

BỘ GIÁO DỤC VÀ ĐÀO TẠO
ĐẠI HỌC ĐÀ NẴNG

NGUYỄN MINH TÂN

ỨNG DỤNG KHAI PHÁ DỮ LIỆU DỰ ĐOÁN
KHÁCH HÀNG RỜI MẠNG VIỄN THÔNG

Chuyên ngành: KHOA HỌC MÁY TÍNH
Mã số: 60.48.01

TÓM TẮT LUẬN VĂN THẠC SĨ KỸ THUẬT

Đà Nẵng - Năm 2011

Công trình được hoàn thành tại
ĐẠI HỌC ĐÀ NẴNG

Người hướng dẫn khoa học: **PGS.TS. Võ Trung Hùng**

Phản biện 1:

Phản biện 2:

Luận văn sẽ được bảo vệ trước Hội đồng chấm Luận
văn tốt nghiệp thạc sĩ kỹ thuật họp tại Đại học Đà Nẵng
vào ngày tháng năm 2011

Có thể tìm hiểu luận văn tại:

- Trung tâm Thông tin - Học liệu, Đại học Đà Nẵng
- Trung tâm Học liệu, Đại học Đà Nẵng.

MỞ ĐẦU

1. Lý do chọn đề tài

Với sự bùng nổ và phát triển của công nghệ thông tin đã mang lại nhiều hiệu quả đối với khoa học cũng như các hoạt động thực tế, trong đó khai phá dữ liệu là một lĩnh vực mang lại hiệu quả thiết thực cho con người. Khai phá dữ liệu đã giúp người sử dụng thu được những tri thức hữu ích từ những cơ sở dữ liệu hoặc các kho dữ liệu khổng lồ khác.

Cơ sở dữ liệu trong các đơn vị, tổ chức kinh doanh, quản lý khoa học chứa đựng nhiều thông tin tiềm ẩn, phong phú và đa dạng, đòi hỏi phải có những phương pháp nhanh, phù hợp, chính xác, hiệu quả để lấy được những thông tin bổ ích. Những “ tri thức ” chiết suất từ nguồn cơ sở dữ liệu trên sẽ là nguồn thông tin hỗ trợ cho lãnh đạo trong việc lên kế hoạch hoạt động hoặc trong việc ra quyết định sản xuất kinh doanh. Tiến hành công việc như vậy chính là thực hiện quá trình phát hiện tri thức trong cơ sở dữ liệu (Knowledge Discovery in Database) mà trong đó kỹ thuật khai phá dữ liệu (Data Mining) cho phép phát hiện những tri thức tiềm ẩn. Để lấy được thông tin mang tính tri thức trong khối dữ liệu khổng lồ, cần thiết phải phát triển các kỹ thuật có khả năng tích hợp các dữ liệu từ các hệ thống giao dịch khác nhau, chuyển chúng thành một tập hợp các cơ sở dữ liệu ổn định có chất lượng.

Một trong các nội dung cơ bản nhất trong khai phá dữ liệu và rất phổ biến là kỹ thuật gom cụm. Phương pháp này nhằm tìm ra các tập thuộc tính thường xuất hiện đồng thời trong cơ sở dữ liệu và rút ra các luật về ảnh hưởng của một tập thuộc tính dẫn đến sự xuất hiện của một (hoặc một tập) thuộc tính khác như thế nào. Ứng dụng

khai phá dữ liệu đã mang lại những lợi ích to lớn trong việc tổng hợp và cung cấp những thông tin trong các nguồn cơ sở dữ liệu lớn.

EVNTelecom là một nhà cung cấp dịch vụ viễn thông mới trên thị trường Việt Nam. Chính thức cung cấp dịch vụ viễn thông công cộng từ cuối năm 2005, đến cuối năm 2007 đã phát triển được hai triệu khách hàng. Đến tháng 6 năm 2008, số lượng khách hàng phát được ở con số bốn triệu. Tuy nhiên tình hình trở nên xấu đi khi các nhà cung cấp ào ạt khuyến mãi và có nhiều đợt đại hạ giá. Kết quả là có nhiều khách hàng rời mạng chuyển sang sử dụng dịch vụ của nhà cung cấp khác, doanh thu ngày một giảm, khó thu hồi vốn đầu tư. Để phát triển được khách hàng, EVNTelecom phải đầu tư gần ba triệu đồng bao gồm chi phí phát triển khách hàng, đầu tư hệ thống và thiết bị đầu cuối. Trong khi đó ARPU là 80.000 đồng và vòng đời trung bình của khách hàng là 20 tháng. Như vậy mỗi khách hàng rời mạng sẽ mất đi 1.4 triệu đồng. Con số thiệt hại sẽ rất lớn khi có hàng trăm ngàn khách hàng rời mạng mỗi năm. Trong bối cảnh đó đồng thời được sự đồng ý của Thầy PGS.TS Võ Trung Hùng, tác giả chọn đề tài “Ứng dụng khai phá dữ liệu dự đoán khách hàng rời mạng viễn thông” cho luận văn tốt nghiệp của mình. Bài toán tập trung tìm phương pháp sử dụng công cụ khai phá dữ liệu để dự đoán được những khách hàng nào nguy cơ rời mạng cao để có biện pháp giữ khách hàng ở lại. Việc triển khai bài toán có ý nghĩa thời sự cao, nhất là đối với EVNTelecom trong hoàn cảnh này.

2. Mục tiêu của đề tài

Mục tiêu của đề tài là nghiên cứu ứng dụng các kỹ thuật khai phá dữ liệu trong công tác dự báo khách hàng rời mạng. Dựa trên kho dữ liệu tích lũy trong những năm gần đây để làm dữ liệu huấn

luyện, tính toán dự báo khách hàng rời mạng. Thực hiện đánh giá kết quả dự đoán trên cơ sở các mẫu dữ liệu trích ra từ dữ liệu thu thập.

3. Nội dung triển khai

Để giải quyết bài toán dự báo khách hàng rời mạng, luận văn tiến hành nghiên cứu các kỹ thuật khai phá dữ liệu, lựa chọn mô hình ứng dụng cho bài toán, tìm hiểu công cụ triển khai ứng dụng khai phá dữ liệu.

Bước tiếp theo sẽ tổ chức thu thập dữ liệu từ các bộ phận: quản lý khách hàng, bộ phận tính cước, bộ phận quản lý nợ và chăm sóc khách hàng. Thực hiện xây dựng cơ sở dữ liệu trên hệ quản trị cơ sở dữ liệu SQL Server 2005. Tiến hành lọc, phân tích và nạp dữ liệu chuẩn bị khai khoáng.

Bước kế tiếp thực hiện nghiên cứu xây dựng mô hình giải quyết bài toán. Trước hết sẽ tiến hành xây dựng mô hình gom cụm để phân khách hàng thành 5 cụm dựa theo các tiêu chí danh sách dịch vụ đang sử dụng, mức độ trung thành, doanh thu hàng tháng, quá trình thanh toán cước dịch vụ và quá trình chăm sóc khách hàng. Sau đó xây dựng cây quyết định dự đoán khách hàng rời mạng cho từng cụm. Thực hiện kiểm tra mô hình trên 15% lượng khách hàng rời mạng. Đây là một quá trình lặp để lựa chọn mô hình hữu ích nhất.

Bước cuối cùng thực hiện dự đoán khả năng khách hàng rời mạng cho lượng khách hàng đang hoạt động. Trong quá trình triển khai luận văn, tôi tiến hành tìm hiểu cơ sở lý thuyết của thuật toán cây quyết định và thuật toán gom cụm – hai thuật toán sẽ được sử dụng trong mô hình ứng dụng dự đoán. Đồng thời cũng sẽ tiến hành nghiên cứu các công cụ khai phá dữ liệu của Microsoft SQL Server 2005 để làm công cụ triển khai mô hình.

4. Bố cục của luận văn

Ngoài phần mở đầu và kết luận, trong luận văn tôi đề cập đến các nội dung chính sau:

Chương 1: Nghiên cứu tổng quan về kho dữ liệu, mô hình tổng quát về kỹ thuật khai phá dữ liệu và kỹ thuật dự báo trong khai phá dữ liệu.

Chương 2: Phân tích thiết kế hệ thống, trong chương này các nội dung tôi đề cập đến đó là: Mô tả ứng dụng, đề xuất giải pháp ứng dụng kỹ thuật khai phá dữ liệu và cuối cùng là phân tích thiết kế hệ thống.

Chương 3: Phát triển và Demo ứng dụng, chương này đề cập đến xây dựng mô hình, kiểm tra và đánh giá mô hình dự đoán.

CHƯƠNG 1. NGHIÊN CỨU TỔNG QUAN

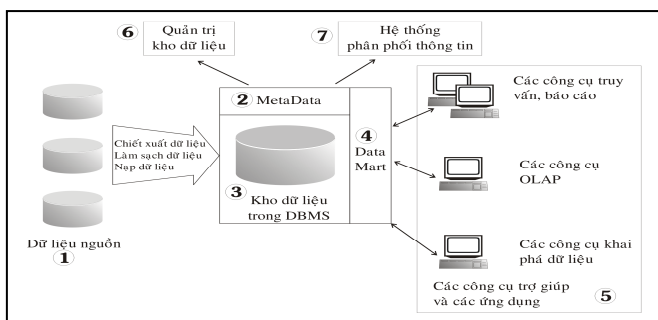
Trong chương này, chúng tôi trình bày một số khái niệm về kho dữ liệu, khai phá dữ liệu và các ứng dụng.

1.1. KHO DỮ LIỆU

1.1.1. Khái niệm

Ngày nay định nghĩa cho kho dữ liệu bao gồm các công cụ thông minh dùng để trích rút, biến đổi và nạp dữ liệu vào kho, cũng như để quản lý và lưu trữ siêu dữ liệu (metadata), các chức năng khác như thanh lọc, thu nạp, phân tích, trích rút, biến đổi (ETL) và quản lý dữ liệu được coi là các thành phần cốt yếu của một kho dữ liệu.

1.1.2. Kiến trúc kho dữ liệu



Hình 1-1: Kiến trúc kho dữ liệu.

1.1.3. Quy trình xây dựng

Trong quá trình xây dựng kho dữ liệu cần chú ý một số vấn đề sau:

- Cần ước lượng kích thước cần thiết của kho dữ liệu.
- Tối thiểu hoá kích thước của bảng sự kiện (fact table).

1.1.4. Ứng dụng kho dữ liệu

- Chiết xuất, tổng hợp và chuyển đổi từ các dữ liệu thô sang dạng các dữ liệu chất lượng cao và có tính ổn định, giúp cho việc nâng cao các kỹ thuật biểu diễn thông tin truyền thống.

- Các kho dữ liệu được sử dụng để hỗ trợ cho phân tích trực tuyến (OLAP), xác định xem giả thuyết đúng hay sai.

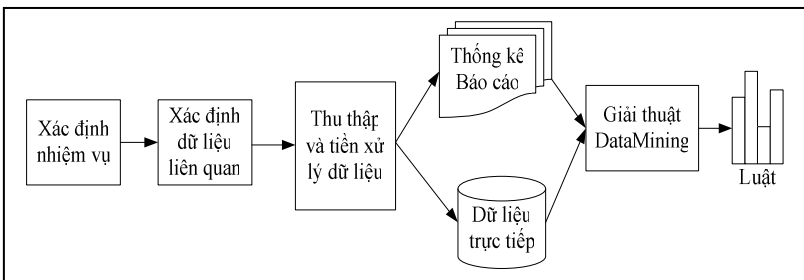
- Hỗ trợ cho công nghệ khai phá dữ liệu (data mining).

1.2. KHAI PHÁ DỮ LIỆU

1.2.1. Khái niệm

Khai phá dữ liệu là quá trình tìm kiếm các mẫu mới, những thông tin tiềm ẩn mang tính dự đoán trong các khối dữ liệu lớn.

1.2.2. Mô hình tổng quát



Hình 1-2: Sơ đồ mô tả quá trình khai phá dữ liệu.

1.2.3. Các công cụ khai phá dữ liệu

Có hai nhóm công cụ khai phá dữ liệu: Các công cụ mã nguồn mở (open-source tools) và nhóm công cụ thương mại.

❖ Các công cụ mã nguồn mở (open-source): R (www.r-project.org); Tanagra (eric.univ-lyon2.fr/~ricco/tanagra/); Weka (www.cs.waikato.ac.nz/ml/weka); YALE (rapid-i.com); KNIME (www.knime.org); Orange (www.ailab.si/orange)

❖ Các công cụ thương mại: Intelligent Miner (IBM); Microsoft data mining tools (MS SQL Server 2000/2005/2008); Oracle Data Mining; Enterprise Miner (SAS Institute)

1.3. DỰ BÁO

1.3.1. Khái niệm:

Dự báo là một khoa học và nghệ thuật tiên đoán những sự việc sẽ xảy ra trong tương lai, trên cơ sở phân tích khoa học về các dữ liệu đã thu thập được. Khi tiến hành dự báo cần căn cứ vào việc thu thập, xử lý số liệu trong quá khứ và hiện tại để xác định xu hướng vận động của các hiện tượng trong tương lai nhờ vào một số mô hình toán học (định lượng). Dự báo cũng có thể là một dự đoán chủ quan hoặc trực giác về tương lai (định tính) và để dự báo định tính được chính xác hơn, người ta cố loại trừ những tính chủ quan của người dự báo.

1.3.2. Các phương pháp dự báo:

Bảng 1-1: Tổng hợp một số phương pháp dự báo

1.	Tiên đoán (Genius forecasting)
2.	Ngoại suy xu hướng (Trend Extrapolation)
3.	Phương pháp chuyên gia (Consensus method)
4.	Phương pháp mô phỏng (Stimulation)
5.	Phương pháp ma trận tác động qua lại
6.	Phương pháp kịch bản (Scenario)
7.	Phương pháp cây quyết định (Decision Tree)
8.	Phương pháp dự báo tổng hợp

1.3.3. Kỹ thuật dự báo dựa vào khai phá dữ liệu

Có nhiều kỹ thuật dự báo vào khai phá dữ liệu, ở đây xin trình bày kỹ thuật gom cụm:

Gom cụm là việc nhóm một tập dữ liệu lớn thành một số nhóm nhỏ, mục đích của gom cụm là tìm những mẫu chung hoặc gom các mẫu dữ liệu tương tự nhau thành nhóm theo một tiêu chuẩn nào đó. Các mẫu dữ liệu trong nhóm thì tương tự nhau hơn các mẫu dữ liệu ở các nhóm khác nhau. Gom cụm được ứng dụng nhiều trong các bài toán thống kê phân tích, phân loại đối tượng. Là công cụ độc lập để xem xét phân bố dữ liệu và là bước tiền xử lý cho các thuật toán.

Để giải bài toán gom cụm, trước hết phải chọn được phép đo khoảng cách và phương pháp gom cụm. Trong đó việc lựa chọn phép đo có ý nghĩa quyết định chất lượng gom cụm.

1.4. Một số nghiên cứu về khai phá dữ liệu trong viễn thông

Đối với các doanh nghiệp viễn thông, trong quá trình sản xuất kinh doanh của mình đã thu thập được khối lượng khổng lồ các loại dữ liệu:

- + Dữ liệu chi tiết cuộc gọi.
 - + Thông tin khách hàng như mức cước sử dụng, nghề nghiệp, giới tính khách hàng, các dịch vụ gia tăng đã sử dụng...
 - + Dữ liệu liên quan đến vận hành hệ thống.
- Ứng dụng khai phá dữ liệu trong viễn thông áp dụng trong ba lĩnh vực chủ yếu:

- ❖ Ứng dụng trong marketing.
- ❖ Ứng dụng trong phát hiện gian lận.
- ❖ Ứng dụng trong quản lý vận hành hệ thống.

CHƯƠNG 2. PHÂN TÍCH THIẾT KẾ HỆ THỐNG

Trong chương này, chúng tôi trình bày các yêu cầu của Hệ thống khai phá dữ liệu ứng dụng cho việc dự báo khách hàng rời mạng viễn thông, tiến hành các bước phân tích và thiết kế của Hệ thống dự báo khách hàng rời mạng.

2.1. MÔ TẢ ỨNG DỤNG

2.1.1. Giới thiệu về EVNTelecom

EVNTelecom là một doanh nghiệp trực thuộc Tập đoàn Điện lực Việt Nam, EVNTelecom được phép cung cấp đầy đủ các dịch vụ viễn thông tại Việt Nam.

2.1.2. Phân tích yêu cầu

Đề tài mang ý nghĩa kỳ vọng tạo ra sự khác biệt trong công tác chăm sóc khách hàng tại EVNTelecom. Việc triển khai thành công sẽ đem lại ý nghĩa thiết thực, giúp nhà cung cấp EVNTelecom trong hoạch định chiến lược phát triển. Mục tiêu của bài toán là dự đoán khả năng rời mạng của từng khách hàng để có biện pháp ứng phó.

2.1.2.1. Xác định yêu cầu

Hệ thống thực hiện được các chức năng :

- Gộp nhóm các đối tượng có nguy cơ rời mạng cao để xây dựng chính sách cho nhóm đối tượng.
- Dự đoán được khả năng rời mạng của từng khách hàng để có biện pháp ứng phó từng trường hợp.
- Tỷ lệ lỗi dự đoán ở mức cho phép.

2.1.2.2. Phạm vi bài toán

Dữ liệu bài toán là dữ liệu kinh doanh của EVNTelecom trên địa bàn Quảng Nam từ khi triển khai kinh doanh từ năm 2005 đến năm 2009.

Mô hình được xây dựng trên lượng thông tin khách hàng tích lũy từ 2005 đến 2009. Các dữ liệu phát sinh theo chu kỳ tháng lấy từ tháng 8 năm 2008 đến tháng 8 năm 2009.

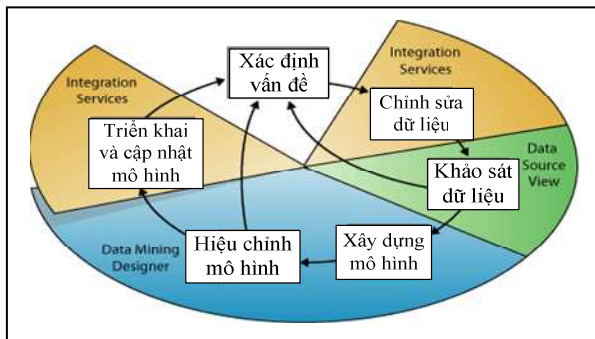
2.1.2.3. Yêu cầu về hệ thống

Bộ xử lý Intel Xeon 2GHz, 4GB bộ nhớ RAM, đĩa cứng 320GB Raid mức 5. Hệ điều hành sử dụng hệ điều hành Microsoft Windows Server 2003 Service Pack 2. Dữ liệu được tổ chức trên hệ quản trị cơ sở dữ liệu Microsoft SQL Server 2005. Công cụ khai phá dữ liệu sử dụng bộ công cụ Analysis Services tích hợp cùng Microsoft SQL Server 2005. Công cụ lập trình sử dụng Business Intelligence Development Studio.

2.2. THIẾT KẾ HỆ THỐNG

2.2.1. Mô hình tổng quát.

Mô hình tổng quát của quá trình xây dựng hệ thống như sau:



2.2.2. Phân tích thiết kế hệ thống:

2.2.2.1. Danh sách các actor:

Danh sách actor:

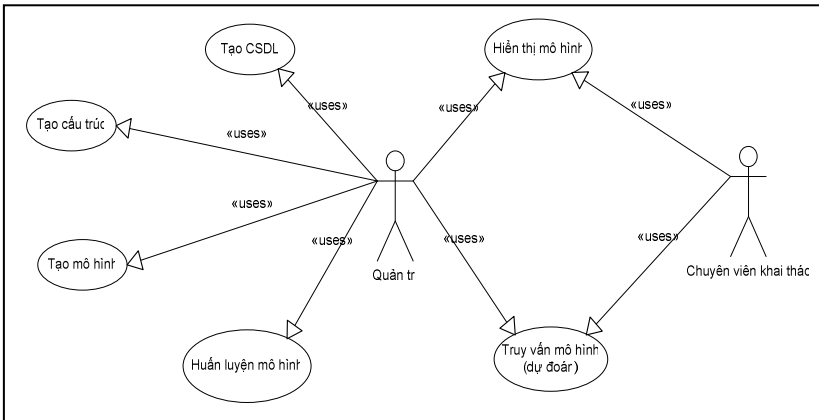
STT	Tên actor	Diễn giải
1	Chuyên viên khai thác hệ thống	Là các cán bộ, chuyên viên sử dụng phần mềm để phân tích dữ liệu và dự báo khả năng rời mạng của khách hàng
2	Quản trị hệ thống	Quản trị dữ liệu

Danh sách use case:

STT	Tên use case	Diễn giải
1	Tạo CSDL	Mở kết nối đến CSDL trên server, tạo các đối tượng datasource, dataview
2	Tạo cấu trúc	Tạo cấu trúc của mô hình khai phá dữ liệu dựa trên các thuộc tính đầu vào (input) được lựa chọn và đầu ra để dự đoán (predict)
3	Tạo mô hình	Áp dụng các thuật toán khai phá dữ liệu như cây quyết định, gom cụm, điều chỉnh các tham số thuật toán để hoàn chỉnh mô hình.
4	Huấn luyện mô hình	Sử dụng dữ liệu từ CSDL đã kết nối đưa vào huấn luyện mô hình được tạo
5	Hiển thị mô hình	Hiển thị kết quả huấn luyện của mô hình (xem cây, xem kết quả gom cụm)
6	Truy vấn mô hình (dự đoán)	Đưa dữ liệu đầu vào và hiển thị kết quả dự đoán.

2.2.2.2. Sơ đồ use case:

Sơ đồ use case của hệ thống như sau:



2.2.3. Các bước xây dựng hệ thống

2.2.3.1. Thu thập dữ liệu

Tiến hành khảo sát các các bộ phận tác nghiệp để thu thập dữ liệu như:

- + Bộ phận quản lý khách hàng
- + Bộ phận cước
- + Bộ phận quản lý nợ
- + Bộ phận chăm sóc khách hàng

2.2.3.2. Mô tả dữ liệu

Dữ liệu khách hàng phát sinh khi có hợp đồng được ký kết với khách hàng. Bảng sau mô tả tên, kiểu giá trị và ý nghĩa của các trường:

Số thứ tự	Tên trường	Kiểu dữ liệu	Ý nghĩa
01	Ma_KH	Varchar	Dùng làm khoá chính
02	Ten_KH	Varchar	Tên của khách hàng
03	Gioitinh	Bit	Giới tính của khách hàng

04	Diachi	Varchar	Địa chỉ khách hàng
05	CMND	Varchar	Số CMND của khách hàng
06	So_HD	Varchar	Số hợp đồng
07	Ngay_HD	Datetime	Ngày ký hợp đồng
08	So_thuebao	Varchar	Số thuê bao
09	Ngayhoamang	Datetime	Ngày hoà mạng
10	Loai_KH	Varchar	Phân loại khách hàng
11	Hinhthuc_tt	Varchar	Hình thức thanh toán
12	Trangthai	Varchar	Trạng thái hoạt động của thuê bao
13	Thietbi	Varchar	Loại thiết bị đầu cuối cung cấp cho khách hàng
14	Nhanvien_BH	Varchar	Nhân viên bán hàng

- + Bảng mô tả dữ liệu cước dịch vụ
- + Bảng mô tả dữ liệu khách hàng rời mạng
- + Bảng mô tả dữ liệu nợ cước dịch vụ
- + Bảng mô tả dữ liệu chăm sóc khách hàng

2.2.3.3. Chọn dữ liệu

Dữ liệu cước phát sinh theo chu kỳ tháng thu thập từ tháng 8/2008 đến tháng 10/2009.

- + Dữ liệu khách hàng
- + Dữ liệu cước sử dụng dịch vụ
- + Dữ liệu khách hàng rời mạng
- + Dữ liệu chăm sóc khách hàng

2.2.3.4. Đánh giá chất lượng dữ liệu và làm sạch dữ liệu

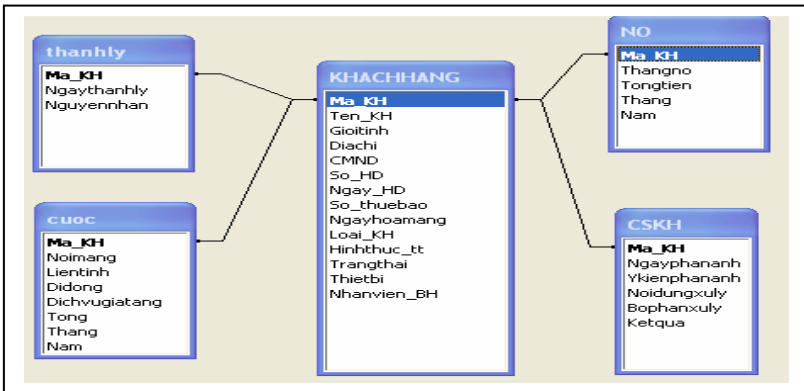
Đánh giá chất lượng dữ liệu

- Dữ liệu cần xử lý là dữ liệu thu thập từ các bộ phận quản lý tác nghiệp. Vì vậy dữ liệu có tính trung thực cao, hầu hết đều phản ánh đúng ngữ nghĩa.

- Dữ liệu có bị thiếu ở một vài trường do việc cập nhập ban đầu không đầy đủ hoặc chưa được nhập liệu. Phần dữ liệu bị thiếu chỉ nằm ở thông tin khách hàng, một số thông tin bổ sung nghiệp vụ. Tuy nhiên các dữ liệu bị thiếu không ảnh hưởng nhiều đến chất lượng của mô hình.

2.2.3.5. *Nạp dữ liệu*

Tạo cơ sở dữ liệu



Hình 2-1: Lược đồ quan hệ cơ sở dữ liệu khai khoáng.

Nạp dữ liệu

Thực hiện thu thập dữ liệu tác nghiệp từ các cơ sở dữ liệu Oracle, SQL Server bằng các công cụ import và export. Để đảm bảo tính nhất quán dữ liệu, sử dụng các câu lệnh transact SQL để nạp dữ liệu từ cơ sở dữ liệu tạm vào cơ sở dữ liệu khai khoáng.

2.3. KHAI PHÁ DỮ LIỆU VỚI Microsoft SQL Server 2005

2.3.1. Microsoft SQL server 2005 Analysis Services

2.3.1.1. Môi trường phát triển ứng dụng

Microsoft cung cấp các công cụ để phát triển ứng dụng khai phá dữ liệu:

- Business Intelligence Development Studio (BI Dev Studio) là môi trường phát triển ứng dụng trong SSAS.

- Giao diện lập trình ứng dụng API giúp lập trình viên dễ dàng phát triển ứng dụng.

2.3.1.2. Các thuật toán data mining trong Microsoft SQL Server 2005

Danh sách và các ứng dụng của thuật toán tích hợp với Microsoft SQL Server 2005:

Bảng 2-1: Ứng dụng các thuật toán

Nhiệm vụ	Thuật toán Microsoft sử dụng
Dự đoán thuộc tính rời rạc	Microsoft Decision Trees Algorithm Microsoft Naive Bayes Algorithm Microsoft Clustering Algorithm Microsoft Neural Network Algorithm
Dự đoán thuộc tính liên tục	Microsoft Decision Trees Algorithm Microsoft Time Series Algorithm
Dự đoán 1 trình tự	Microsoft Sequence Clustering Algorithm
Luật kết hợp	Microsoft Association Algorithm Microsoft Decision Trees Algorithm
Gom cụm	Microsoft Clustering Algorithm Microsoft Sequence Clustering Algorithm

2.3.2. OLE DB for Data Mining

2.3.2.1. Giới thiệu OLE DB for Data Mining

2.3.2.2. Các khái niệm cơ bản trong OLE DB for DM

- + Case
- + Case key
- + Nested key

+ Mô hình data mining

DMX

Việc xây dựng mô hình trải qua ba bước cơ bản: tạo mô hình, huấn luyện mô hình và dự đoán.

Ngoài câu lệnh truy vấn như trên, DMX định nghĩa thêm nhiều hàm có thể sử dụng kết hợp với câu truy vấn để cho kết quả mong muốn.

2.4. MỘT SỐ THUẬT TOÁN ĐẶC BIỆT

Trong phần này luận văn chỉ trình bày chi tiết các thuật toán sẽ được sử dụng cho mô hình ứng dụng của luận văn.

2.4.1. Thuật toán gom cụm

Thuật toán Microsoft Clustering cho phép sử dụng hai phương pháp đo khoảng cách cụm : K-Means và Expectation Maximization (EM).

Thuật toán K-Means xác định đối tượng trở thành thành viên của cụm có khoảng cách từ đối tượng đến tâm cụm nhỏ nhất, sử dụng phép đo khoảng cách O Clit. Kết thúc thuật toán, mỗi đối tượng thuộc về một cụm duy nhất. Thuật toán EM sử dụng phép đo xác suất để xác định tư cách thành viên bằng cách xem xét một đường cong cho mỗi chiều với điểm trung tâm và độ lệch chuẩn. Nếu một điểm nằm bên trong đường cong, nó thuộc về một cụm với xác suất chắc chắn. Microsoft Decision Trees

Thuật toán cây quyết định của Microsoft (Microsoft Decision Trees) là thuật toán cây quyết định lai, hỗ trợ phân lớp và hồi quy. Tùy thuộc tham số, cây quyết định có thể có sự phân nhánh và hình dạng khác nhau. Một mô hình có thể có nhiều cây, các cây này có liên kết với nhau.

Microsoft Decision Trees sử dụng cây phân lớp theo xác suất, mặc định dùng Bayesian score làm tiêu chuẩn rẽ nhánh thay vì Entropy. Không thực hiện bước tỉa cây, giới hạn các biến nhập có ít hơn 100 trạng thái.

Khi làm việc với số liệu liên tục, thường là dữ liệu kiểu số, đều được rời rạc hoá.

CHƯƠNG 3. PHÁT TRIỂN HỆ THỐNG

Chương này chúng tôi trình bày các bước phát triển và thử nghiệm Mô hình khai phá dữ liệu phục vụ cho việc dự đoán khách hàng rời mạng viễn thông.

3.1. XÂY DỰNG CƠ SỞ DỮ LIỆU

Cơ sở dữ liệu được xây dựng trên môi trường SQL Server 2005. Sau khi xử lý và làm sạch dữ liệu, dữ liệu được nạp vào CSDL theo đúng lược đồ đã thiết kế ở trên, với các thông tin như thông tin thuê bao, cước, thông tin chăm sóc khách hàng.

3.2. XÂY DỰNG MODULE KHAI PHÁ DỮ LIỆU

3.2.1. Xây dựng mô hình

3.2.1.1. Khảo sát dữ liệu

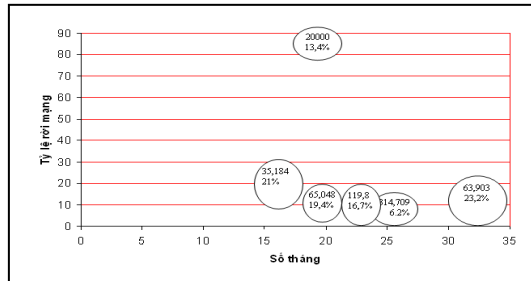
3.2.1.2. Chuẩn bị dữ liệu

3.2.1.3. Xây dựng mô hình

Mô hình gom cụm

Thực hiện gom khách hàng theo tiêu chí : mức độ trung thành, khu vực tiềm năng, mức cước trung bình, tình hình trả nợ cước để phân loại khách hàng. Mô hình được thực thi trên thuật toán Expectation Maximization (EM) trong họ thuật toán Microsoft Clustering đã trình bày ở chương 2.

Để số lượng thành viên mỗi cụm không quá lớn và số lượng các cụm đủ để nhận thấy được sự khác biệt, dự kiến phân thành 4 đến 8 cụm. Sau khi thử lần lược các trường hợp lựa chọn mô hình có 6 cụm cho kết quả tốt nhất. Hình sau thể hiện sự phân bố của các cụm



Hình 3-1: Biểu đồ phân bố cụm.

Các vòng tròn trên biểu đồ biểu diễn các cụm theo hai chiều là số tháng sử dụng trung bình của các thành viên cụm và tỉ lệ rời mạng của cụm. Số liệu bên trong các vòng tròn chỉ ARPU bình quân của cụm và tỉ lệ phần trăm số thành viên của cụm với toàn mô hình

Mô hình dự đoán khách hàng rời mạng

Qua đánh giá bằng các công cụ thống kê, khảo sát các nhà quản lý, để dự đoán khách hàng rời mạng hay không có thể xem xét các yếu tố :

- + Thông tin cá nhân khách hàng
- + Doanh thu phát sinh
- + Tình hình trả nợ cước
- + Chất lượng dịch vụ
- + Chất lượng thiết bị đầu cuối (máy điện thoại)
- + Quá trình phản hồi, giải quyết khiếu nại

Luận văn lựa chọn mô hình dự đoán trên cây phân loại cho thuộc tính trạng thái thuê bao sử dụng thuật toán Microsoft Decision Trees. Nếu giá trị dự đoán của thuộc tính trạng thái thuê bao là “1” thì thuê bao đó được dự đoán là rời mạng. Trường hợp ngược lại là thuê bao không rời mạng.

Ngoài việc dự đoán khách hàng thuộc lớp “rời mạng” hoặc “đang hoạt động” theo luật cây quyết định còn phải xác định thêm xác suất mỗi khách hàng thuộc lớp nào là bao nhiêu. Thuật toán Microsoft Decision Tree hỗ trợ tính toán giá trị này bằng công thức xác suất Bayesian.

3.2.2. Kiểm tra và đánh giá mô hình

Để kiểm tra tính chính xác mô hình dự đoán, cần thực hiện truy vấn tri thức mô hình bằng dữ liệu kiểm tra. Sử dụng 15% dữ liệu đã trích ra từ bước huấn luyện mô hình để kiểm tra các mô hình.

3.2.3. Triển khai mô hình

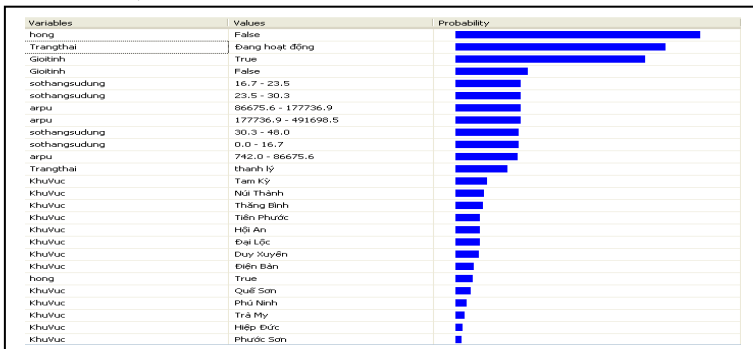
Việc triển khai mô hình của luận văn có thể thực hiện trên hai cách:

- Sử dụng câu lệnh truy vấn DMX trong môi trường BI Dev Studio.
- Sử dụng chương trình demo được xây dựng từ ngôn ngữ C# trên môi trường Visual Studio 2005.

3.2.3.1. Sử dụng môi trường BI Dev Studio

Một số kết quả thu thập được trên môi trường BI Dev Studio như sau:

Mô hình gom cụm: Đặc trưng chung của các cụm (Cluster Characteristics):



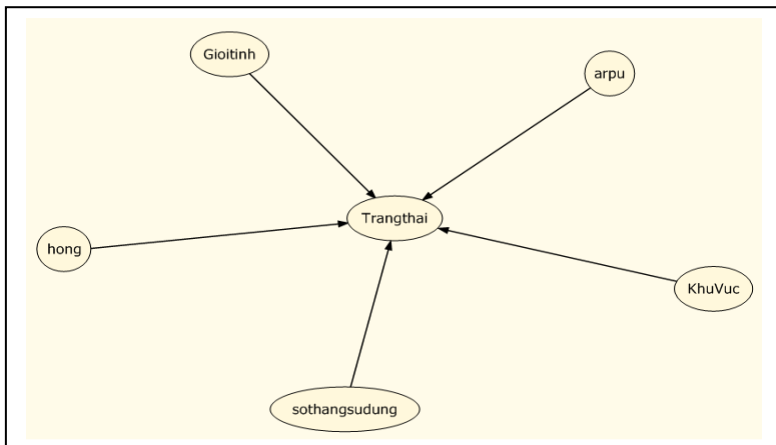
So sánh các cụm (cluster Discrimination):

Hình sau so sánh cụm 5 (tỉ lệ rời mạng cao nhất) khác biệt với cụm 1 ở những thuộc tính sau:

Variables	Values	Favors Cluster 5	Favors Cluster 1
arpu	20000.0		
arpu	20000.0 - 6347518.0		
Trangthai	thanh lý		
Trangthai	Đang hoạt động		
sothangsudung	25.3 - 41.2		
sothangsudung	0.0 - 25.3		
hong	False		
hong	True		
sothangsudung	41.2 - 48.0		

Mô hình cây quyết định (dự đoán):

Mô hình phụ thuộc giữa các thuộc tính (Dependency Network):



Sơ đồ cây quyết định:

KẾT LUẬN

Đánh giá kết quả đề tài

Về lý thuyết, luận văn đã trình bày được cơ sở lý thuyết liên quan đến hai thuật toán gom cụm và thuật toán cây quyết định. Sau đó đi sâu trình bày phương pháp tính khoảng cách cụm và phương pháp lựa chọn thuộc tính rẽ nhánh. Nghiên cứu quy trình triển khai ứng dụng khai phá dữ liệu. Ngoài ra luận văn cũng tìm hiểu các vấn đề cơ bản về công cụ Microsoft SQL Server 2005 Analysis Services và các công cụ liên quan.

Đối với bài toán dự đoán khách hàng rời mạng, luận văn đề xuất xây dựng hai mô hình dự đoán: mô hình gom cụm và mô hình dự đoán theo cây quyết định. Mô hình gom cụm thực hiện gom nhóm các khách hàng dựa trên sự tương tự của các độ đo. Kết quả gom cụm là đối tượng của nghiên cứu của các bộ phận chuyên môn, đồng thời cũng là đầu vào của thuật toán cây quyết định nhằm có kết quả dự đoán tin cậy hơn. Mô hình dự đoán đã cho kết quả tương đối ổn định ở các mẫu kiểm tra khác nhau, tỉ lệ lỗi ở mức chấp nhận được và có thể triển khai thực tế. Kết quả dự đoán của mô hình được sử dụng bởi các bộ phận nghiên cứu thị trường, chăm sóc khách hàng nhằm tìm ra giải pháp giảm thiểu khách hàng rời mạng.

Ngoài ra luận văn cũng đã xây dựng chương trình demo minh họa khả năng triển khai lập trình ứng dụng khai phá dữ liệu đến người dùng cuối.

Nhìn chung, luận văn đã cơ bản hoàn thành các mục tiêu đề ra. Để đưa mô hình dự đoán vào ứng dụng cần tiếp tục đầu tư thu thập một vài dữ liệu cá nhân của khách hàng mà đề tài chưa có điều kiện thu thập, tổ chức lưu trữ và thu thập dữ liệu bộ phận chăm sóc khách

hàng. Triển khai dự đoán, kiểm chứng thực tế và đánh giá kết quả một cách thường xuyên. Bản thân nhận thấy đây là hướng tiếp cận đúng đắn và có tính thực tiễn cao.

Hạn chế

Đề tài chỉ mới thực hiện dự đoán khách hàng rời mạng nhưng chưa dự đoán được thời điểm rời mạng.

Phạm vi áp dụng của đề tài

Với việc triển khai hệ thống thử nghiệm cho thấy khả năng ứng dụng kết quả này trong việc dự đoán khách hàng rời mạng tại Viễn thông điện lực Quảng Nam, từ đó giúp cho đơn vị tìm hiểu nguyên nhân khách hàng rời mạng, cũng như có các chính sách chăm sóc, hỗ trợ khách hàng tốt hơn.

Qua nghiên cứu này tôi hy vọng được đóng góp phần nhỏ trong công tác chăm sóc khách hàng, giảm thiểu thuê bao rời mạng. Không riêng gì khách hàng của EVNTelecom, các đơn vị cung cấp dịch vụ khác cũng có thể xem đây như là tư liệu tham khảo trong quá trình triển khai ứng dụng của mình.

Hướng phát triển

Hướng phát triển trong thời gian tới của đề tài là nghiên cứu sự tác động của các yếu tố bên ngoài như chính sách điều tiết vĩ mô, sự tác động của các đối thủ cạnh tranh trên thị trường, tâm lý tiêu dùng... lên kết quả của mô hình. Các yếu tố này được xem xét như các tham số điều chỉnh mô hình để có kết quả dự đoán chính xác hơn, tiến đến thực hiện mô hình dự đoán thời điểm rời mạng.