho việc ra quyết định nâng cấp phát triển dịch vụ của công ty BCVT Việt Nam (điện thoại IP 171) và công ty điện tử viễn thông Quân đội (điện thoại IP 178).

KẾT LUẬN

Lĩnh vực khai phá dữ liệu là lĩnh vực còn mới ở Việt nam. Đặc biệt đối với sinh viên chúng em. Với thời gian hạn chế đồ án đã đạt được một số kết quả như: Tìm hiểu tổng quan về khai phá dữ liệu; ứng dụng của khai phá dữ liệu để phát hiện tri thức; cấu trúc của CDR (Call Detail Records) trong tổng đài điện thoại…Trong đồ án này em thực hiện quy trình khai phá dữ liệu trên tập dữ liệu CDR với phần mềm SPSS để phân tích dữ liệu cuộc gọi điện thoại cho các thuê bao VNPT Hải Phòng từ đó nắm bắt được quy luật sử dụng cũng như nhu cầu của khách hàng để doanh nghiệp có thể đảm bảo chất lượng dịch vụ hoặc triển khai thêm các dịch vụ mới.

Do dữ liệu của CDR chưa bao gồm tất cả thông tin khách hàng nên chưa thể đánh giá được tiềm năng khách hàng theo vùng.

Khai phá dữ liệu là một ngành khoa học có ứng dụng rất nhiều trong thực tế, và đem lại nhiều lợi ích.

Hướng nghiên cứu tiếp theo: Tiếp tục tìm thêm nhiều tiêu chí đánh giá; có thể gom nhóm khách hàng theo tổng thời gian sử dụng dịch vụ; áp dụng khai phá dữ liệu rộng rãi trong nhiều ngành khác…

TÀI LIỆU THAM KHẢO

Tài liệu tiếng Việt

[1] Nguyễn Đức Cường, *Tổng quan về khai phá dữ liệu*, Khoa Công Nghệ Thông Tin – Đại học Bách Khoa Thành Phố Hồ Chí Minh

[2] Nguyễn Anh Trung, *Ứng dụng các kỹ thuật khai phá dữ liệu vào lĩnh vực viễn thông*, Trung tâm công nghệ thông tin, Học viện Công Nghệ Bưu Chính Viễn Thông.

[3] Hà Văn Sơn, *Giáo tình lý thuyết thống kê*, Bộ môn lý thuyết thống kê, thống kê kinh tế Đại học Kinh Tế Thành Phố Hồ Chí Minh.

[4] Trương Ngọc Châu, Phan Văn Dũng, Nghiêm cứu tính ứng dụng của khai thác luật kết hợp trong cơ sở dữ liệu giao dich, Trường Đại học Bách Khoa Đà Nẵng

[5] Hoàng Hải Xanh, *Các kỹ thuật phân cụm trong Data Mining,* Luận văn, Đại Học Công Nghệ - Đại học Quốc Gia Hà Nội.

[6] Lê Bá Phương, *Ứng dụng khai khoáng dữ liệu trong phân tích số liệu cuộc gọi điện thoại,* Luận văn thạc sĩ, Đại học Quốc Gia Thành Phố Hồ Chí Minh.

Tài liệu tiếng anh

[7] Alan Rea (1995), *Data Mining – An Introduction*. The Parallel Computer Centre, Nor of The Queen's University of Belfast.

[8] U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth and R. Uthurusamy: *Dsvances in Knowledge Dicovery and Data Mining* (1996).

[9] Jiawei Han and Micheline Kamber (2001), *Data Mining: Concepts and Techniques*, Hacours Science and Technology Company, USA.

[10] T.Mitchell, *Machine Learning and Data Mining*, Communication of the ACM, Vol. 42 (1990).

Trang web:

[11] http://en.wikipedia.org/wiki/Call_Detail_Record

[12]http://forum.mait.vn/ebook-tai-lieu/12060-e-book-huong-dan-su-dung-spss.html