ĐẠI HỌC THÁI NGUYÊN

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

NGUYỄN NGỌC HẢI

NGHIÊN CỨU MỘT SỐ THUẬT TOÁN KHAI PHÁ LUẬT KẾT HỢP VÀ THỬ NGHIỆM ỨNG DỤNG VÀO KHAI PHÁ CƠ SỞ DỮ LIỆU BẢO HIỂM Y TẾ TẠI BẢO HIỂM XÃ HỘI BẮC GIANG

LUẬN VĂN THẠC SĨ KHOA HỌC MÁY TÍNH

THÁI NGUYÊN - 2013

ĐẠI HỌC THÁI NGUYÊN

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

NGUYỄN NGỌC HẢI

NGHIÊN CỨU MỘT SỐ THUẬT TOÁN KHAI PHÁ LUẬT KẾT HỢP VÀ THỬ NGHIỆM ỨNG DỤNG VÀO KHAI PHÁ CƠ SỞ DỮ LIỆU BẢO HIỂM Y TẾ TẠI BẢO HIỂM XÃ HỘI BẮC GIANG

Chuyên ngành :KHOA HỌC MÁY TÍNH

Mã số :60 48 01

LUẬN VĂN THẠC SĨ KHOA HỌC MÁY TÍNH

Người hướng dẫn khoa học:TS. PHÙNG VĂN ỔN

THÁI NGUYÊN - 2013

MỤC LỤC

Trang

LỜI CẢM ƠN

Tác giả xin chân thành cảm ơn các thầy giáo, cô giáo Trường Đại học Công nghệ thông tin và Truyền thông Thái Nguyên và các thầy Viện Công nghệ thông tin - Đại học quốc gia Hà Nội, đã tận tâm giảng dạy các kiến thức trong hai năm học qua cùng với sự cố gắng hết mực của bản thân.

Đặc biệt tôi xin bày tỏ sự biết ơn sâu sắc đến thầy giáo Tiến sĩ Phùng Văn Ổn, PGS. TS Ngô Quốc Tạo người đã tận tình giảng dạy và hướng dẫn tôi thực hiện luận văn này.

Tác giả cũng xin chân thành cảm ơn lãnh đạo Bảo hiểm xã hội tỉnh Bắc Giang, các bạn đồng nghiệp, các bạn trong lớp cao học CK10B đã tạo điều kiện, giúp đỡ tôi trong suốt thời gian qua.

Rất mong nhận được sự góp ý của các thầy, cô, bạn bè, đồng nghiệp để luận văn có thể phát triển và hoàn thiện hơn.

LỜI CAM ĐOAN

Tôi xin cam đoan đây là công trình nghiên cứu của riêng tôi.

Các số liệu, kết quả nêu trong luận văn và chưa từng được ai công bố trong bất kỳ công trình nào khác.

Thái Nguyên, tháng 09 năm 2013

TÁC GIẢ

Nguyễn Ngọc Hải

DANH MỤC CÁC KÝ HIỆU, CÁC CHỮ VIẾT TẮT

Từ viết tắt	Tiếng Anh	Tiếng Việt
C_k	C_k	Tập các K – itemset ứng cử
Conf	Confidence	Độ tin cậy
CSDL	Database	Cơ sở dữ liệu
DW	Data Warehouse	Kho dữ liệu
Item	Item	Khoản mục
Itemset	Itemset	Tập các khoản mục
K- itemset	K- itemset	Tập gồm K mục
KDD	Knowledge Discovery and Data Mining	Kỹ thuật phát hiện tri thức và khai phá dữ liệu
L_k	L_k	Tập các K - itemset phổ biến
Minconf	Minimum Confidence	Độ tin cậy tối thiểu
Minsup	Minimum Support	Độ hỗ trợ tối thiểu
OLAP	On Line Analytical Processing	Phân tích trực tuyến
MOLAP	Multidimensional OLAP	Phân tích đa chiều trực tuyến
ROLAP	Relational OLAP	Phân tích quan hệ trực tuyến
pre(k, s)	pre(k, s)	Tiếp đầu dãy có độ dài **k** của **s**
Record	Record	Bản ghi
Supp	Support	Độ hỗ trợ
TID	Transaction Indentification	Định danh giao tác
SQL	Structured Query Language	Ngôn ngữ truy vấn có cấu trúc
SQO	Semantic Query Optimization	Tối ưu truy vấn ngữ nghĩa
DBSCAN	Density Based Spatial Clustering of Application with Noise	Thuật toán phân lớp dựa vào vị trí địa phương
DENCLUE	DENsity Based CLUstEring	Thuật toán phân lớp cơ bản (tổng quát)
ADO	Activate X Data Object	Đối tượng dữ liệu Active X
DFS	Depth First Search	Tìm kiếm theo chiều sâu
BFS	Breadth First Search	Tìm kiếm theo chiều rộng
DHP	Direct Hashing and Pruning	Bảng băm trực tiếp và sự cắt tỉa
PHP	Perfect Hashing and Pruning	Bảng băm lý tưởng và sự cắt tỉa
I/O	Input/Output	Vào/ra
BHXH		Bảo hiểm xã hội
BHYT		Bảo hiểm y tế
KCB		Khám chữa bệnh

DANH MỤC CÁC BẢNG

Trang

DANH MỤC CÁC HÌNH VẼ, ĐỒ THỊ

Trang

MỞ ĐẦU

Ngày nay, thông tin được coi là tài sản quan trọng của các tổ chức, doanh nghiệp và các cá nhân. Cá nhân hoặc tổ chức nào thu thập và hiểu được thông tin, và hành động kịp thời dựa trên các thông tin đó sẽ có được kết quả tốt trong lĩnh vực sản xuất, kinh doanh, quản lý ... của mình. Chính vì lý do đó, việc tạo ra thông tin, tổ chức lưu trữ và khai thác thông tin ngày càng trở nên quan trọng và gia tăng không ngừng.

Sự tăng trưởng vượt bậc của các cơ sở dữ liệu (CSDL) trong các hoạt động như: sản xuất kinh doanh, thương mại, quản lý đã làm nảy sinh và thúc đẩy sự phát triển của kỹ thuật thu thập, lưu trữ, phân tích và khai phá dữ liệu… không chỉ bằng các phương pháp thông thường như: thống kê mà đòi hỏi cách xử lý thông minh hơn, hiệu quả hơn. Từ đó các nhà quản lý có được thông tin hữu ích để tác động lại quá trình sản xuất, kinh doanh của mình… đó là tri thức. Các kỹ thuật cho phép ta khai thác được tri thức hữu dụng từ CSDL (lớn) được gọi là các kỹ thuật ***khai phá dữ liệu*** (DM – Data Mining). Khai phá luật kết hợp là một nội dung quan trọng trong khai phá dữ liệu.

Luận văn tìm hiểu về luật kết hợp và ứng dụng thử nghiệm khai phá cơ sở dữ liệu Bảo hiểm y tế nhằm hỗ trợ cho công tác quản lý, sử dụng quỹ BHYT tại tỉnh Bắc Giang .

Chương 1
TỔNG QUAN VỀ KHAI PHÁ DỮ LIỆU

1.1. Tổ chức và khai thác cơ sở dữ liệu truyền thống

Việc dùng các phương tiện tin học để tổ chức và khai thác cơ sở dữ liệu (CSDL) đã được phát triển từ những năm 60 của thế kỉ trước. Từ đó cho đến nay, rất nhiều CSDL đã được tổ chức, phát triển và khai thác ở mọi quy mô và các lĩnh vực hoạt động của con người và xã hội. Cho đến nay, số lượng CSDL đã trở nên khổng lồ bao gồm các CSDL cực lớn cỡ gigabytes và thậm chí terabytes lưu trữ các dữ liệu kinh doanh ví dụ như dữ liệu thông tin khách hàng, dữ liệu bán hàng, dữ liệu các tài khoản, ... Nhiều hệ quản trị CSDL mạnh với các công cụ phong phú và thuận tiện đã giúp con người khai thác có hiệu quả nguồn tài nguyên dữ liệu. Mô hình CSDL quan hệ và ngôn ngữ (SQL) đã có vai trò hết sức quan trọng trong việc tổ chức và khai thác CSDL.

Tuy nhiên bên cạnh chức năng khai thác dữ liệu có tính chất tác nghiệp, sự thành công trong công việc không còn là năng suất của các hệ thống thông tin nữa mà là tính linh hoạt và sẵn sàng đáp ứng những yêu cầu trong thực tế, CSDL cần đem lại những "tri thức" hơn là chính những dữ liệu trong đó. Lúc này, các mô hình CSDL truyền thống và ngôn ngữ SQL đã cho thấy không có khả năng thực hiện công việc này. Để lấy thông tin có tính "tri thức" trong khối dữ liệu khổng lồ này, người ta đã tìm ra những kỹ thuật có khả năng hợp nhất các dữ liệu từ các hệ thống giao dịch khác nhau, chuyển đổi thành một tập hợp các CSDL ổn định, có chất lượng được sử dụng chỉ cho riêng một vài mục đích nào đó. Các kỹ thuật đó gọi chung là kỹ thuật tạo kho dữ liệu (data warehousing) và môi trường các dữ liệu có được gọi là các kho dữ liệu (data warehouse).

Đồng thời, Công nghệ khai phá dữ liệu (data mining) ra đời đáp ứng những đòi hỏi trong khoa học cũng như trong hoạt động thực tiễn. Đây chính là một ứng dụng chính để khai phá kho dữ liệu nhằm phát hiện tri thức (Knowledge Discovery) phục vụ công tác quản lý, kinh doanh,….

1.2. Tổng quan về kỹ thuật phát hiện tri thức và khai phá dữ liệu

Chúng ta có thể xem tri thức như là các thông tin tích hợp, bao gồm các sự kiện và các mối quan hệ giữa chúng. Các mối quan hệ này có thể được hiểu ra, có thể được phát hiện, hoặc có thể được học. Nói cách khác, tri thức có thể được coi là dữ liệu có độ trừu tượng và tổ chức cao.

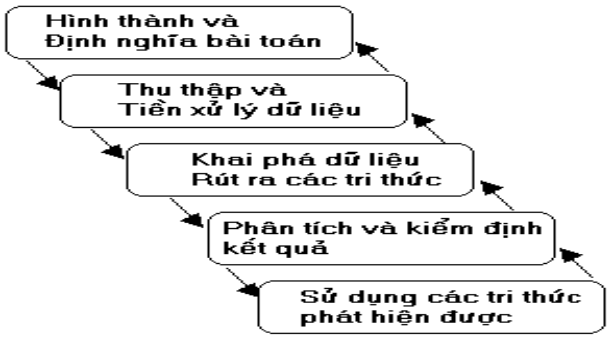
Phát hiện tri thức trong các cơ sở dữ liệu là một qui trình nhận biết các mẫu hoặc các mô hình trong dữ liệu với các tính năng: hợp thức, mới, khả ích, và có thể hiểu được. Còn khai phá dữ liệu là một bước trong qui trình phát hiện tri thức gồm có các thuật toán khai thác dữ liệu chuyên dùng dưới một số qui định về hiệu quả tính toán chấp nhận được để tìm ra các mẫu hoặc các mô hình trong dữ liệu. Nói một cách khác, mục đích của phát hiện tri thức và khai phá dữ liệu chính là tìm ra các mẫu và/hoặc các mô hình đang tồn tại trong các cơ sở dữ liệu nhưng vẫn còn bị che khuất bởi hàng núi dữ liệu.

Định nghĩa: Phát hiện tri thức và khai phá dữ liệu (KDD: Knowledge Discovery and Data Mining) là quá trình không tầm thường nhận ra những mẫu có giá trị, mới, hữu ích tiềm năng và hiểu được trong dữ liệu [7].

Còn các nhà thống kê thì xem Khai phá dữ liệu như là một qui trình phân tích được thiết kế để thăm dò một lượng cực lớn các dữ liệu nhằm phát hiện ra các mẫu thích hợp và/hoặc các mối quan hệ mang tính hệ thống giữa các biến và sau đó sẽ hợp thức hoá các kết quả tìm được bằng cách áp dụng các mẫu đã phát hiện được cho các tập con mới của dữ liệu. Qui trình này bao gồm ba giai đoạn cơ bản: thăm dò, xây dựng mô hình hoặc định nghĩa mẫu, hợp thức/kiểm chứng.

1.2.1. Qui trình khai phá dữ liệu và phát hiện tri thức.

Qui trình phát hiện tri thức được mô tả tóm tắt trên Hình 1:



Hình 1.1. Quy trình phát hiện tri thức

Bước thứ nhất: Hình thành, xác định và định nghĩa bài toán. Là tìm hiểu lĩnh vực ứng dụng từ đó hình thành bài toán, xác định các nhiệm vụ cần phải hoàn thành. Bước này sẽ quyết định cho việc rút ra được các tri thức hữu ích và cho phép chọn các phương pháp khai phá dữ liệu thích hợp với mục đích ứng dụng và bản chất của dữ liệu.

Bước thứ hai: Thu thập và tiền xử lý dữ liệu. Là thu thập và xử lý thô, còn được gọi là tiền xử lý dữ liệu nhằm loại bỏ nhiễu, xử lý việc thiếu dữ liệu, biến đổi dữ liệu và rút gọn dữ liệu nếu cần thiết, bước này thường chiếm nhiều thời gian nhất trong toàn bộ qui trình phát hiện tri thức.

Bước thứ ba: Khai phá dữ liệu, rút ra các tri thức. Là trích ra các mẫu và/hoặc các mô hình ẩn dưới các dữ liệu. Giai đoạn này rất quan trọng, bao gồm các công đoạn như: chức năng, nhiệm vụ và mục đích của khai phá dữ liệu, dùng phương pháp khai phá nào?

Bước thứ tư: Sử dụng các tri thức phát hiện được. Là hiểu tri thức đã tìm được, đặc biệt là làm sáng tỏ các mô tả và dự đoán. Các bước trên có thể lặp đi lặp lại một số lần, kết quả thu được có thể được lấy trung bình trên tất cả các lần thực hiện.

Tóm lại: KDD là một quá trình chiết xuất ra tri thức từ kho dữ liệu mà trong đó khai phá dữ liệu là công đoạn quan trọng nhất.

1.2.2. Các lĩnh vực liên quan đến khai phá dữ liệu và phát hiện tri thức

Khai phá dữ liệu và phát hiện tri thức liên quan đến nhiều ngành, nhiều lĩnh vực: thống kê, trí tuệ nhân tạo, cơ sở dữ liệu, thuật toán học, tính toán song song và tốc độ cao, thu thập tri thức cho các hệ chuyên gia, quan sát dữ liệu... Đặc biệt Phát hiện tri thức và khai phá dữ liệu rất gần gũi với lĩnh vực thống kê, sử dụng các phương pháp thống kê để mô hình dữ liệu và phát hiện các mẫu, luật... Kho dữ liệu (Data Warehousing) và các công cụ phân tích trực tuyến (OLAP) cũng liên quan rất chặt chẽ với Phát hiện tri thức và khai phá dữ liệu.

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