ĐẠI HỌC QUỐC GIA HÀ NỘI
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

---- 🕮 ----

ĐINH CHUNG DŨNG

NGHIÊN CỨU VÀ ÁP DỤNG KỸ THUẬT KHAI PHÁ DỮ LIỆU TRÊN BỘ DỮ LIỆU SINH VIÊN ĐẠI HỌC PHỤC VỤ CÔNG TÁC CỐ VẤN HỌC TẬP

LUẬN VĂN THẠC SĨ CÔNG NGHỆ THÔNG TIN

HÀ NỘI, 2017

ĐẠI HỌC QUỐC GIA HÀ NỘI

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

ĐINH CHUNG DŨNG

NGHIÊN CỨU VÀ ÁP DỤNG KỸ THUẬT KHAI PHÁ DỮ LIỆU TRÊN BỘ DỮ LIỆU SINH VIÊN ĐẠI HỌC PHỤC VỤ CÔNG TÁC CỐ VẤN HỌC TẬP

Ngành : Công nghệ thông tin
Chuyên ngành : Truyền dữ liệu và mạng máy tính
Mã số : Chuyên ngành đào tạo thí điểm

LUẬN VĂN THẠC SĨ CÔNG NGHỆ THÔNG TIN

HƯỚNG DẪN KHOA HỌC: TS. NGUYỄN TRUNG TUẤN

HÀ NỘI, 2017

LỜI CẢM ƠN

Tác giả luận văn xin chân thành cảm ơn đến người hướng dẫn khoa học là TS. Nguyễn Trung Tuấn, Viện Công nghệ Thông tin Kinh tế, Trường Đại học Kinh tế Quốc dân. Thầy đã dành nhiều thời gian và tâm huyết để hướng dẫn và giúp đỡ tác giả hoàn thành luận văn này. Tác giả cũng xin cảm ơn các Thầy, Cô trong Khoa Công nghệ Thông tin, Trường Đại học Công nghệ, Đại học Quốc gia Hà Nội đã tạo điều kiện thuận lợi, giúp đỡ và có những đóng góp quý báu trong thời gian nghiên cứu và hoàn thành luận văn của tác giả.

Xin chân thành cảm ơn gia đình, bạn bè và đồng nghiệp đã giúp đỡ, động viên tác giả trong suốt thời gian nghiên cứu luận văn.

Hà Nội, Ngày……tháng….. năm 2017

Đinh Chung Dũng

Lời cam đoan

Tôi xin cam đoan đây là công trình nghiên cứu của tôi dưới sự hướng dẫn khoa học của TS Nguyễn Trung Tuấn. Các số liệu và kết quả nghiên cứu, công bố trong luận văn là trung thực và chưa từng được ai công bố trong bất kỳ công trình nào khác.

Hà Nội, Ngày……tháng….. năm 2017

Đinh Chung Dũng

MỤC LỤC

DANH MỤC CÁC THUẬT NGỮ VIẾT TẮT

Từ viết tắt	Tiếng Anh	Tiếng Việt
BI	Business Intelligence	Kinh doanh thông minh/trí tuệ doanh nghiệp
BIDS	Business Intelligence Development Studio	Bộ công cụ phân tích dữ liệu trong MicroSoft SQL Server
DA/PA	Data/Pattern analysis	Phân tích dữ liệu/mẫu
DBMS	Database Management System	Hệ quản trị cơ sở dữ liệu
KDD	Knowledge Discovery and Data Mining	Phát hiện tri thức và Khai phá dữ liệu
KE	Knowledge Extraction	Trích chọn tri thức
ML	Machine Learning	Học máy
SQL	Structured Query Language	Ngôn ngữ truy vấn cấu trúc

DANH MỤC CÁC HÌNH

MỞ ĐẦU

1. Lý do lựa chọn đề tài

Khai phá dữ liệu và phát hiện tri thức đang là lĩnh vực được các nhà khoa học quan tâm nghiên cứu trong nhiều năm gần đây. Ứng dụng khai phá dữ liệu được thực hiện trong nhiều lĩnh vực khác nhau như giáo dục, y tế, tài chính, ngân hàng, kinh doanh… Đặc biệt, trong thời gian gần đây, khai phá dữ liệu và phát hiện tri thức trong lĩnh vực giáo dục đang được quan tâm nghiên cứu. Đối với bậc giáo dục Đại học hiện nay, sinh viên đang học tập tại các trường Đại học theo hình thức đào tạo tín chỉ. Đối với hình thức đào tạo này yêu cầu sinh viên phải có sự chủ động cao, có nhiều sự lựa chọn mềm dẻo các môn học trong chuyên ngành đào tạo. Sinh viên sẽ phải tự mình phân bổ các môn học cho từng kỳ sao cho đủ số tín chỉ theo quy chế đào tạo, sinh viên có thể học nhanh để ra trường sớm hoặc đúng hạn với số điểm cao. Trên thực tế đã có rất nhiều trường hợp thời gian học đã hết nhưng các em vẫn chưa hoàn thành đủ tín chỉ, còn nợ môn chuyên ngành. Các sinh viên chưa quen và gặp rất nhiều khó khăn trong định hướng học tập, làm ảnh hưởng đến quá trình học tập của mình cũng như ảnh hưởng đến kết quả đào tạo của nhà trường. Chính vì vậy công tác cố vấn học tập cho sinh viên đã được đặt ra là một công việc quan trọng trong hình thức đào tạo theo tín chỉ. Đây cũng là bài toán được đặt ra cho lĩnh vực khai phá dữ liệu khi có số liệu lớn về sinh viên và quá trình học tập của sinh viên trong nhà trường nhằm trợ giúp cho cố vấn học tập đạt được hiệu quả cao hơn.

Hiện nay tôi đang công tác tại Trường Đại học Kinh tế quốc dân, trước những thực trạng đang tồn tại ở nơi làm việc cùng với lĩnh vực tôi đang theo học, được sự đồng ý của TS. Nguyễn Trung Tuấn tôi chọn đề tài luận văn: "***Nghiên cứu và áp dụng kỹ thuật khai phá dữ liệu trên bộ dữ liệu sinh viên đại học phục vụ công tác cố vấn học tập***", luận văn góp phần vào việc giải quyết các vấn đề hết sức cấp bách và cần thiết trong thực tế.

2. Mục tiêu nghiên cứu của luận văn

Mục tiêu nghiên cứu của luận văn là để hiểu các kỹ thuật khai phá dữ liệu và phát hiện tri thức cơ bản, tập trung chủ yếu vào hai kỹ thuật chính là kỹ thuật khai phá luật kết hợp và cây quyết định. Đây là kỹ thuật đã có nhiều nhà khoa học nghiên cứu và có nhiều đóng góp vào thực tiễn. Hiểu các quy chế, quy định, thông tư hướng dẫn về triển khai thực hiện đào tạo đại học chính quy theo hệ thống tín chỉ, các văn bản liên quan đến quy định về cố vấn học tập, chương trình đào tạo chính quy theo học chế tín chỉ thuộc các chuyên ngành của Trường Đại học Kinh

tế Quốc dân. Đặc biệt tập trung vào các vấn đề cố vấn học tập cho sinh viên trong quá trình học tập tại trường. Kết quả đạt được là phát hiện một số luật trong cố vấn học tập thông qua bộ dữ liệu quản lý thông tin sinh viên hiện tại của Trường Đại học Kinh tế Quốc dân bằng việc áp dụng kỹ thuật khai phá luật kết hợp và cây quyết định với sự trợ giúp của các công cụ có sẵn.

3. Đối tượng, phạm vi nghiên cứu

Nghiên cứu các vấn đề cơ bản của khai phá dữ liệu và phát hiện tri thức; kỹ thuật khai phá luật kết hợp, cây quyết định trong phát hiện tri thức và khai phá dữ liệu; bài toán cố vấn học tập tại trường Đại học Kinh tế quốc dân. Các nội dung nghiên cứu sẽ được thực nghiệm trên bộ dữ liệu sinh viên đại học chính quy của trường Đại học Kinh tế quốc dân.

4. Phương pháp nghiên cứu

Tổng hợp các vấn đề lý thuyết liên quan từ các nguồn là giáo trình, bài giảng, chuyên đề, luận văn, luận án, internet và tìm hiểu văn bản quy định tại trường Đại học Kinh tế Quốc dân về các vấn đề liên quan đến luận văn. Cùng với đó là phương pháp nghiên cứu thực nghiệm. Sử dụng bộ dữ liệu thực tế về thông tin và quá trình học của một khóa sinh viên chính quy đã ra trường, xử lý dữ liệu trên những công cụ quản trị cơ sở dữ liệu và công cụ khai phá dữ liệu của Microsoft SQL Server. So sánh các kết quả thu được từ các mô hình khai phá để rút ra các kết luận và quy trình sử dụng kỹ thuật khai phá dữ liệu cho bài toán cố vấn học tập.

5. Những đóng góp của luận văn

Thực hiện mục tiêu nghiên cứu đã nêu ở trên, ngoài việc tổng hợp và tổng quan các kiến thức liên quan cần thiết, luận văn đưa ra các đóng góp chính sau đây:

+ Đề xuất quy trình xử lý dữ liệu cho các bài toán cố vấn học tập tại trường Đại học kinh tế quốc dân

+ Thực nghiệm với bộ dữ liệu thực tế và đánh giá các kết quả đã tìm được từ các kỹ thuật khai phá dữ liệu cho các bài toán cố vấn học tập đã nêu.

6. Kết cấu của luận văn

Luận văn được trình bày trong ba chương chính:

Chương 1. Tổng quan về phát hiện tri thức và khai phá dữ liệu

Trong chương này sẽ trình bày những vấn đề cơ bản về phát hiện tri thức và khai phá dữ liệu, bao gồm những nội dung cơ bản: tổng quan về khai phá dữ liệu và phát hiện tri thức; ứng dụng của khai phá dữ liệu và phát hiện tri thức; các phương pháp và kỹ thuật khai phá dữ liệu và phát hiện tri thức.

Chương 2. Bài toán cố vấn học tập và đặc điểm bộ dữ liệu sinh viên tại trường Đại học Kinh tế Quốc dân

Nội dung của chương này sẽ trình bày những vấn đề về cố vấn học tập trong đào tạo đại học chính quy theo hình thức tín chỉ, những vấn đề gặp phải trong quá trình cố vấn học tập. Từ đó phân tích và hình thành bài toán cần giải quyết trong công tác cố vấn học tập tại trường Đại học Kinh tế Quốc dân. Giới thiệu và mô tả đặc điểm của bộ dữ liệu sinh viên chính quy đã thu thập được trường Đại học Kinh tế quốc dân để phục vụ cho quá trình thực nghiệm; mô tả về các bộ dữ liệu con được trích rút dữ liệu bộ dữ liệu lớn phục vụ cho các mục đích phân tích khác nhau theo yêu cầu của bài toán cố vấn học tập.

Chương 3. Ứng dụng thử nghiệm khai phá dữ liệu sinh viên phục vụ cố vấn học tập tại trường Đại học Kinh tế Quốc dân

Chương này sẽ giới thiệu về một số công cụ khai phá dữ liệu và phát hiện tri thức thông dụng và công cụ BIDS của Microsoft SQL Server 2008. Quy trình thực nghiệm khai phá và phát hiện tri thức với bài toán cố vấn học tập. Trình bày và đánh giá các kết quả khai phá dữ liệu trên 02 bài toán cố vấn học tập: Tư vấn lựa chọn môn học theo tổ hợp lựa chọn từng ngành; Phân lớp, dự đoán sinh viên có ra trường đúng thời hạn hay không.

Ngoài ra, phần Mở đầu của luận văn sẽ giới thiệu chung về những nội dung và phương pháp thực hiện nghiên cứu đề tài luận văn. Phần Kết luận của luận văn sẽ trình bày về tóm tắt về những kết quả đã đạt được, những hạn chế và hướng nghiên cứu tiếp theo của đề tài luận văn.

CHƯƠNG 1

TỔNG QUAN VỀ PHÁT HIỆN TRI THỨC VÀ KHAI PHÁ DỮ LIỆU

1.1 Giới thiệu chương

Mục tiêu của chương này là nhằm trình bày về cơ sở lý luận, lý thuyết nền tảng phục vụ cho những nghiên cứu sâu hơn trong luận văn. Nội dung chính của chương bao gồm những khái niệm, các kỹ thuật, ứng dụng và những vấn đề cần nghiên cứu trong phát hiện tri thức và khai phá dữ liệu. Chương 1 được bố cục gồm có 9 mục, mục kế tiếp sẽ đề cập đến những vấn đề cơ bản về phát hiện tri thức và khai phá dữ liệu. Mục 3 trong chương sẽ tóm tắt quá trình phát hiện tri thức và khai phá dữ liệu. Mục 4 trình bày các phương pháp khai phá dữ liệu. Mục 5 sẽ trình bày về các vấn đề cần nghiên cứu của phát hiện tri thức và khai phá dữ liệu. Mục 6 là các lĩnh vực ứng dụng của phát hiện tri thức và khai phá dữ liệu. Mục 7 là kỹ thuật khai phá luật kết hợp. Mục 8 tóm tắt lý thuyết cây quyết định. Cuối cùng là tổng kết những vấn đề đã được tác giả thể hiện trong chương.

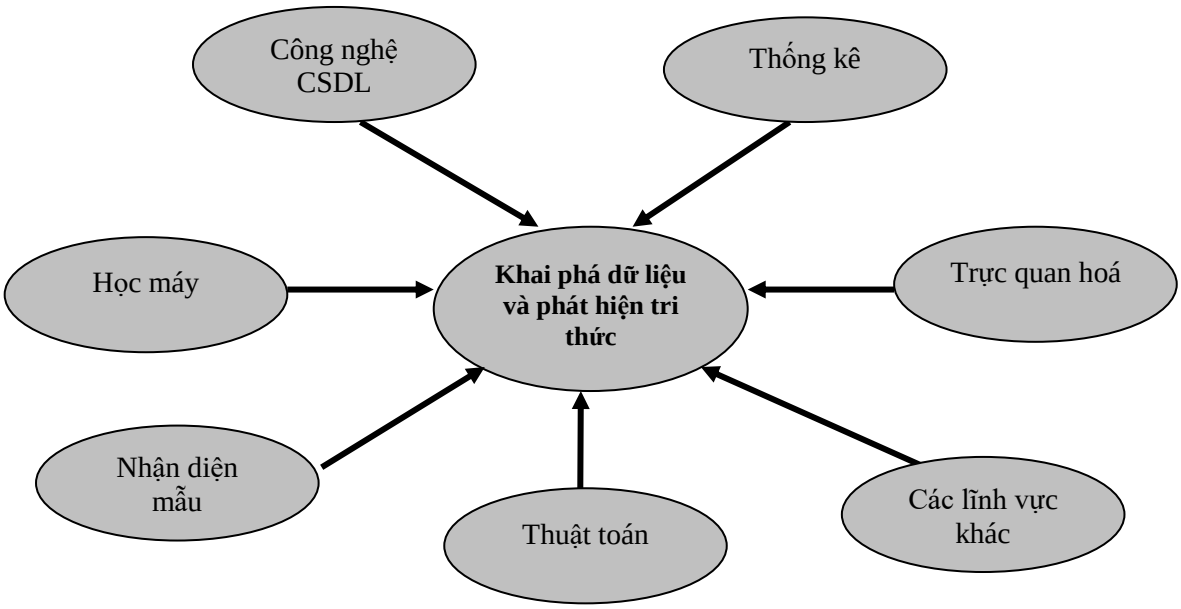
1.2 Tổng quan về phát hiện tri thức và khai phá dữ liệu

Cùng với sự phát triển của các ngành khoa học, các dữ liệu con người thu nhận, lưu trữ thông qua các hoạt động kinh tế - xã hội, các hoạt động nghiên cứu khoa học ngày một lớn, chúng được lưu trữ trên các hệ thống máy tính với dung lượng lên đến hàng *terabyte*, thậm chí đến hàng *petabyte*. Tuy nhiên, việc hiểu và sử dụng hết được những dữ liệu đó đối với con người rất khó khăn. Trước thực tế như vậy, một hướng nghiên cứu mới về phát hiện tri thức và khai phá dữ liệu đã hình thành và phát triển nhanh chóng trong gần 20 năm qua. Tác giả sẽ trình bày lại một số khái niệm liên quan đến lĩnh vực Phát hiện tri thức và Khai phá dữ liệu *(KDD - Knowledge Discovery and Data mining)* được đề cập trong [1], [4], [9], [10], [11], [12], nhằm hệ thống hóa những kiến thức nền tảng về lĩnh vực này. Trong thực tế, Phát hiện tri thức và Khai phá dữ liệu còn có thể được sử dụng với cụm từ Khai phá dữ liệu và Phát hiện tri thức.

Dữ liệu *(data)* là số liệu về các hiện tượng, sự vật mà người ta thu thập được thông qua quan sát, khảo sát trực tiếp hoặc thông qua các thiết bị hỗ trợ, chúng có thể là các con số, các chuỗi ký tự, các biểu tượng hoặc các đối tượng có ý nghĩa nhất định. Dữ liệu có thể được đưa vào các chương trình máy tính theo một định dạng nào đó. Thông tin (*information*) là các dữ liệu đã qua một quá trình xử lý, chắt lọc và thường mang những ý nghĩa nhất định đối với những đối tượng tiếp nhận thông tin, người ta cũng có thể coi thông tin là những dữ liệu đã được

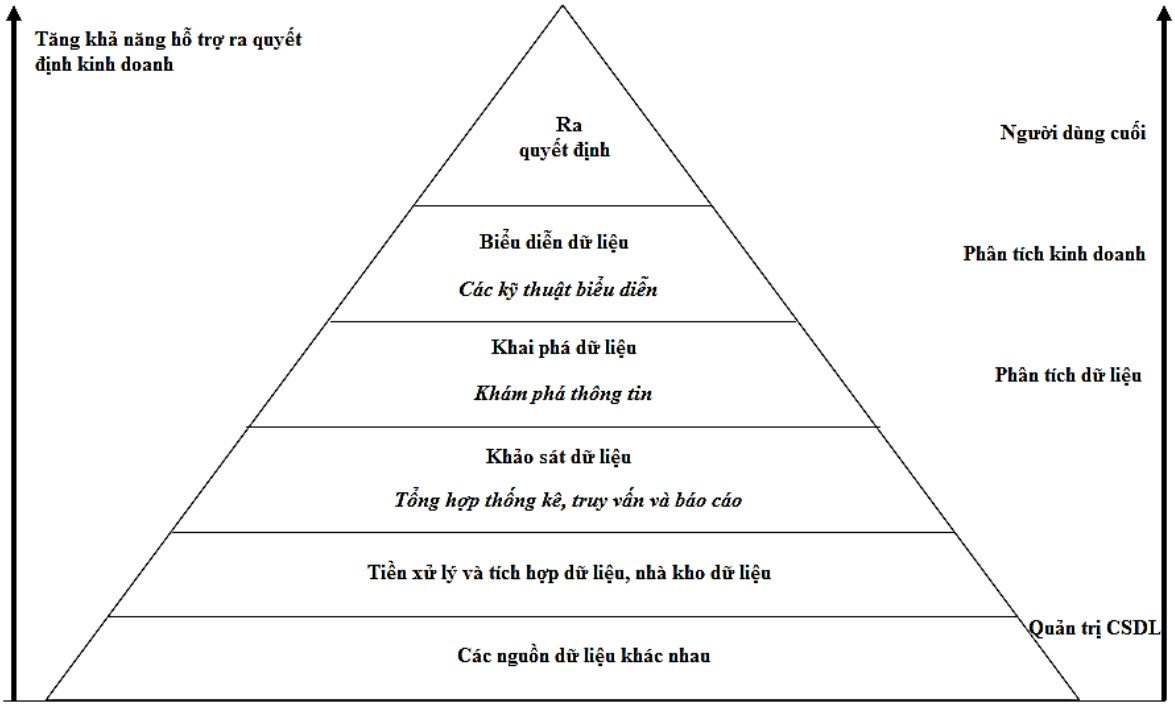
phiên dịch theo một phương pháp nào đó. Thông tin của quá trình xử lý này có thể lại trở thành dữ liệu cho một quá trình xử lý khác. Tri thức *(knowledge)* là các thông tin được tích hợp bao gồm cả các cơ sở lập luận và những vấn đề liên quan, được nhận biết, khám phá, phản ánh trong trí óc và tinh thần. Tri thức còn được hiểu đó là dữ liệu đã được trừu tượng hoá và tổng quát hoá ở mức cao. Tri thức có đặc điểm là có thể được tái tạo, phát triển qua các quá trình học, suy luận và vận dụng, tri thức sẽ không mất đi trong quá trình sử dụng mà ngược lại nó càng gia tăng và phát triển lên một mức độ mới nếu càng được sử dụng nhiều.

Phát hiện tri thức và khai phá dữ liệu là quá trình tự động trích rút các tri thức (*knowledge*) hoặc các mẫu (*pattern*), mô hình (*model*) có đặc điểm không tầm thường, ẩn, chưa biết trước, có khả năng sử dụng và hiểu được từ khối lượng lớn dữ liệu [4]. Phát hiện tri thức và khai phá dữ liệu là một lĩnh vực phát triển rất nhanh chóng, là lĩnh vực giao thoa giữa nhiều lĩnh vực liên quan như: công nghệ cơ sở dữ liệu, thống kê, học máy, thuật toán học và các lĩnh vực liên quan khác nhằm trích rút ra những tri thức hữu ích từ những tập dữ liệu rất lớn. Người ta cũng có thể sử dụng những tên khác cho khai phá dữ liệu và khám phá tri thức như: khám phá tri thức trong cơ sở dữ liệu (*Knowledge discovery in databases - KDD*), trích chọn tri thức (*Knowledge extraction - KE*), phân tích dữ liệu hay mẫu (*Data/pattern analysis - DA/PA*) hay kinh doanh thông minh hoặc tri thức doanh nghiệp (*Business Intelligence - BI*) [4]...



Hình 1.1 Mối quan hệ của KDD với các lĩnh vực khác [4]

Khai phá dữ liệu *(Data mining)* là một khâu trong quá trình khám phá tri thức mà trong đó ta có thể áp dụng những thuật toán khai phá dữ liệu với những giới hạn có thể chấp nhận được về độ phức tạp tính toán để tìm ra những mẫu hoặc mô hình trong dữ liệu [4]. Khai phá dữ liệu có hai chức năng chính là: mô tả dữ liệu và dự báo dữ liệu, trong đó mô tả dữ liệu tập trung vào tìm kiếm các đặc tính, đặc trưng của dữ liệu, còn dự báo dữ liệu tập trung vào việc phân tích, suy diễn dữ liệu quá khứ, hiện tại để dự báo giá trị dữ liệu tương lai. Như vậy mục đích của phát hiện tri thức và khai phá dữ liệu là để tìm ra những mẫu và/hoặc những mô hình tồn tại trong cơ sở dữ liệu mà chúng có thể đang ẩn trong khối dữ liệu rất lớn.



Hình 1.2 Mối quan hệ của KDD và kinh doanh thông minh [4]

Phát hiện tri thức và khai phá dữ liệu được ứng dụng trong nhiều lĩnh vực khác nhau [4]. Trong phân tích dữ liệu và hỗ trợ quyết định, phát hiện tri thức và khai phá dữ liệu được ứng dụng vào quản trị kinh doanh và phân tích thị trường (còn được coi là các lĩnh vực kinh doanh thông minh hay trí tuệ doanh nghiệp - Hình 1.2) như: định hướng thị trường, quản trị quan hệ khách hàng (*Customer Relation Management - CRM*), phân tích giỏ hàng, phân mảng thị trường và kinh doanh đa chiều; quản trị và phân tích rủi ro: dự báo, duy trì khách hàng, kiểm soát chất lượng, phân tích cạnh tranh...; phát hiện gian lận và dò tìm những mẫu không

bình thường, phân tích cá biệt (*outlier*). Trong các lĩnh vực khác, người ta áp dụng vào khai phá dữ liệu văn bản (bản tin, thư điện tử, tài liệu), khai phá dữ liệu Web, khai phá dữ liệu theo luồng và các dữ liệu sinh học...

Theo [4], người ta thường sử dụng một số tiêu chí sau để phân loại mức độ hấp dẫn của kết quả: Tính căn cứ *(Evidence)* chỉ ra ý nghĩa của kết quả tìm kiếm được và thường đo bằng các tiêu chí thống kê. Độ dư thừa *(Redundancy)* để chỉ sự tương tự của kết quả tìm được so với các kết quả tìm kiếm khác và các độ đo xác định mức độ tương tự của một kết quả với các kết quả khác. Tính hữu dụng *(Usefulness)* để chỉ mối quan hệ giữa kết quả tìm được và mục tiêu của người dùng. Tính mới *(Novelty)* để chỉ ra sự khác biệt của kết quả với những tri thức có trước của người sử dụng hay của hệ thống, người ta còn gọi đó là tính bất ngờ. Tính đơn giản *(Simplicity)* để chỉ độ phức tạp về cú pháp biểu diễn kết quả tìm kiếm và khả năng tổng quát hoá. Ta cụ thể hoá một số các thuật ngữ như sau:

- *Dữ liệu (Data):* là một tập hợp các thể hiện của các đối tượng hoặc tập hợp các giá trị của các biến (ví dụ là các bản ghi trong cơ sở dữ liệu).
- *Mẫu (Pattern):* là mô tả một tập con của không gian kết quả hoặc không gian dữ liệu, các mẫu và mô hình thường được biểu diễn thông qua một hàm $F(v_1, v_2, ...,v_n)$ trong đó v_i là các tham số, các tham số này có giá trị là các tập con của dữ liệu.
- *Tiến trình (Process)*: thông thường trong tiến trình KDD là quá trình đa bước, bao gồm chuẩn bị và tiền xử lý dữ liệu, tìm kiếm hình mẫu, đánh giá tri thức và tinh chỉnh, được lặp đi lặp lại kèm theo sự sửa đổi nào đó, quá trình này có thể được thực hiện một cách tự động hoặc bán tự động.
- *Hợp lệ (Validity)*: Những mẫu hoặc mô hình được khám phá từ một tập dữ liệu huấn luyện phải đúng trên tập dữ liệu mới với một mức độ chắc chắn nào đó, mức độ chắc chắn này xác định khả năng đúng đắn của mẫu hoặc mô hình tìm được, thông thường người ta xác định một ngưỡng tối thiểu cho độ đo chắc chắn để lọc ra các kết quả phù hợp.
- *Mới (Novelty)*: Các mẫu tìm được phải có tính mới hoặc bất ngờ (ít nhất là đối với hệ thống). Tính mới có thể được đo đối với sự thay đổi trong dữ liệu (bằng việc so sánh các giá trị hiện tại với các giá trị trước hoặc các giá trị mong muốn) hoặc tri thức (kết quả tìm kiếm mới có quan hệ như thế nào đối với kết quả cũ).

- *Hữu dụng tiềm năng (Potentially Useful)*: Các mẫu có thể có khả năng hữu dụng, nó thể hiện các kết quả tìm được có phù hợp với mục tiêu của người dùng không. Tính hữu dụng thường được đo bằng các hàm tiện ích là ánh xạ từ không gian kết quả đến không gian mục tiêu với một độ đo nào đó.
- *Khả năng có thể hiểu được (Understandability)*: Mục đích của KDD là tạo ra các mẫu mà con người có khả năng hiểu được để có thể nắm bắt tốt hơn về dữ liệu. Điều này rất khó xác định một cách chính xác do vậy người ta sử dụng một thông số khác là độ đo tính đơn giản *(Simplicity)*. Có nhiều độ đo tính đơn giản được sử dụng, từ việc đo về cú pháp (ví dụ là kích thước của mẫu) đến ngữ nghĩa (ví dụ như con người có dễ nhận thức được không trong một số tình huống).

Một độ đo khác rất quan trọng được gọi là mức độ hấp dẫn *(Interestingness)* thường là độ đo tổng thể kết hợp các độ đo trên của các mẫu hoặc mô hình tìm được, tuỳ theo mục đích của người sử dụng mà mỗi độ đo riêng biệt được gán một trọng số nhất định khi kết hợp trong độ đo tổng thể.

1.3 Quá trình phát hiện tri thức và khai phá dữ liệu

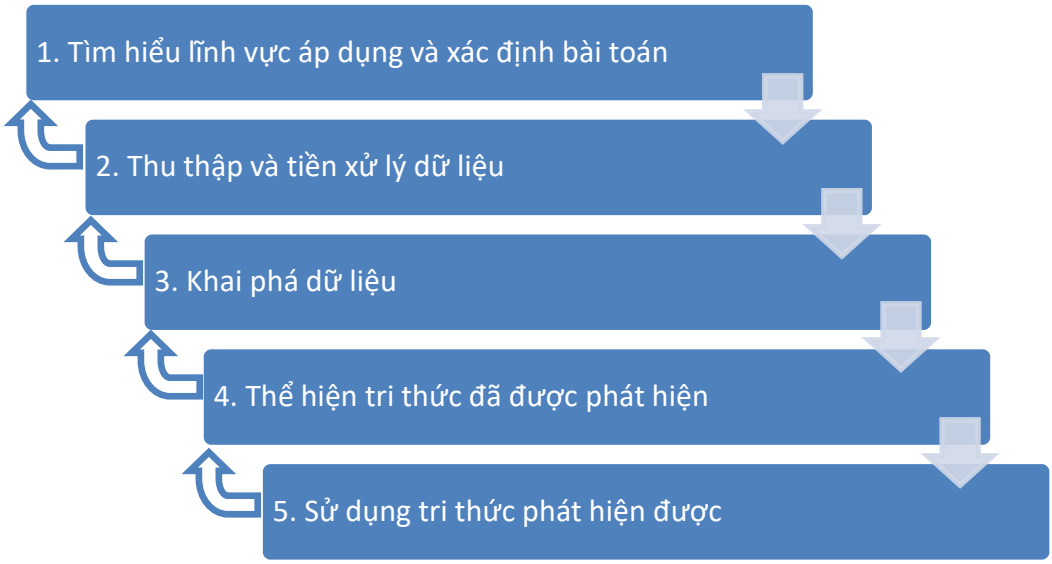
Theo [1], quá trình phát hiện tri thức và khai phá dữ liệu được thực hiện thông qua nhiều bước và được lặp đi lặp lại. Dưới đây là tóm tắt lại những bước cơ bản trong quá trình này đã được mô tả chi tiết trong [1].

Bước 1: **Tìm hiểu lĩnh vực áp dụng và xác định bài toán**, còn được gọi là tìm hiểu tri thức lĩnh vực. Đây là bước tiên quyết để có thể trích rút ra được những tri thức hữu dụng và lựa chọn được các phương pháp khai phá dữ liệu thích hợp cho **bước 3** tuỳ thuộc vào mục đích sử dụng và bản chất của dữ liệu.

Bước 2: **Thu thập và tiền xử lý dữ liệu**: Lựa chọn các nguồn dữ liệu, xử lý nhiễu hoặc loại những dữ liệu dư thừa, xử lý dữ liệu khiếm khuyết, chuyển đổi dữ liệu và rút gọn dữ liệu... Bước này thường chiếm phần lớn thời gian trong cả tiến trình KDD.

Bước 3: **Khai phá dữ liệu**: Tìm kiếm các mẫu/mô hình ẩn chứa trong dữ liệu bằng các thuật toán khai phá dữ liệu nào đó phù hợp với từng loại dữ liệu đầu vào. Các lớp bài toán quan trọng của khai phá dữ liệu là mô hình hoá dự báo như phân lớp và hồi qui; phân đoạn và phân cụm; mô hình hoá sự phụ thuộc như các mô hình đồ thị hoặc dự tính mật độ; tổng quát hoá như

tìm mối quan hệ giữa các trường, sự liên kết, biểu diễn trực quan; mô hình hoá hoặc phát hiện sự thay đổi và sự chênh lệch trong dữ liệu và tri thức.



Hình 1.3 Qui trình 5 bước khai phá dữ liệu

Bước 4: **Thể hiện tri thức đã được phát hiện**: Thể hiện các tri thức đã được phát hiện theo các phương pháp mô tả và dự báo, đây là hai đích cơ bản nhất của các hệ thống phát hiện tri thức. Các thí nghiệm chỉ ra rằng các mẫu hoặc mô hình phát hiện được từ dữ liệu thường không được quan tâm hoặc trực tiếp sử dụng ngay và tiến trình KDD cần thiết phải lặp lại với sự đánh giá tri thức được phát hiện. Để đánh giá các luật thu được, người ta thường chia dữ liệu ra thành hai tập, huấn luyện trên một tập và kiểm tra trên tập kia. Quá trình này có thể lặp đi lặp lại nhiều lần với những cách phân chia khác nhau, kết quả trung bình có thể dự tính được độ mạnh của các luật. Một phương pháp đánh giá thường được sử dụng có tên gọi *m-fold cross validation*. Với phương pháp này, người ta phân chia tập dữ liệu thành *m* tập con một cách ngẫu nhiên và có số lượng phần tử tương đối đều nhau, sau đó sử dụng 1 tập làm tập kiểm tra và *m-1* tập con còn lại làm tập huấn luyện để thực hiện thuật toán, quá trình này được thực hiện *m* lần cho *m* tập con khác nhau, kết quả cuối cùng sẽ được tính là trung bình cộng của *m* lần thực hiện thuật toán. Người ta thường chọn *m=10*, như vậy phép thử này sẽ là *10-fold cross validation* (xác nhận chéo 10 lần).

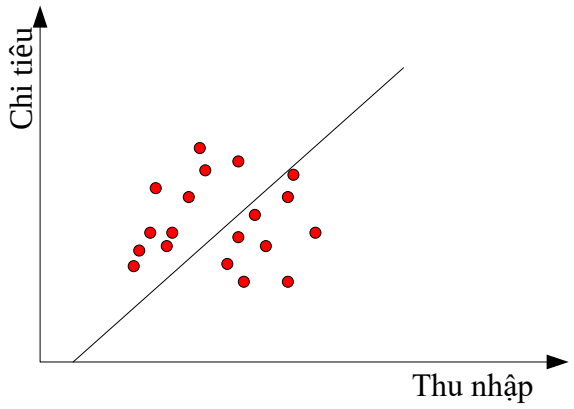
Bước 5: **Sử dụng tri thức phát hiện được**: đây là bước cuối cùng trong quá trình KDD. Trong một số trường hợp, các tri thức có thể được sử dụng mà không cần đưa vào trong hệ thống máy tính. Trong một số trường hợp khác, người sử dụng mong muốn các tri thức đã phát hiện có thể đưa vào máy tính để một số chương trình khai thác được ngay. Việc đưa các kết quả của KDD vào sử dụng trong thực tế là đích cao nhất của khám phá tri thức.

Không gian các mẫu thường rất lớn và việc liệt kê các mẫu đòi hỏi một số phương pháp tìm kiếm trong không gian này. Các ràng buộc về khả năng tính toán sẽ xác định những giới hạn trong không gian con mà các thuật toán có thể thực hiện. Công việc của khai phá dữ liệu trong tiến trình KDD tập trung chủ yếu vào các công cụ được sử dụng để trích và liệt kê các mẫu từ dữ liệu. Phát hiện tri thức bao gồm đánh giá và thể hiện các mẫu để quyết định mẫu nào là tri thức, mẫu nào không là tri thức, cũng bao gồm việc lựa chọn các cách mã hoá, tiền xử lý, lấy mẫu và chiếu trên các thuộc tính trước khi thực hiện khai phá dữ liệu.

1.4 Các phương pháp khai phá dữ liệu

Trong [4] đã chỉ ra hai mục tiêu cơ bản của khai phá dữ liệu là nhằm dự báo và mô tả. Dự báo đòi hỏi sử dụng một số biến hoặc trường trong cơ sở dữ liệu để tìm giá trị chưa biết hoặc giá trị tương lai của các biến cần quan tâm. Mô tả tập trung vào việc tìm kiếm các mẫu thể hiện dữ liệu mà con người có thể hiểu được. Với các ứng dụng khai phá dữ liệu khác nhau thì mức độ quan trọng của việc dự báo hay mô tả cũng sẽ khác nhau. Ở đây ta sẽ tìm hiểu chi tiết các phương pháp khai phá dữ liệu thông dụng được nêu trong [4]:

- *Phân lớp* là việc xác định một hàm ánh xạ các mục dữ liệu vào một trong nhiều lớp đã được xác định trước. Ví dụ dưới đây thể hiện phân lớp theo hai chỉ tiêu là thu nhập và mức độ chi tiêu của các hộ gia đình.



Hình 1.4 Phân lớp dựa theo mức chi tiêu và thu nhập của các hộ gia đình

- *Hồi qui* là việc xác định một hàm ánh xạ một mục dữ liệu đến một giá trị dữ liệu thực của biến dự báo.
- *Phân cụm* là công việc mang tính mô tả thông thường, nó sẽ xác định tập hữu hạn các nhóm hoặc các cụm để mô tả dữ liệu. Các nhóm đó có thể là duy nhất và chi tiết hoặc có thể có các cách biểu diễn phong phú hơn như phân cấp hoặc chồng nhau.
- *Tổng quát hoá* bao gồm các phương pháp để tìm kiếm một mô tả ngắn gọn và tổng quát cho một tập con dữ liệu. Một ví dụ đơn giản là có thể lập bảng về trung vị và độ lệch chuẩn cho tất cả các trường. Các phương pháp phức tạp hơn bao gồm suy dẫn ra các luật tổng kết, các kỹ thuật thể hiện đa biến và phát hiện những mối quan hệ giữa các biến. Các kỹ thuật tổng kết thường sử dụng các phương pháp khai phá dữ liệu tương tác và tự động sinh ra báo cáo.
- *Mô hình hoá sự phụ thuộc* bao gồm việc tìm một mô hình mô tả những sự phụ thuộc cơ bản giữa các biến. Các mô hình phụ thuộc tồn tại ở hai mức: mức cấu trúc (thường trong dạng đồ hoạ) của các mô hình xác định các biến nào phụ thuộc cục bộ lẫn nhau, mức định lượng của các mô hình xác định độ mạnh của các phụ thuộc và thường sử dụng độ đo số.
- *Phát hiện thay đổi và chênh lệch* tập trung vào việc phát hiện những thay đổi đáng chú ý trên dữ liệu từ những giá trị được đo trước đó.
- *Biểu diễn mô hình* là phương pháp để mô tả những mẫu hoặc mô hình có thể được phát hiện. Nếu biểu diễn này bị hạn chế và có nhiều ràng buộc thì khi đó không thể tìm được mô hình đúng đắn cho dữ liệu. Do vậy điều quan trọng là người phân tích dữ liệu phải hiểu đầy đủ các giả thiết của mô hình biểu diễn và người thiết kế thuật toán phải xác định rõ các giả thiết của mô hình biểu diễn được thực hiện trong thuật toán cụ thể.
- *Đánh giá mô hình* dự tính khả năng đáp ứng của một mẫu hoặc mô hình và các tham số của nó với các tiêu chí kết quả của tiến trình *KDD*. Đánh giá độ chính xác dự báo (tính hợp lệ) được dựa trên kiểm tra chéo. Đánh giá chất lượng mô tả bao gồm độ chính xác dự báo, tính mới, tính hữu dụng và khả năng có thể hiểu được của mô hình. Cả các tiêu chí logic và thống kê có thể được sử dụng để đánh giá mô hình.

- *Phương pháp tìm kiếm* có hai bài toán là tìm kiếm tham số và tìm kiếm mô hình. Trong tìm kiếm tham số, thuật toán phải tìm ra các tham số mà nó tối ưu các tiêu chí đánh giá mô hình cho những dữ liệu được quan sát cho trước và một biểu diễn mô hình cố định. Tìm kiếm mô hình tập trung vào việc lặp trên phương pháp tìm kiếm tham số: biểu diễn mô hình được thay đổi vì vậy một họ các mô hình được xem xét. Với mỗi biểu diễn mô hình cụ thể, phương pháp tìm kiếm tham số được thực hiện để xác định chất lượng của mô hình. Thiết lập các phương pháp tìm kiếm mô hình thường theo xu hướng sử dụng các kỹ thuật tìm kiếm gần đúng khi không gian mô hình rất lớn.

1.5 Các vấn đề cần nghiên cứu của phát hiện tri thức và khai phá dữ liệu

Phát hiện tri thức và khai phá dữ liệu là một lĩnh vực mới và đang được nghiên cứu, ứng dụng một cách nhanh chóng, mạnh mẽ, tuy nhiên vẫn còn nhiều bài toán và nhiều thách thức đặt ra cho các nhà nghiên cứu, các thách thức này có thể được phân chia theo các nhóm vấn đề [4]:

- ***Phương pháp luận khai phá dữ liệu***:
    - Khai phá các loại tri thức từ các loại dữ liệu khác nhau như dữ liệu sinh học, web... Các dữ liệu được thu thập từ nhiều nguồn khác nhau và được thể hiện dưới nhiều dạng khác nhau một cách hỗn hợp, hỗn tạp vì thế yêu cầu các hệ thống khai phá dữ liệu phải có khả năng khai phá và tích hợp các dữ liệu này trong cùng một quá trình khai phá. Mặt khác, có những loại dữ liệu có nhiều đặc điểm riêng như dữ liệu dạng chuỗi (*sequence*), chuỗi thời gian (*time-series*)... do vậy cần phải có các phương pháp khai phá dữ liệu khác nhau cho chúng. Các thuộc tính của dữ liệu cũng có các cấu trúc và mối quan hệ lẫn nhau phức tạp. Cấu trúc phân cấp của các thuộc tính hoặc các giá trị, các quan hệ giữa các thuộc tính và các cách phức tạp hơn nữa để biểu diễn tri thức về nội dung của cơ sở dữ liệu sẽ yêu cầu các thuật toán có thể sử dụng hiệu quả trên những thông tin này. Các thuật toán khai phá dữ liệu trước đây đã được phát triển cho những bản ghi có thuộc tính và giá trị đơn giản, nhưng những kỹ thuật mới để suy dẫn các quan hệ giữa các biến cũng đang được phát triển.
    - Khả năng thực hiện. Hiệu suất, hiệu lực và khả năng mở rộng của thuật toán. Khối lượng dữ liệu ngày càng lớn có kích thước lên đến hàng *terabyte* hay *petabyte* do vậy cần thiết phải nghiên cứu các thuật

toán có hiệu năng cao để có thể xử lý được những dữ liệu như vậy. Ngoài đặc điểm về số lượng dữ liệu lớn, số chiều của dữ liệu cũng có khả năng lớn, do vậy cần có các phương pháp hay thuật toán để xử lý và trích chọn những đặc trưng, tìm kiếm mô hình trên đó.

- Kiểm định mẫu/mô hình. Nghiên cứu các phương pháp kiểm tra và đánh giá các mẫu/mô hình đã khai phá được có khả năng ứng dụng hay phù hợp với lĩnh vực yêu cầu hay không.
- Xử lý dữ liệu nhiễu và dữ liệu không đầy đủ. Đây là một bài toán rất dễ gặp trong các cơ sở dữ liệu kinh doanh. Các thuộc tính quan trọng có thể bị khuyết nếu như cơ sở dữ liệu không được thiết kế với ý tưởng phát hiện tri thức. Các giải pháp có thể bao gồm các sử dụng nhiều chiến lược thống kê phức tạp để xác định những biến ẩn và những sự phụ thuộc.
- Các phương pháp khai phá song song, phân tán và gia tăng. Với sự gia tăng dữ liệu ngày một nhiều và số chiều dữ liệu ngày một lớn cần thiết phải nghiên cứu các kỹ thuật và phương pháp khai phá dữ liệu trên các hệ thống song song, phân tán. Các phương pháp khai phá dữ liệu gia tăng (*incremental*) cũng cần được nghiên cứu để có thể thực hiện trên các bộ dữ liệu lớn và cải thiện tốc độ khai phá dữ liệu. Khai phá dữ liệu gia tăng cũng nhằm đáp ứng cho việc dữ liệu thay đổi nhanh chóng có thể làm cho các mẫu tìm được trước đó không đúng. Hơn nữa, các biến được đo trong một cơ sở dữ liệu ứng dụng đã cho có thể bị thay đổi, xoá hoặc thêm vào với những thước đo mới theo thời gian. Các giải pháp có thể bao gồm các phương pháp mang tính gia tăng để cập nhật các mẫu và xử lý các thay đổi như một cơ hội để phát hiện bằng cách sử dụng nó để gợi ý chỉ cho các mẫu của sự thay đổi đó
- Tích hợp tri thức đã khai phá được với các hệ thống đã tồn tại, kết hợp với tri thức cơ sở đã có. Nhiều phương pháp *KDD*, nhiều công cụ hiện tại không thực sự tương tác, không dễ dàng tích hợp với tri thức có trước của bài toán, ngoại trừ những cách đơn giản như sử dụng tri thức lĩnh vực với vai trờ là một điều kiện quan trọng trong toàn bộ các bước của tiến trình *KDD*.

- ***Tương tác với người dùng***:
    - Ngôn ngữ truy vấn cho khai phá dữ liệu và tích hợp với các hệ thống truy vấn khác. Một hệ thống phát hiện độc lập thường không hữu dụng nhiều. Vấn đề nổi bật của sự tích hợp bao gồm tích hợp với Hệ quản trị cơ sở dữ liệu (*DBMS*) thông qua giao diện truy vấn, tích hợp với các bảng tính và công cụ thể hiện trực quan, trợ giúp các bộ đọc cảm biến thời gian thực.
    - Thể hiện các kết quả đã khai phá được dưới dạng trực quan hoặc biểu thức. Điểm quan trọng trong các ứng dụng là làm cho con người có thể hiểu được các mẫu đã phát hiện. Các giải pháp có thể là: biểu diễn đồ hoạ, cấu trúc luật với những đồ thị có hướng nối, biểu diễn bằng ngôn ngữ tự nhiên, sử dụng các kỹ thuật trực quan hoá *(visualization)* dữ liệu và tri thức.
    - Khai phá dữ liệu tương tác với các mức độ trừu tượng tri thức khác nhau. Các hệ thống phát hiện tri thức và khai phá dữ liệu cần phải có sự tương tác với người sử dụng để có thể chọn lọc được những thông tin hay tri thức dưới nhiều dạng và mức độ trừu tượng khác nhau tuỳ thuộc đối tượng người sử dụng và mục đích sử dụng. Quá trình tương tác này có thể được thực hiện thông qua các giao diện người sử dụng để truyền và dự đoán các tham số cho hệ thống khai phá dữ liệu.

1.6 Các lĩnh vực ứng dụng của phát hiện tri thức và khai phá dữ liệu

Mặc dù khai phá dữ liệu và phát hiện tri thức là một xu hướng nghiên cứu tương đối mới, nhưng thu hút nhiều nhà nghiên cứu bởi vì các ứng dụng thực tế của nó trong nhiều lĩnh vực. Sau đây là một số ứng dụng tiêu biểu:

- Phân tích dữ liệu và hỗ trợ ra quyết định: ứng dụng này phổ biến trong thương mại, tài chính và thị trường chứng khoán…

- Giáo dục: phân tích dữ liệu sinh viên đại học để cố vấn lộ trình học tập, dự đoán khả năng ra trường sớm hay muộn….

- Y tế: Tìm kiếm sự liên quan tiềm năng giữa các triệu chứng, chẩn đoán và phương pháp điều trị (dinh dưỡng, bác sĩ phẫu thuật, toa thuốc)

- Khai phá dữ liệu văn bản và web: Tóm tắt tài liệu, khôi phục văn bản và tìm kiếm văn bản, phân lớp văn bản và siêu văn bản.

- Tin sinh học: Tìm kiếm và so sánh thông tin di truyền điển hình hoặc đặc biệt như bộ gen và DNA, các mối quan hệ ngầm giữa một số gen và một số bệnh di truyền, ….

- Tài chính và thị trường chứng khoán: kiểm tra dữ liệu để trích xuất thông tin dự đoán cho giá của các loại cổ phiếu

- Những ứng dụng khác trong các lĩnh vực viễn thông, bảo hiểm y tế, thiên văn học, chống khủng bố, thể thao,…

1.7 Kỹ thuật khai phá luật kết hợp

1.7.1 Lý thuyết về luật kết hợp

Từ khi được giới thiệu vào năm 1993 trở đi, bài toán khai phá luật kết hợp nhận được rất nhiều sự quan tâm của nhiều nhà khoa học. Ngày nay việc khai thác các luật như thế vẫn là một trong những phương pháp khai phá mẫu phổ biến nhất trong khai phá dữ liệu và phát hiện tri thức.

Cho một tập I = {I1, I2, ...,Im} các tập m khoản mục (item), một giao dịch (transaction) T được định nghĩa như một tập con (subset) của các khoản mục trong I ($T \in I$). Tương tự như khái niệm tập hợp, các giao dịch không được trùng lặp, nhưng có thể nới rộng tính chất này của tập hợp và trong các thuật toán sau này, người ta đều giả thiết rằng các khoản mục trong một giao dịch và trong tất cả các tập mục (item set) khác, có thể coi chúng đã được sắp xếp theo thứ tự từ điển của các item.

Gọi D là cơ sở dữ liệu của n giao dịch và mỗi giao dịch được đánh nhãn với một định danh duy nhất (Unique Transasction Identifier). Nói rằng, một giao dịch $T \in D$ hỗ trợ (support) cho một tập $X \subseteq I$ nếu nó chứa tất cả các item của X, nghĩa là $X \subseteq T$, trong một số trường hợp người ta dùng ký hiệu T(X) để chỉ tập các giao dịch hỗ trợ cho X. Kí hiệu support(X) (hoặc sup(X), s(X)) là tỷ lệ phần trăm của các giao dịch hỗ trợ X trên tổng các giao dịch trong D, nghĩa là:

$$sup(X) = \{ T \in D, X \subseteq T\} \ / \ D$$

Độ hỗ trợ tối thiểu (minimum support) minsup là một giá trị cho trước bởi người sử dụng. Nếu tập mục X có $sup(X) \geq$ minsup thì ta nói X là một tập các mục phổ biến (hoặc large itemset). Một tập phổ biến được sử dụng như một tập đáng quan tâm trong các thuật toán, ngược lại, những tập không phải tập phổ biến là những tập không đáng quan tâm. Trong các trình bày sau này, ta sẽ sử dụng những cụm từ khác như "X có độ hỗ trợ tối thiểu", hay "X không có độ hỗ trợ

tối thiểu" cũng để nói lên rằng X thỏa mãn hay không thỏa mãn điều kiện support(X) ≥ minsup.

1.7.2 Định nghĩa luật kết hợp

Một luật kết hợp có dạng R: X => Y, trong đó X, Y là tập các mục, X, Y $\subseteq$ I và X $\cap$Y = $\varnothing$. X được gọi là tiên đề và Y được gọi là hệ quả của luật.

Luật X => Y tồn tại một độ tin cậy c (confidence-conf). Độ tin cậy c được định nghĩa là khả năng giao dịch T hỗ trợ X thì cũng hỗ trợ Y. Ta có công thức tính độ tin cậy c như sau:

$$\text{conf}(X \Rightarrow Y) = p(Y \subseteq I \mid X \subseteq I) = \frac{p(Y \subseteq T \wedge X \subseteq T)}{p(X \subseteq T)} = \frac{\sup(X \cup Y)}{\sup(X)}$$

Tuy nhiên, không phải bất cứ luật kết hợp nào có mặt trong tập các luật có thể được sinh ra cũng đều có ý nghĩa trên thực tế. Mà các luật đều phải thoả mãn một ngưỡng hỗ trợ và tin cậy cụ thể. Thực vậy, cho một tập các giao dịch D, bài toán phát hiện luật kết hợp là sinh ra tất cả các luật kết hợp mà có độ tin cậy (conf) lớn hơn độ tin cậy tối thiểu (minconf) và độ hỗ trợ (supp) lớn hơn độ hỗ trợ tối thiểu (minsupp) tương ứng do người dùng xác định.

Quy trình khai phá luật kết hợp được thực hiện lần lượt theo hai bài toán sau:

- Bài toán 1: Tìm tất cả các tập mục mà có độ hỗ trợ lớn hơn độ hỗ trợ tốt thiểu do người dùng xác định. Các tập mục thoả mãn độ hỗ trợ tối thiểu được gọi là các tập mục phổ biến (theo ngưỡng minsupp).
- Bài toán 2: Dùng các tập mục phổ biến để sinh ra các luật mong muốn. Ý tưởng chung là nếu gọi ABCD và AB là các tập mục phổ biến, thì chúng ta có thể xác định luật nếu AB => CD giữ lại với tỷ lệ độ tin cậy:

$$conf = \frac{\sup(ABCD)}{\sup(AB)}$$

  nếu conf ≥ minconf thì luật được giữ lại (luật này sẽ thoả mãn độ hỗ trợ tối thiểu vì ABCD là phổ biến)

Các thành phần cấu tạo nên một luật bao gồm:

- Phần tiền đề (antecedent): Thông thường gồm nhiều mệnh đề, thường được kết hợp với nhau bởi toán tử AND.

- Phần kết quả (consequent): Thông thường là 1 mệnh đề với mục đích dễ dàng trong việc dự đoán hay ra quyết định.

Vì các luật phải xác định xem luật hữu dụng nên có hai tiêu chí để đánh giá một luật thu được:

- Độ tin cậy (confidence): xác suất mà nếu nguyên nhân đúng thì kết quả đúng trong CSDL. Độ chính xác cao cho thấy luật có độ tin cậy cao, có thể dựa vào đó để dự đoán hoặc ra quyết định. Trong kinh doanh, độ chính xác của luật vô cùng quan trọng vì nó sẽ là một thông tin dự doán hữu ích có thể đem ra áp dụng. Nếu áp dụng một luật có độ chính xác thấp thì dự đoán hay quyết định mang tính chất phỏng đoán nhiều hơn.
- Độ hỗ trợ (support): được đo bằng tỷ lệ giữa số bản ghi chứa các thành phần của luật và tổng số luật. Độ hỗ trợ trả lời câu hỏi về tuần suất xuất hiện của luật. Độ hỗ trợ càng cao thì khả năng áp dụng luật đó vào thực tế càng cao, bởi vì các triển vọng thực tế sẽ xảy ra như thế - dữ liệu lịch sử đã nói lên điều đó.

Ví dụ ta có một luật: Nếu mua lốp thì mua phanh (ta có thể kí hiệu hình thức như sau Lốp $\rightarrow$ Phanh). Bảng thống kê từ cơ sở dữ liệu như sau:

MÔ TẢ	SỐ LƯỢNG
Khách mua lốp	50
Khách mua phanh	30
Khách mua cả lốp và phanh	20
Tổng số khách	**100**

Các thông số của luật có thể được tính như sau:

confidence = 20/50 = 0.4 (40%)

support = 20/100 = 0.2 (20%)

Ngoài ra, còn có một số các tiêu chuẩn khác để đo tính hữu dụng của luật, đó là:

- *Độ phủ (coverage):* Trả lời câu hỏi luật thường được dùng như thế nào để dự đoán và được tính bằng tỷ lệ giữa số bản ghi có chứa tiền đề trên tổng số bản ghi.
- *Độ đo ý nghĩa (significant):* So sánh giữa một mẫu cho trước và một trường hợp ngẫu nhiên.
- *Độ đơn giản (simplicity):* Rõ ràng người dùng quan tâm đến các luật đơn giản mà có hiệu quả hơn là các luật phức tạp bởi vì chúng gợi tả hơn, dễ hiểu hơn. Có nhiều cách để đơn giản hoá các luật như phép tách...
- *Độ bất ngờ (novelty):* Như đã đề cập ở trên, chúng ta chỉ đi tìm các luật có tính bất ngờ, tất nhiên phụ thuộc vào khung quy chiều, có thể cho hệ thống hoặc cho người dùng.

1.7.3 Một số hướng tiếp cận trong khai phá luật kết hợp

Lĩnh vực khai thác luật kết hợp cho đến nay đã được nghiên cứu và phát triển theo nhiều hướng khác nhau. Có những đề xuất nhằm cải tiến tốc độ thuật toán, có những đề xuất nhằm tìm kiếm luật có ý nghĩa hơn… và có một số hướng chính sau đây.

- **Luật kết hợp nhị phân** (binary association rule hoặc boolean association rule) : là hướng nghiên cứu đầu tiên của luật kết hợp. Hầu hết các nghiên cứu ở thời kỳ đầu về luật kết hợp đều liên quan đến luật kết hợp nhị phân. Trong dạng luật kết hợp này, các mục (thuộc tính) chỉ được quan tâm là có hay không xuất hiện trong giao tác của cơ sở dữ liệu chứ không quan tâm về "mức độ" xuất hiện.

  Ví dụ: Trong hệ thống tính cước điện thoại thì việc gọi 10 cuộc điện thoại và 1 cuộc được xem là giống nhau. Thuật toán tiêu biểu nhất khai phá dạng luật này là thuật toán **Apriori** và các biến thể của nó. Đây là dạng luật đơn giản và các luật khác cũng có thể chuyển về dạng luật này nhờ một số phương pháp như rời rạc hoá, mờ hoá, …

  Một ví dụ về dạng luật này: *gọi liên tỉnh= "yes" AND gọi di động= "yes"* => *gọi quốc tế = "yes" AND gọi dịch vụ 108 = "yes", với độ hỗ trợ 20% và độ tin cậy 80%*.

- **Luật kết hợp có thuộc tính số và thuộc tính hạng mục** (quantitative and categorial association rule) : Các thuộc tính của các cơ sở dữ liệu thực tế có kiểu rất đa dạng (nhị phân - binary, số - quantitative, hạng mục - categorial,…). Để phát hiện luật kết hợp với các thuộc tính này, các nhà nghiên cứu đã đề xuất một số phương pháp rời rạc hoá nhằm chuyển dạng luật này về dạng nhị phân để có thể áp dụng các thuật toán đã có. Một ví dụ về dạng luật này: *phương thức gọi = "Tự động" AND giờ gọi IN ["23:00:39.. 23:00:59"] AND Thời gian đàm thoại IN ["200.. 300"] => gọi liên tỉnh = "có" , với độ hỗ trợ là 23. 53% , và độ tin cậy là 80%* .
- **Luật kết hợp tiếp cận theo hướng tập thô** (mining association rules base on rough set) : Tìm kiếm luật kết hợp dựa trên lý thuyết tập thô.
- **Luật kết hợp nhiều mức** (multi-level association rule) : Với cách tiếp cận theo luật này sẽ tìm kiếm thêm những luật có dạng: *"mua máy tính PC => mua hệ điều hành AND mua phần mềm tiện ích văn phòng,…"* thay vì chỉ những luật quá cụ thể như *"mua máy tính IBM PC => mua hệ điều hành Microsoft Windows AND mua phần mềm tiện ích văn phòng Microsoft Office,…".* Như vậy dạng luật đầu là dạng luật tổng quát hoá của dạng luật sau và tổng quát theo nhiều mức khác nhau.
- **Luật kết hợp mờ** (fuzzy association rule) : Với những hạn chế còn gặp phải trong quá trình rời rạc hoá các thuộc tính số (quantitave attributes), các nhà nghiên cứu đã đề xuất luật kết hợp mờ nhằm khắc phục các hạn chế trên và chuyển luật kết hợp về một dạng tự nhiên hơn, gần gũi hơn với người sử dụng, một ví dụ của dạng này là : *"thuê bao tư nhân = 'yes' AND thời gian đàm thoại lớn AND cước nội tỉnh = 'yes' => cước không hợp lệ = 'yes', với độ hỗ trợ 4% và độ tin cậy 85%"*. Trong luật trên, điều kiện "*thời gian đàm thoại lớn"* ở vế trái của luật là một thuộc tính đã được mờ hoá.
- **Luật kết hợp với thuộc tính được đánh trọng số** (association rule with weighted items) : Trong thực tế, các thuộc tính trong cơ sở dữ liệu không phải lúc nào cũng có vai trò như nhau. Có một số thuộc tính được chú trọng hơn và có mức độ quan trọng cao hơn các thuộc tính khác. Ví dụ: *khi khảo sát về doanh thu hàng tháng, thông tin về thời gian đàm thoại, vùng cước là quan trọng hơn nhiều so với thông tin về phương thức gọi*... Trong quá trình tìm kiếm luật, chúng ta sẽ gán thời gian gọi, vùng

cước các trọng số lớn hơn thuộc tính phương thức gọi. Đây là hướng nghiên cứu rất thú vị và đã được một số nhà nghiên cứu đề xuất cách giải quyết bài toán này. Với luật kết hợp có thuộc tính được đánh trọng số, chúng ta sẽ khai thác được những luật "hiếm" (tức là có độ hỗ trợ thấp, nhưng có ý nghĩa đặc biệt hoặc mang rất nhiều ý nghĩa).

- **Khai thác Luật kết hợp song song** (parallel mining of association rules): Bên cạnh khai thác luật kết hợp tuần tự, các nhà làm tin học cũng tập trung vào nghiên cứu các thuật giải song song cho quá trình phát hiện luật kết hợp. Nhu cầu song song hoá và xử lý phân tán là cần thiết bởi kích thước dữ liệu ngày càng lớn hơn nên đòi hỏi tốc độ xử lý cũng như dung lượng bộ nhớ của hệ thống phải được đảm bảo. Có rất nhiều thuật toán song song khác nhau đã đề xuất để có thể không phụ thuộc vào phần cứng.

Bên cạnh những nghiên cứu về những biến thể của luật kết hợp, các nhà nghiên cứu còn chú trọng đề xuất những thuật toán nhằm tăng tốc quá trình tìm kiếm tập phổ biến từ cơ sở dữ liệu. Ngoài ra, còn có một số hướng nghiên cứu khác về khai thác luật kết hợp như: khai thác luật kết hợp trực tuyến, khai thác luật kết hợp được kết nối trực tuyến đến các kho dữ liệu đa chiều (Multidimensional data, data warehouse) thông qua công nghệ OLAP (Online Analysis Processing), MOLAP (multidimensional OLAP), ROLAP (Relational OLAP), ADO (Active X Data Object) for OLAP...

1.8 Cây quyết định

1.8.1 Sơ lược về cây quyết định

Cuối những năm 70 đầu những năm 80, J.Ross Quinlan đã phát triển một thuật toán sinh cây quyết định. Đây là một tiếp cận tham lam, trong đó nó xác định một cây quyết định được xây dựng từ trên xuống một cách đệ quy theo hướng chia để trị. Hầu hết các thuật toán sinh cây quyết định đều dựa trên tiếp cận top-down trình bày sau đây, trong đó nó bắt đầu từ một tập các bộ huấn luyện và các nhãn phân lớp của chúng. Tập huấn luyện được chia nhỏ một các đệ quy thành các tập con trong quá trình cây được xây dựng.

Cây quyết định là một mô tả tri thức dạng đơn giản nhằm phân các đối tượng dữ liệu thành một số lớp nhất định. Các nút của cây được gán nhãn là tên của các thuộc tính, các cạnh được gán các giá trị có thể của các thuộc tính, các lá

mô tả các lớp khác nhau. Các đối tượng được phân lớp theo các đường đi trên cây, qua các cạnh tương ứng với giá trị của thuộc tính của đối tượng tới lá.

Quá trình xây dựng cây quyết định là quá trình phát hiện ra các luật phân chia tập dữ liệu đã cho thành các lớp đã được định nghĩa trước. Trong thực tế, tập các cây quyết định có thể có đối với bài toán này rất lớn và rất khó có thể duyệt hết được một cách tường tận.

1.8.2 Định nghĩa cây quyết định

Một cây quyết định là một cấu trúc hình cây, trong đó:

- Mỗi đỉnh trong (đỉnh có thể khai triển được) biểu thị cho một phép thử đối với một thuộc tính.
- Mỗi nhánh biểu thị cho một kết quả của phép thử.
- Các đỉnh lá (các đỉnh không khai triển được) biểu thị các lớp hoặc các phân bổ lớp.
- Đỉnh trên cùng trong một cây được gọi là gốc.

1.8.3 Xây dựng cây quyết định

Việc sinh cây quyết định bao gồm hai giai đoạn:

+ Giai đoạn 1: Xây dựng cây

  - Tại thời điểm khởi đầu, tất cả các cây (case) dữ liệu học đều nằm tại gốc.
  - Các cây dữ liệu được phân chia đệ quy trên cơ sở các thuộc tính được chọn.

+ Giai đoạn 2: Rút gọn cây

  - Phát hiện và bỏ đi các nhánh chứa các điểm dị thường và nhiều trong dữ liệu.

1.8.4 Một số thuật toán xây dựng cây quyết định

Ý tưởng chính của các thuật toán xây dựng cây quyết định là dựa vào phương pháp tham lam (greedy), phân chia tập mẫu dựa trên thuộc tính cho kết quả tối ưu hóa tiêu chuẩn, thuộc tính được chọn là thuộc tính cho độ đo tốt nhất, có lợi nhất trong quá trình phân lớp. Độ đo để đánh giá chất lượng phân chia là độ đo sự đồng nhất như độ đo Entropy (Information Gain), Information Gain Ratio, Gini Index. Cách phân chia các mẫu dựa trên độ đo sự đồng nhất của dữ liệu, tức là tạo ra các nhóm sao cho một lớp chiếm ưu thế trong từng nhóm. Vấn đề điều kiện dừng là

khi tất cả các mẫu rơi vào một nút thuộc về cùng một lớp, không còn thuộc tính nào có thể dùng để phân chia mẫu nữa, không còn lại mẫu nào tại nút

- *Thuật toán CLS:* là một trong những thuật toán xây dựng cây quyết đinh ra đời sớm nhất. CLS thường chỉ áp dụng cho các CSDL có số lượng thuộc tính nhỏ, mối quan hệ giữa các thuộc tính không quá phức tạp, giá trị thuộc tính thuộc dạng phân loại rời rạc. Còn đối với các CSDL lớn và có chứa các thuộc tính mà giá trị của nó là liên tục thì CLS làm việc không hiệu quả. Do thuật toán CLS chưa có tiêu chuẩn lựa chọn thuộc tính trong quá trình xây dựng cây mà với cùng một tập dữ liệu đầu vào áp dụng thuật toán CLS có thể cho ra nhiều cây kết quả khác nhau. Nhưng đây là thuật toán đơn giản, dễ cài đặt, phù hợp trong việc hình thành ý tưởng và giải quyết những nhiệm vụ đơn giản.

Xây dựng cây quyết định lần đầu tiên được Hoveland và Hint giới thiệu trong Concept Learning System (CLS) vào cuối những năm 50 của thế kỷ 20. Sau đó gọi tắt là thuật toán CLS. Thuật toán CLS được thiết kế theo chiến lược chia để trị từ trên xuống và gồm các bước sau:

B1: Tạo một nút T, nút này gồm tất cả các mẫu của tập huấn luyện.

B2: Nếu tất cả các mẫu trong T có thuộc tính quyết định mang giá trị "yes" (hay thuộc cùng một lớp), thì gán nhãn cho nút T là "yes" và dừng lại. T lúc này là nút lá.

B3: Nếu tất cả các mẫu trong T có thuộc tính quyết định mang giá trị "no" (hay thuộc cùng một lớp), thì gán nhãn cho nút T là "no" và dừng lại. T lúc này là nút lá.

B4: Trường hợp ngược lại các mẫu của tập huấn luyện thuộc cả hai lớp "yes" và "no" thì:

a. Chọn một thuộc tính X trong tập thuộc tính của tập mẫu dữ liệu, X có các giá trị v1,v2, …vn.

b. Chia tập mẫu trong T thành các tập con T1, T2,….,Tn. chia theo giá trị của X.

c. Tạo n nút con Ti (i=1,2…n) với nút cha là nút T.

d. Tạo các nhánh nối từ nút T đến các nút Ti (i=1,2…n)

B5: Thực hiện lặp cho các nút con Ti(i =1,2..n) và quay lại bước 2.

Ta nhận thấy trong bước 4 của thuật toán, thuộc tính được chọn để triển khai cây là tuỳ ý. Do vậy cùng với một tập mẫu dữ liệu huấn luyện nếu áp dụng thuật toán CLS với thứ tự chọn thuộc tính triển khai cây khác nhau, sẽ cho ra các cây có hình dạng khác nhau. Việc lựa chọn thuộc tính sẽ ảnh hưởng tới độ rộng, độ sâu, độ phức tạp của cây. Vì vậy một câu hỏi đặt ra là thứ tự thuộc tính nào được chọn để triển khai cây sẽ là tốt nhất. Vấn đề này sẽ được giải quyết trong thuật toán ID3 dưới đây.

- Thuật toán ID3: được phát biểu bởi Quinlan (trường đại học Syney, Australia) và được công bố vào cuối thập niên 70 của thế kỷ 20. Sau đó, thuật toán ID3 được giới thiệu và trình bày trong mục Induction on Decision Trees, Machine learning năm 1986. Quinlan đã khắc phục được hạn chế của thuật toán CLS. ID3 cho cây kết quả tối ưu hơn thuật toán CLS. Khi áp dụng thuật toán ID3 cho cùng một tập dữ liệu đầu vào và thử nhiều lần thì cho cùng một kết quả bởi vì thuộc tính ứng viên ở mỗi bước trong quá trình xây dựng cây được lựa chọn trước. Tuy nhiên thuật toán này cũng chưa giải quyết được về vấn đề thuộc tính số, liên tục, số lượng các thuộc tính còn bị hạn chế và ID3 làm việc không hiệu quả với dữ liệu bị nhiễu hoặc bị thiếu.

Giải thuật quy nạp cây ID3 (gọi tắt là ID3) là một giải thuật học đơn giản nhưng tỏ ra thành công trong nhiều lĩnh vực. ID3 biểu diễn các khái niệm (concept) ở dạng cây quyết định (decision tree). Biểu diễn này cho phép chúng ta xác định phân loại đối tượng bằng cách kiểm tra các giá trị của nó trên một số thuộc tính nào đó. Như vậy, nhiệm vụ của giải thuật ID3 là học cây quyết định từ tập dữ liệu rèn luyện (training data). Hay nói khác hơn, giải thuật có:

Đầu vào: Một tập hợp các ví dụ. Mỗi ví dụ bao gồm các thuộc tính mô tả một tình huống, hay một đối tượng nào đó, và một giá trị phân loại của nó.

Đầu ra: Cây quyết định có khả năng phân loại đúng đắn các ví dụ trong tập dữ liệu rèn luyện, và hy vọng là phân loại đúng cho cả các ví dụ chưa gặp trong tương lai.

Thuật toán ID3 xây dựng cây quyết định dựa vào sự phân lớp các đối tượng (mẫu huấn luyện) bằng cách kiểm tra giá trị các thuộc tính. ID3 xây dựng cây quyết định từ trên xuống (*top -down*) bắt đầu từ một tập các đối tượng và các thuộc tính của nó. Tại mỗi nút của cây, tiến hành việc kiểm tra các thuộc tính để tìm ra thuộc

tính tốt nhất được sử dụng để phân chia tập các đối tượng mẫu, theo các giá trị của thuộc tính được chọn để mở rộng. Quá trình này được thực hiện một cách đệ quy cho đến khi mọi đối tượng của phân vùng đều thuộc cùng một lớp; lớp đó trở thành nút lá của cây. Để làm được việc này thuật toán ID3 có sử dụng tới hai hàm Entropy và Gain

Hàm entropy

Khái niệm entropy của một tập S được định nghĩa trong Lý thuyết thông tin là số lượng mong đợi các bít cần thiết để mã hóa thông tin về lớp của một thành viên rút ra một cách ngẫu nhiên từ tập S. Trong trường hợp tối ưu, mã có độ dài ngắn nhất. Theo lý thuyết thông tin, mã có độ dài tối ưu là mã gán $-\log_2 p$ bits cho thông điệp có xác suất là p.

Trong trường hợp S là tập ví dụ, thì thành viên của S là một ví dụ, mỗi ví dụ thuộc một lớp hay có một giá trị phân loại.

– Entropy có giá trị nằm trong khoảng [0..1],

– Entropy(S) = 0 ⇔ tập ví dụ S chỉ toàn ví dụ thuộc cùng một loại, hay S là thuần nhất.

– Entropy(S) = 1 ⇔ tập ví dụ S có các ví dụ thuộc các loại khác nhau với độ pha trộn là cao nhất.

– 0 < Entropy(S) < 1 ⇔ tập ví dụ S có số lượng ví dụ thuộc các loại khác nhau là không bằng nhau.

Để đơn giản ta xét trường hợp các ví dụ của S chỉ thuộc loại âm (-) hoặc dương (+)

- p_+ là phần các ví dụ dương trong tập S.
- p_- là phần các ví dụ âm trong tập S.

Khi đó, entropy đo độ pha trộn của tập S theo công thức sau:

$$\text{Entropy}(S) = -p_+\log_2 p_+ - p_-\log_2 p_-$$

Một cách tổng quát hơn, nếu các ví dụ của tập S thuộc nhiều hơn hai loại, giả sử là có c giá trị phân loại thì công thức entropy tổng quát là:

$$\text{Entropy}(S) = \sum_{i=1}^{c} - p_i \log_2 p_i$$

Trong đó: p_i là tỷ lệ mẫu thuộc lớp i trên tập hợp S các mẫu kiểm tra. ID3 được xem là một cải tiến của CLS. Tuy nhiên thuật toán ID3 không có khả năng xử lý đối với những dữ liệu có chứa thuộc tính số - thuộc tính liên tục (numeric attribute) và khó khăn trong việc xử lý các dữ liệu thiếu (missing data) và dữ liệu nhiễu (noisy data). Vấn đề này được giải quyết bởi cải tiến C4.5 sau đây.

- *Thuật toán C4.5:* để khắc phục những hạn chế của thuật toán ID3, Quinlan đã đưa ra thuật toán C4.5. Thuật toán này có thể phân lớp các dữ liệu có chứa thuộc tính số (hoặc thuộc tính liên tục), thuộc tính đa trị và giải quyết được vấn đề dữ liệu bị nhiễu hoặc bị thiếu. Tuy nhiên C4.5 vẫn có hạn chế đó là làm việc không hiệu quả với những cơ sở dữ liệu rất lớn vì chưa giải quyết được vấn đề bộ nhớ.

Mô tả thuật toán dưới dạng giả mã như sau [1]:

Function xay_dung_cay(T)

{

1. <Tính toán tần xuất các giá trị trong các lớp của T>;

2. If <Kiểm tra các mẫu, nếu thuộc cùng một lớp hoặc có rất ít mẫu khác lớp>Then <Trả về 1 nút lá>

Else <Tạo một nút quyết định N>;

3. For <Với mỗi thuộc tính A> Do <Tính giá trị Gain(A)>;

4. <Tại nút N, thực hiện việc kiểm tra để chọn ra thuộc tính có giá trị Gain tốt nhất (lớn nhất). Gọi N.test là thuộc tính có Gain lớn nhất>;

5. If <Nếu N.test là thuộc tính liên tục> Then <Tìm ngưỡng cho phép tách của N.test>;

6. For <Với mỗi tập con T` được tách ra từ tập T> Do

(T` được tách ra theo quy tắc:

- Nếu N.test là thuộc tính liên tục tách theo ngưỡng ở bước 5

- Nếu N.test là thuộc tính phân loại rời rạc tách theo các giá trị của thuộc tính này.

)

7. { If <Kiểm tra, nếu T' rỗng>} Then

<Gán nút con này của nút N là nút lá>;

Else

8. <Gán nút con này là nút được trả về bằng cách gọi đệ qui lại đối với hàm xay_dung_cay(T'), với tập T'>;

}

9. <Tính toán các lỗi của nút N>;

<Trả về nút N>;

}

T là tập dữ liệu ban đầu, số lượng mẫu được ký hiệu là |T|. Quá trình xây dựng cây được tiến hành từ trên xuống. Đầu tiên ta xác định nút gốc, sau đó xác định các nhánh xuất phát từ gốc này. Tập T được chia thành các tập con theo các giá trị của thuộc tính được xét tại nút gốc. Nếu T có m thuộc tính thì có m khả năng để lựa chọn thuộc tính. Một số thuật toán thì trong quá trình xây dựng cây mỗi thuộc tính chỉ được xét một lần, nhưng với thuật toán này một thuộc tính có thể được xét nhiều lần.

Xét thuộc tính X có n giá trị lần lượt là $L_1, L_2, ..L_n$. Khi đó, ta có thể chia tập T ra thành n tập con $X_i (i=1..n)$ theo các giá trị của X. Tần xuất freq (C_j, T) là số lượng mẫu của tập T nào đó được xếp vào lớp con C_j. Xác xuất để một mẫu được lấy bất kỳ từ T thuộc lớp C_j là:

$$P = \frac{freq(C_j, T)}{|T|}$$

Khi đó Information (T) được tính theo công thức sau:

$$\text{Info(T)} = -\sum_{j=1}^{k} \frac{\text{freq}(C_j, T)}{|T|} * \log_2\left(\frac{freq(C_j, T)}{|T|}\right) \qquad = \text{Entropy (T)}$$

Công thức này đánh giá số lượng thông tin trung bình cần thiết để phân lớp các mẫu trong tập hợp T. Khi đó:

$$\text{Info}_x(T) = -\sum_{i-1}^{n} \frac{|T_i|}{|T|} * \text{Info}(T_i)$$

$$Gain(X,T) = \text{Entropy(T)} - \text{Info}_X(T)$$

Thuộc tính được lựa chọn tại một nút là thuộc tính có Gain lớn nhất. Thuộc tính được chọn sẽ được dùng để phân lớp tập mẫu dữ liệu tại nút đó. Quá trình phân chia được tiếp tục cho đến khi các mẫu trong tập dữ liệu được phân lớp hoàn toàn.

1.8.5 Ưu điểm của cây quyết định

So với các phương pháp khai phá dữ liệu khác, cây quyết định là phương pháp có một số ưu điểm:

- Cây quyết định dễ hiểu. Người ta có thể hiểu mô hình cây quyết định sau khi được giải thích ngắn.
- Việc chuẩn bị dữ liệu cho một cây quyết định là cơ bản hoặc không cần thiết. Các kỹ thuật khác thường đòi hỏi chuẩn hóa dữ liệu, cần tạo các biến phụ (dummy variable) và loại bỏ các giá trị rỗng.
- Cây quyết định có thể xử lý cả dữ liệu có giá trị bằng số và dữ liệu có giá trị là tên thể loại. Các kỹ thuật khác thường chuyên để phân tích các bộ dữ liệu chỉ gồm một loại biến. Chẳng hạn, các luật quan hệ chỉ có thể dùng cho các biến tên, trong khi mạng nơ-ron chỉ có thể dùng cho các biến có giá trị bằng số.
- Cây quyết định là một mô hình hộp trắng. Nếu có thể quan sát một tình huống cho trước trong một mô hình, thì có thể dễ dàng giải thích điều kiện đó bằng logic Boolean. Mạng nơ-ron là một ví dụ về mô hình hộp đen, do lời giải thích cho kết quả quá phức tạp để có thể hiểu được.
- Có thể thẩm định một mô hình bằng các kiểm tra thống kê. Điều này làm cho ta có thể tin tưởng vào mô hình.
- Cây quyết định có thể xử lý tốt một lượng dữ liệu lớn trong thời gian ngắn. Có thể dùng máy tính cá nhân để phân tích các lượng dữ liệu lớn trong một thời gian đủ ngắn để cho phép các nhà chiến lược đưa ra quyết định dựa trên phân tích của cây quyết định.

1.9 Tổng kết chương 1

Chương 1 đã tóm tắt được cơ sở lý thuyết, xu hướng liên quan đến phát hiện tri thức và khai phá dữ liệu. Hiểu được quy trình, phương pháp, lĩnh vực khai phá dữ liệu. Hiểu những ý tưởng chính trong hai kỹ thuật khai phá luật kết hợp và cây quyết định như; định nghĩa, các bước thực hiện trong luật kết hợp, cây quyết định, các tham số và công thức tính các tham số, ý tưởng của các thuật toán xây dựng cây quyết định. Việc nắm được những vấn đề cơ bản về hai kỹ thuật khai phá dữ liệu chính là tiền đề cho việc lựa chọn phương pháp giải bài toán cố vấn học tập.

CHƯƠNG 2

BÀI TOÁN CỐ VẤN HỌC TẬP VÀ ĐẶC TRƯNG BỘ DỮ LIỆU SINH VIÊN ĐẠI HỌC TẠI TRƯỜNG ĐẠI HỌC KINH TẾ QUỐC DÂN

2.1 Giới thiệu chương

Nội dung trong chương này là tóm tắt lại những vấn đề xung quanh công tác cố vấn học tập, quy chế đào tạo đại học hệ chính quy theo tín chỉ. Qua quá trình thu thập dữ liệu và tìm hiểu về hệ thống quản lý đào tạo, đã mô tả những đặc trưng của bộ dữ liệu sinh viên trường đại học kinh tế quốc dân. Từ đó đề xuất phát biểu hai bài toán cố vấn học tập cụ thể. Đưa ra mục tiêu đạt được và ý nghĩa của bài toán đối với các hoạt động cố vấn học tập tại trường Đại học kinh tế quốc dân. Phác thảo sơ đồ giải quyết bài toán. Chương 2 gồm có 5 mục lớn, mục tiếp theo sẽ trình bày về các vấn đề cố vấn học tập theo hình thức đào tạo tín chỉ tại trường Đại học kinh tế quốc dân. Mục 2.3 sẽ phát biểu đề xuất 2 bài toán cố vấn học tập, mục tiêu, ý nghĩa, sơ đồ phác thảo giải 2 quyết bài toán. Mục 2.4 nêu lên đặc trưng dữ liệu sinh viên, hệ thống quản lý đào tạo của trường đại học kinh tế quốc dân. Cuối cùng là tổng kết chương.

2.2 Những vấn đề về cố vấn học tập theo hình thức đào tạo tín chỉ tại trường Đại học Kinh tế Quốc dân

2.2.1 Tổ chức hệ thống cố vấn học tập

- Cố vấn học tập (CVHT) *(theo [2])*

Là người tư vấn và hỗ trợ sinh viên phát huy tối đa khả năng học tập, rèn luyện và nghiên cứu khoa học, lựa chọn đăng ký học phần phù hợp để đáp ứng mục tiêu tốt nghiệp và khả năng tìm được việc làm sau khi ra trường, theo dõi quá trình học tập, rèn luyện của sinh viên nhằm giúp sinh viên điều chỉnh kịp thời hoặc đưa ra một lựa chọn đúng trong quá trình học tập, đồng thời quản lý, hướng dẫn và chỉ đạo lớp được phân công phụ trách.

- CVHT được tổ chức theo mô hình 2 cấp:

- *CVHT chuyên trách*: Là những cán bộ làm nhiệm vụ CVHT thuộc Phòng Thanh tra, Đảm bảo CLGD và Khảo thí;

- *CVHT kiêm nhiệm:* Là những cán bộ, giảng viên làm nhiệm vụ CVHT ở Khoa/Viện chuyên ngành, được lựa chọn từ Trưởng/Phó Bộ môn, trợ lý Khoa/Viện và một số giảng viên có kinh nghiệm .

- *CVHT chuyên trách và CVHT kiêm nhiệm* có mối quan hệ hỗ trợ nhau cùng thực hiện nhiệm vụ.

2.2.2 Chức năng của cố vấn học tập

- Tư vấn, hỗ trợ thông tin và định hướng quá trình học tập, rèn luyện, thực hiện quyền và nghĩa vụ của sinh viên.

- Theo dõi quá trình học tập và rèn luyện của sinh viên.

- Đề xuất phương án xử lý đối với các tình huống phát sinh trong quá trình đào tạo sinh viên.

- Tham mưu cho lãnh đạo Trường, Khoa/Viện chuyên ngành các vấn đề liên quan đến công tác GD&ĐT, NCKH của sinh viên và đào tạo theo nhu cầu xã hội.

2.2.3 Nhiệm vụ của cố vấn học tập

2.2.3.1 Nhiệm vụ chung của CVHT chuyên trách và kiêm nhiệm

a. Tư vấn về học tập và rèn luyện

1) Hướng dẫn sinh viên nắm vững các quy chế đào tạo của Bộ GD&ĐT và các quy định về đào tạo của Nhà trường.

2) Nắm danh sách sinh viên các lớp được giao làm CVHT, từ đó, hướng dẫn sinh viên xây dựng kế hoạch học tập riêng trên cơ sở lựa chọn các học phần được Nhà trường tổ chức giảng dạy từng học kỳ, vừa phù hợp với yêu cầu của chuyên ngành đào tạo, vừa phù hợp với năng lực, sở thích, điều kiện sinh hoạt, hoàn cảnh kinh tế của từng sinh viên.

3) Tư vấn cho sinh viên về chương trình học tập: mục tiêu, nội dung...và cách lựa chọn các học phần.

4) Tư vấn cho sinh viên đăng ký các học phần của từng học kỳ theo chuyên ngành đào tạo và hướng dẫn cho sinh viên phấn đấu để hoàn thành khối lượng học tập đã đăng ký. Tư vấn cho sinh viên cách thức xây dựng kế hoạch học tập cá nhân cho toàn khóa học với tiến độ mục tiêu (*học nhanh hay chậm)*, và tư vấn kế hoạch cụ thể từng học kỳ.

5) Tư vấn cho sinh viên sử dụng phần mềm quản lý đào tạo.

6) Tư vấn và hướng dẫn cho sinh viên về phương pháp học tập và nghiên cứu khoa học; hướng dẫn, khuyến khích, tạo điều kiện cho sinh viên tham gia các hoạt động học tập và nghiên cứu khoa học; hướng dẫn sinh viên giải quyết những khó

khăn trong quá trình học tập và NCKH.

7) Thường xuyên theo dõi kết quả học tập của sinh viên. Nhắc nhở sinh viên khi thấy kết quả học tập của họ giảm sút.

8) Thông qua tình hình, kết quả học tập của sinh viên để tư vấn, hướng dẫn sinh viên trong việc đăng ký, điều chỉnh kế hoạch học tập cho phù hợp với năng lực và hoàn cảnh của từng sinh viên.

9) Thảo luận và hướng dẫn sinh viên cách chọn để học thành công song song hai chương trình, học nâng điểm, cách tính điểm học tập và rèn luyện.

10) Phối hợp và hỗ trợ các Khoa/Viện chuyên ngành, các phòng chức năng, các tổ chức ĐTN và HSV của Nhà trường trong việc tổ chức các phong trào, các hoạt động ngoại khóa và tham gia các hoạt động đoàn thể, hoạt động xã hội khác của sinh viên, đồng thời theo dõi, đánh giá toàn diện về học tập và rèn luyện của sinh viên. Tham dự các hội nghị lớp và chi đoàn sinh viên. Nhận xét và tham gia đánh giá rèn luyện cùng với Ban cán sự lớp và Chi đoàn sinh viên.

b. Tư vấn trong lĩnh vực khác

1) Hướng dẫn sinh viên tham gia các hoạt động ngoại khóa và thực hiện các nội quy sinh hoạt trong Trường.

2) Góp ý cho sinh viên về các vấn đề xã hội như rèn luyện bản thân, xây dựng các mối quan hệ và các vấn đề về nghề nghiệp như đặc tính nghề nghiệp, môi trường làm việc, thị trường lao động, sự lựa chọn nghề nghiệp và cơ hội thăng tiến trong tương lai.

2.2.3.2 Nhiệm vụ cụ thể

Ngoài các nhiệm vụ quy định ở trên, CVHT chuyên trách và CVHT kiêm nhiệm còn phải thực hiện những nhiệm vụ cụ thể sau:

a. CVHT chuyên trách

1) Đầu mối xây dựng, kiện toàn hệ thống, hoàn thiện quy trình làm việc và vận hành của bộ máy CVHT để hệ thống này hoạt động ngày càng hiệu quả hơn, đảm bảo là kênh liên hệ của sinh viên với các Khoa/Viện chuyên ngành, các phòng chức năng và các bộ phận liên quan của Nhà trường.

2) Là đầu mối liên lạc giữa hệ thống CVHT kiêm nhiệm với các phòng chức năng của Trường trong việc xử lý các vấn đề liên quan.

3) Tập hợp và chuẩn bị tài liệu cho việc tư vấn, hướng dẫn đội ngũ CVHT kiêm nhiệm. Phối hợp với đội ngũ CVHT kiêm nhiệm trong việc thực hiện nhiệm vụ.

4) Nắm vững phần mềm quản lý đào tạo để hỗ trợ công tác CVHT.

5) Tổ chức các khóa tập huấn về nghiệp vụ CVHT cho các CVHT.

6) Phối hợp với CVHT kiêm nhiệm trong việc tổ chức họp lớp sinh viên đầu kỳ và cuối kỳ.

7) Phối hợp với các Khoa/Viện chuyên ngành, các phòng chức năng trong Trường để hỗ trợ và tư vấn, tạo điều kiện cho sinh viên học tập. Thường xuyên trao đổi với Khoa/Viện chuyên ngành về tình hình sinh viên, tổ chức các hoạt động hỗ trợ cho sinh viên, giải quyết chế độ, chính sách cho sinh viên.

8) Giới thiệu cho sinh viên địa chỉ (*cán bộ, đơn vị*) để được nhận tư vấn.

9) Biên soạn và hoàn thiện tài liệu hướng dẫn sinh viên và các biểu mẫu.

10) Định kỳ (*cuối học kỳ, cuối năm học*) hoặc đột xuất báo cáo Nhà trường về sinh viên và lớp sinh viên.

b. CVHT kiêm nhiệm

1) Làm đầu mối giải quyết trực tiếp các công việc liên quan đến học tập và rèn luyện của sinh viên mà mình phụ trách.

2) Chủ trì tổ chức họp đầu và cuối kỳ với lớp sinh viên mà mình phụ trách

3) Tham dự các cuộc họp của Hội đồng cấp Khoa/Viện liên quan đến sinh viên lớp mình làm CVHT.

4) Thường xuyên liên hệ với CVHT chuyên trách để được hỗ trợ các điều kiện trong việc thực hiện chức năng và nhiệm vụ được giao.

5) Quy định thời gian tiếp sinh viên tại Khoa/Viện chuyên ngành để họ có thể thường xuyên đến nhận ý kiến tư vấn.

6) Cuối mỗi học kỳ, báo cáo tình hình học tập, rèn luyện của sinh viên với BCN Khoa/Viện chuyên ngành để phục vụ công tác quản lý. Nắm rõ tình hình của sinh viên thuộc diện yếu kém, thông báo cho gia đình biết để phối hợp với Khoa/Viện và Nhà trường trong việc giáo dục và quản lý sinh viên.

2.3 Bài toán cố vấn học tập tại trường Đại học kinh tế quốc dân

2.3.1 Vấn đề thực tế xung quanh bài toán

Từ những vấn đề về cố vấn học tập cho sinh viên đang theo học theo hình thức đào tạo tín chỉ tại các trường đại học nói chung và trường Đại học kinh tế quốc dân nói riêng. Cùng với những quy chế đào tạo theo hình thức mới, hàng năm sinh viên mới nhập học thường khó khăn trong việc thích nghi với hình thức đào tạo này. Bộ phận cố vấn học tập của trường phải có rất nhiều phương pháp để cố vấn cho sinh viên trên các vấn đề đã nói ở trên như; hướng dẫn sinh viên xây dựng kế hoạch học tập riêng trên cơ sở chương trình học từng chuyên ngành được Nhà trường tổ chức giảng dạy từng học kỳ, vừa phù hợp với yêu cầu của chuyên ngành đào tạo, vừa phù hợp với năng lực, sở thích, hoàn cảnh kinh tế của từng sinh viên. Tư vấn cho sinh viên về chương trình học tập: mục tiêu, nội dung...và cách lựa chọn các học phần, đặc biệt là các học phần lựa chọn của ngành và chuyên ngành. Tư vấn cho sinh viên đăng ký các học phần của từng học kỳ theo chuyên ngành đào tạo và hướng dẫn cho sinh viên phấn đấu để hoàn thành khối lượng tín chỉ đã đăng ký với kết quả tốt. Tư vấn cho sinh viên cách thức xây dựng kế hoạch học tập cá nhân cho toàn khóa học với tiến độ mục tiêu (*học nhanh hay chậm)*, và tư vấn kế hoạch cụ thể từng học kỳ.

Vấn đề về quy định đào tạo tín chỉ tại trường đại học kinh tế quốc dân, sinh viên thuộc 45 chuyên ngành phải hoàn thành tất cả 126, 127,128, 129 hoặc 130 tín chỉ tùy từng chuyên ngành, trước mỗi kỳ học sinh viên ngoài việc tự chủ động đăng ký học phần bắt buộc ra thì còn phải đăng ký học 11, 10, 9, 6 hoặc 7, 8 học phần tự chọn tùy từng chuyên ngành. Trên mỗi một tổ hợp tự chọn bao gồm 4, 3 hoặc 2 học phần trong đó sinh viên phải tự chọn 1 học phần (một học phần có 2 hoặc 3 tín). Vấn đề là khi lựa chọn học phần tự chọn sinh viên thường băn khoăn không biết với tổ hợp này thì sẽ nên đăng ký môn học nào, tổ hợp kia nên đăng ký môn học khác, hoặc chọn học phần có kiến thức bổ trợ cho nhau, phù hợp năng lực sở thích, hoặc có lợi để học song ngành. Mỗi sinh viên thường có những lựa chọn linh hoạt khác nhau, họ thường tìm đến với cố vấn học tập để tìm câu trả lời, hoặc tham khảo các anh chị khóa trước, do đó họ thường đăng ký không dựa vào quy tắc nào, có thể hỏi bạn bè đã đăng ký trước, có nhiều trường hợp chọn môn học không phù hợp dẫn đến ảnh hưởng tiến độ và tình trạng tốt nghiệp của sinh viên, những học phần nên học trước thì lại đăng ký sau, đăng ký quá nhiều học phần ảnh hưởng đến kết quả học tập.

Bộ phận cố vấn học tập trước mỗi kỳ, họ thường phải dựa vào nhiều thông tin để cố vấn cho mỗi sinh viên của mỗi chuyên ngành, họ phải trả lời nhiều sinh viên để tổng hợp được nhu cầu đăng ký, hoặc dựa vào các báo cáo. Họ gặp rất nhiều khó khăn trong việc tổng hợp. Trước vấn đề này bài toán tư vấn cho sinh viên thuộc các chuyên ngành khác nhau đăng ký các học phần tự chọn phù hợp là vô cùng quan trọng. Sinh viên có tư vấn kịp thời, từ đó sinh viên có những quyết định hợp lý đảm bảo thời gian học tập của mình.

2.3.2 Phát biểu bài toán

Xuất phát từ những vấn đề tồn tại trong hệ đào tạo tín chỉ, vấn đề cố vấn học tập cho sinh viên như đã nêu ở trên, hai bài toán được đề xuất phát biểu như sau:

***Bài toán 1**:* Cố vấn cho sinh viên đăng ký các học phần tự chọn theo các tổ hợp trên định hướng chuyên ngành. Vào đầu mỗi kỳ học khi phòng đào tạo thông báo mở các lớp học phần, sinh viên thường phải tự sắp xếp thời khóa biểu của mình và chủ động đăng ký môn học. Họ thường gặp khó khăn trong việc lựa chọn, băn khoăn không biết nên học môn nào trong một tổ hợp, và đa số phải tham khảo ý kiến của cán bộ cố vấn học tập để xin tư vấn, định hướng lựa chọn các học phần tự chọn trong kỳ học đó sao cho phù hợp với năng lực sở thích và quy chế đào tạo.

***Bài toán 2**:* Phân lớp, dự báo cho sinh viên có khả năng ra trường đúng thời hạn hay không đúng hạn. Theo thống kê của phòng đào tạo, hàng năm có từ 10 đến 15 phần trăm sinh viên ra trường muộn. Để giải quyết vấn đề này thì vai trò của cán bộ cố vấn học tập là phải đưa ra quyết định cảnh báo học tập kịp thời. Sau khi kết thúc mỗi kỳ học, cán bộ cố vấn học tập thường phải theo dõi kết quả học tập của sinh viên, tổng hợp kết quả từng kỳ học. Nếu phát hiện những sinh viên chưa đủ số tín chỉ và xếp loại học lực yếu thì phải thông báo cho sinh viên biết sớm, giúp sinh viên nhanh chóng điều chỉnh kế hoạch và thái độ học tập, bổ sung đủ tín chỉ, cải thiện điểm thì mới hoàn thành tốt nghiệp đúng thời hạn theo quy chế đào tạo của nhà trường.

2.3.3 Mục tiêu và ý nghĩa của bài toán

Mục tiêu, ý nghĩa bài toán 1: Làm thế nào có thêm nhiều cơ sở thông tin giúp cho cán bộ cố vấn học tập dựa vào đó để làm phương tiện cố vấn, giải quyết những vấn đề thực tế của sinh viên. Bằng phương pháp khai phá dữ liệu dựa trên luật kết hợp, tìm ra mối quan hệ kết hợp giữa các môn học (môn học nào hay được sinh viên kết hợp đăng ký cùng nhau), kết quả sinh ra được một tập luật kết hợp giữa

các môn học, luật này mạnh và có ích với khả năng xảy ra cao. Ý nghĩa từ bảng tập luật đó giúp cán bộ cố vấn trả lời hai câu hỏi của sinh viên.

- Nếu đăng ký học phần A ở tổ hợp này, và học phần C ở tổ hợp kia, thì thường hay đăng ký học phần nào ở tổ hợp khác, theo từng chuyên ngành khác nhau.
- Trong các tổ hợp học phần lựa chọn, học phần lựa chọn nào hay được chọn đăng ký cùng với nhau.

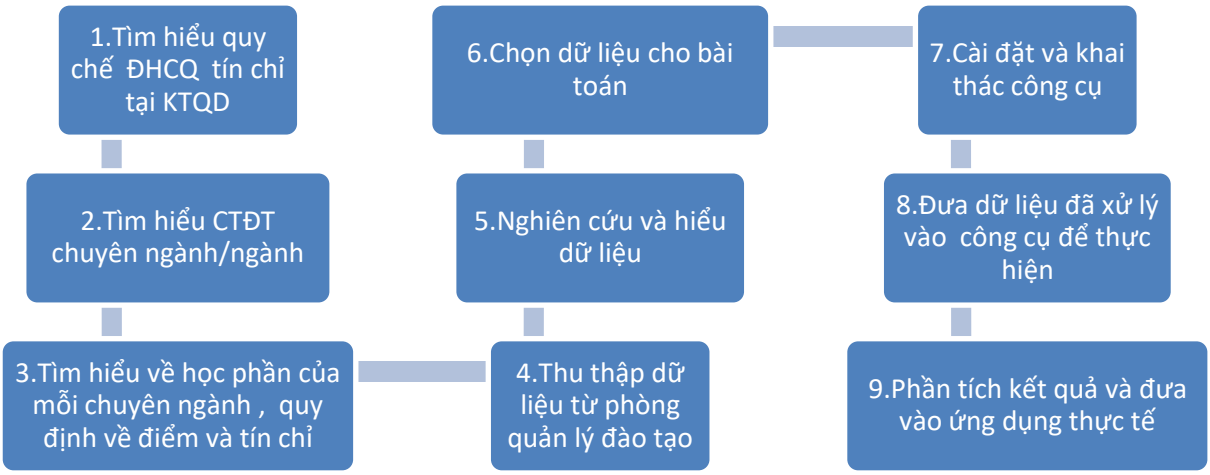
Từ đó cán bộ cố vấn học tập sẽ nắm được xu hướng lựa chọn học phần, phân tích xem nhu cầu ra sao, cố vấn cho phòng đào tạo điều chỉnh tăng, giảm, thay đổi số lượng lớp học phần cho phù hợp phân bổ chương trình môn học hợp lý cho giáo viên và sinh viên.

Mục tiêu, ý nghĩa bài toán 2: Từ kết quả phân lớp xác định được sinh viên nào đang bị rơi vào tình trạng cảnh báo ra trường không đúng hạn, đúng hạn. Nếu cán bộ cố vấn học tập có những cảnh bảo nhanh chóng và kịp thời thì hàng năm tỉ lệ ra trường muộn sẽ giảm đi rất nhiều. Bằng phương pháp khai phá dữ liệu, phân lớp sinh viên dựa trên cây quyết định. Giúp cán bộ cố vấn học tập trong quá trình theo dõi kết quả học tập của sinh viên chính xác hơn, cảnh báo, dự báo tiến trình học tập cho sinh viên sau mỗi kỳ học, góp phần tăng tỷ lệ ra trường đúng hạn của nhà Trường đạt tối đa lên mục tiêu 100% sinh viên ra trường đúng hạn.

Sau khi khai phá dữ liệu bằng kỹ thuật phân lớp dựa vào cây quyết định. Ý nghĩa kết quả sau khi thực hiện phân lớp sinh viên là dựa vào số tín chỉ đã tích lũy và điểm chung bình chung tích lũy của các kỳ học sẽ giúp cán bộ cố vấn học tập có khả năng ra quyết định cảnh báo, dự báo sinh viên A có khả năng rơi vào trường hợp ra trường đúng hạn hay không đúng hạn, nếu không đúng hạn thì sinh viên đó sớm đăng ký học phần bổ sung cho kịp ra trường.

2.3.4 Quy trình giải quyết bài toán

Từ phát biểu và mục tiêu của bài toán cố vấn học tập luận văn đề xuất xây dựng mô hình khai phá dựa vào luật kết hợp và cây quyết định trên công cụ BIDS để thực hiện giải quyết hai bài toán đó theo sơ đồ phác thảo sau.



Hình 2.1 Quy trình giải quyết bài toán

2.4 Đặc trưng dữ liệu sinh viên trường Đại học kinh tế quốc dân

2.4.1 Hệ thống quản lý đào tạo, quản lý sinh viên

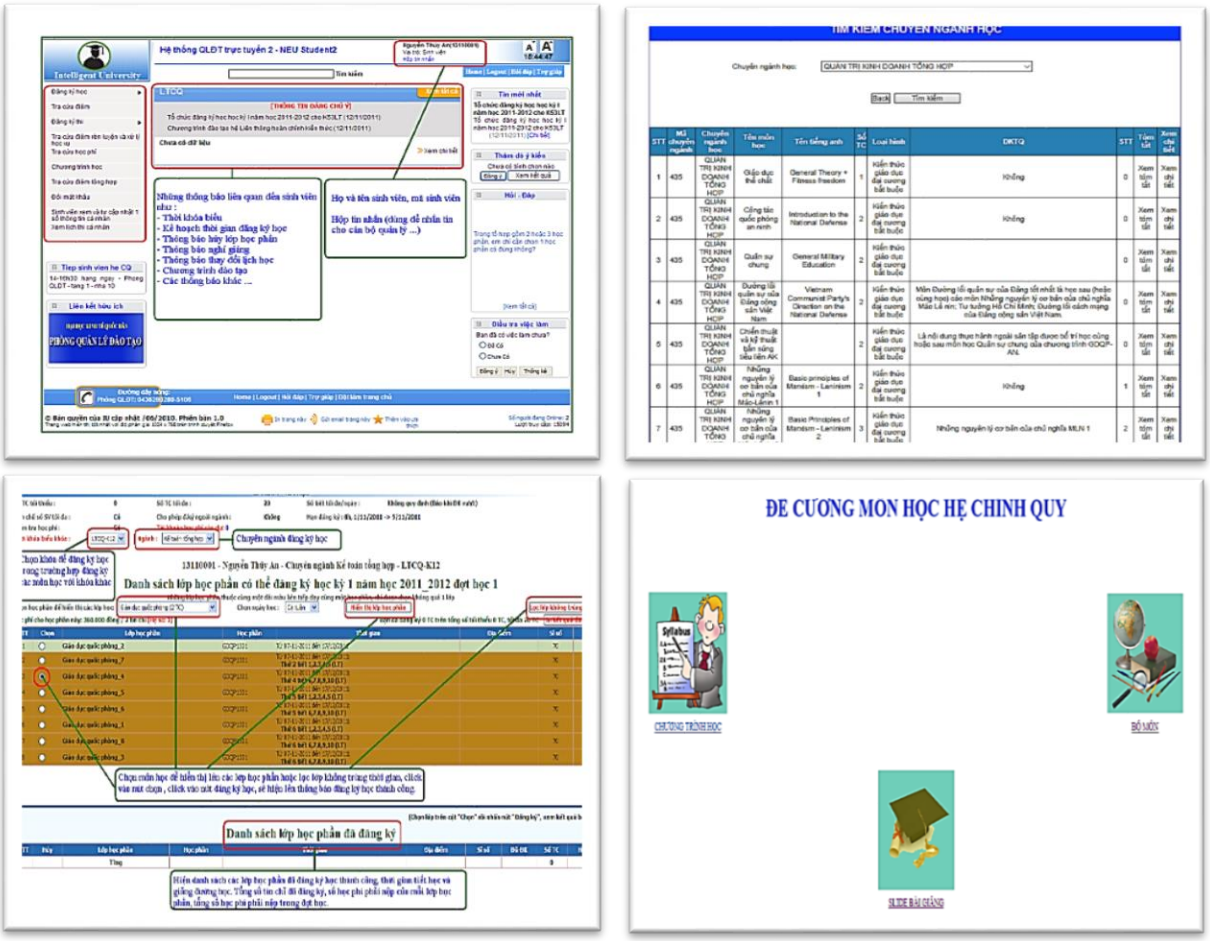
Nhằm nâng cao hiệu quả khai thác thông tin đào tạo đại học, quản lý sinh viên, cung cấp các dịch vụ trực tuyến cũng như đẩy mạnh công tác cải cách hành chính của Trường, hệ thống thông tin quản lý đào tạo (đại học, sau đại học) đã được xây dựng.

Các thông tin đào tạo được đăng tải và cập nhật thường xuyên trên Cổng thông tin điện tử của Trường *(http://www.neu.edu.vn)*. Người truy cập có thể nhận được đầy đủ và cập nhật các thông tin về tuyển sinh (hệ đại học, liên thông, văn bằng II, sau đại học), các thông tin về học bổng (học bổng khuyến khích học tập, học bổng tài trợ, học bổng du học), thông tin giới thiệu việc làm, chương trình đào tạo, kế hoạch học tập, thời khóa biểu, lịch thi, thông tin giáo trình, các quy định, quy chế, thông tin học phí, bảng điểm, kết quả quá trình học tập và xử lý học tập...

Hệ thống quản lý đào tạo (QLĐT) trực tuyến được đưa vào sử dụng bắt đầu từ khi Nhà trường áp dụng hình thức đào tạo theo học chế tín chỉ (năm học 2006 - 2007). Với hình thức đào tạo theo niên chế, sinh viên có thể không cần truy cập vào mạng để tra cứu thông tin (có thể thông qua CVHT hoặc BCS lớp), nhưng với hình thức đào tạo theo học chế tín chỉ, thì công việc này bắt buộc đối với mỗi sinh viên. Hệ thống thông tin này cho phép sinh viên truy cập để đăng ký học phần, đăng ký lớp học, tra cứu điểm (điểm quá trình, điểm thi kết thúc học phần) và theo dõi các thông tin liên quan đến học tập...sau khi sinh viên có tài khoản cá nhân (được cấp sau khi hoàn thành các thủ tục nhập Trường).

Tài khoản cá nhân này sẽ được dùng để truy cập vào hệ thống thông tin khác của Trường như Thư viện... Hệ thống QLĐT của Trường Đại học Kinh tế Quốc dân luôn được cập nhật và chỉnh sửa để ngày càng phù hợp hơn với hình thức đào tạo mới, cho phép giảng viên có thể đăng ký giảng dạy, quản lý lớp, cung cấp thêm các chức năng cho các cố vấn học tập để quản lý lớp sinh viên, theo dõi kết quả, cảnh báo học tập cho từng sinh viên. Ngoài ra, hệ thống cũng là nơi cung cấp đầy đủ thông tin tham khảo về nội dung từng môn học phần để sinh viên dễ dàng có thể lựa chọn theo nhu cầu cá nhân.

Chương trình đào tạo Trường ĐH Kinh tế Quốc dân gồm có 47 chuyên ngành thuộc 22 nhóm ngành khác nhau. Mỗi năm tuyển sinh khoảng 4000 sinh viên hệ chính quy.



Hình 2.2 Hệ thống quản lý đào tạo

2.4.2 Mô tả một phần cơ sở dữ liệu quản lý sinh viên dựa trên những thông tin đã thu thập

Trường Kinh tế quốc dân sử dụng hệ quản trị cơ sở dữ liệu Oracle. Mô tả bằng cơ sở dữ liệu quan hệ với cấu trúc các bảng khác nhau:

- Sinh viên : Mã sinh viên, tên sinh viên, ngày sinh, giới tính, địa chỉ, quê quán, số điện thoại,…

- Khóa đào tạo: Mã khóa, tên khóa…

- Bộ môn: Mã bộ môn, tên bộ môn…

- Bảng điểm: Điểm lần 1, lần 2, điểm đạt lớn nhất,…

- Học phần: Mã học phàn, tên học phần, số tín chỉ,…

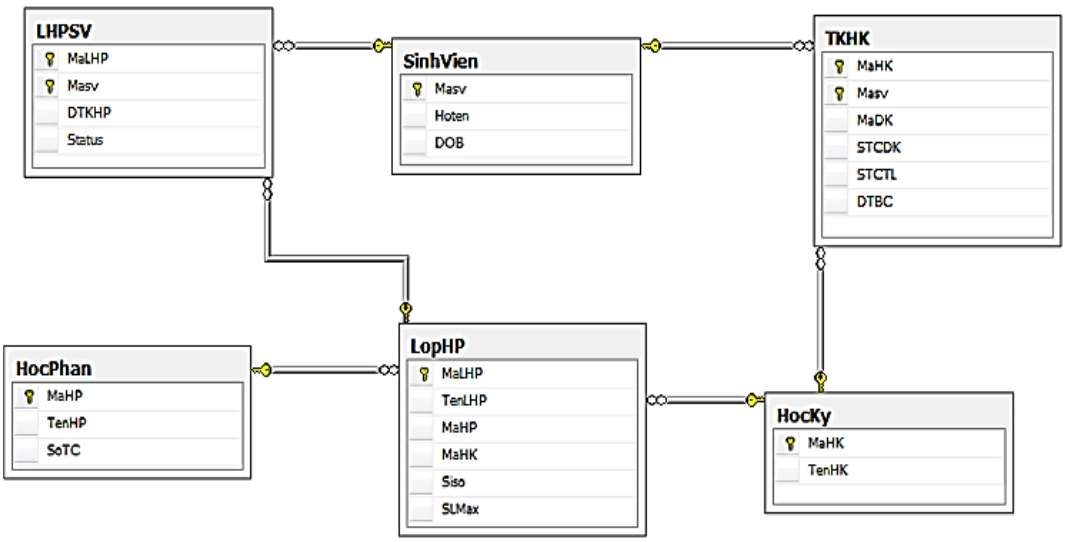
- Lớp học phần sinh viên: Mã lớp học phần, Mã sinh viên…

- Ngành: Mã ngành, tên ngành,…

- Chuyên ngành: Mã chuyên ngành, tên chuyên ngành,…

- Học kỳ: Mã học kỳ, tên học kỳ…

- Tổng kết học kỳ: Mã học kỳ, mã sinh viên, số tín chỉ đăng ký, số tín chỉ tích lũy, điểm chung bình chung, điểm chung bình chung tích lũy…



Hình 2.3 Cơ sở dữ liệu quản lý sinh viên

Phòng đào tạo là nơi có quyền cao nhất trong việc quản lý, lưu trữ, xử lý thông tin liên quan đến điểm sinh viên, quá trình học, xét tốt nghiệp và ra trường.

2.5 Tổng kết chương 2

Qua nội dung đã trình bày trong chương 2 tác giả đã hiểu được vấn đề chính trong cố vấn học tập, hiểu về quy chế đào tạo theo tín chỉ. Tầm quan trọng của việc cán bộ cố vấn thường xuyên phải cố vấn học tập cho sinh viên trước và sau mỗi kỳ học. Từ 2 bài toán đã đề xuất là tư vấn chọn môn học theo tổ hợp và phân lớp dự báo khả năng sinh viên ra trường đúng hạn hay không. Đặt mục tiêu và ý nghĩa rõ ràng đó là góp phần có thêm nhiều cơ sở thông tin để giúp ích cho bộ phận CVHT. Mô tả lại được về cơ sở dữ liệu quan hệ sinh viên qua dữ liệu đã thu thập được. Định hình và đưa ra công việc cần phải làm tiếp theo trong phần thực nghiệm qua sơ đồ phác thảo. Đó là phải xử lý dữ liệu, biến đổi, lọc bỏ dư thừa, trùng lặp sao cho phù hợp với bài toán và phương pháp khai phá. Nắm được những vấn đề cốt lõi, chuẩn bị dữ liệu đầy đủ cho thực nghiệm giải bài toán đã đề xuất.

CHƯƠNG 3

ỨNG DỤNG THỬ NGHIỆM GIẢI BÀI TOÁN CỐ VẤN HỌC TẬP TẠI TRƯỜNG ĐẠI HỌC KINH TẾ QUỐC DÂN

3.1 Giới thiệu chương

Chương 3 sẽ giới thiệu sơ lược một số công cụ khai phá dữ liệu, quy trình thực hiện khai phá dữ liệu giải bài toán cố vấn học tập. Nội dung chủ yếu là trình bày các lần thực nghiệm tiến hành giải 2 bài toán, từ dữ liệu thu thập đến biến đổi dữ liệu, tiến hành khai phá, giải thích kết quả đạt được có ý nghĩa với bài toán đề xuất. Chương gồm 7 mục chính, mục tiếp theo sẽ trình bày qua về công cụ khai phá dữ liệu. Mục 3.3 là quy trình thực hiện khai phá dữ liệu sinh viên và phát hiện tri thức với bài toán cố vấn học tập tại Trường Đại học Kinh tế Quốc dân. Mục 3.4 quá trình thực nghiệm khai phá dữ liệu trên công cụ BIDS giải bài toán 1 bằng kỹ thuật khai phá luật kết hợp, nêu ý nghĩa kết quả đạt được. Mục 3.5 quá trình thực nghiệm KPDL giải quyết bài toán 2 bằng kỹ thuật phân lớp dựa vào cây quyết định, làm rõ ý nghĩa của kết quả đạt được với bài toán đề xuất. Mục 3.6 là đưa ra một số đề xuất kiến nghị sau khi thực nghiệm. Cuối cùng là tổng kết chương.

3.2 Giới thiệu một số công cụ khai phá dữ liệu và phát hiện tri thức

3.2.1 Weka

Weka (Waikato Environment for Knowledge Analysis), là bộ phần mềm học máy, mã nguồn mở, do đại học Waikato phát triển bằng Java, nhằm phục vụ cho các nhiệm vụ chuyên về khai phá dữ liệu. Weka chứa các công cụ phục vụ cho tiền xử lý dữ liệu, phân loại, hồi quy, phân cụm, các luật liên quan và trực quan hóa. Nó cũng phù hợp cho việc phát triển, xây dựng các mô hình học máy và có khả năng chạy được trên nhiều hệ điều hành khác nhau như Windows, Mac, Linux.3.1.2.

Các tính năng chính: Những tính năng vượt trội trong Weka có thể kể đến là:

- Mã nguồn mở
- Hỗ trợ các thuật toán học máy (machine learning) và khai phá dữ liệu
- Trực quan hóa, dễ dàng xây dựng các ứng dụng thực nghiệm
- Do sử dụng JVM nên Weka độc lập với môi trường

3.2.2 Ngôn ngữ R

Theo một nhà nghiên cứu, Ngôn ngữ lập trình R đang nhanh chóng trở thành ngôn ngữ phổ biến trong các gói ngôn ngữ dữ liệu truyền thống như SPSS, SAS và MATLAB, ít nhất là theo một nhà phân tích dữ liệu về ngôn ngữ lập trình.

“Trong suốt mùa hè vừa qua, R là phần mềm phân tích được sử dụng nhiều nhất trong các bài viết học thuật chuyên sâu, chấm dứt một kỷ nguyên 16-năm của SPSS”, ông Robert Muenchen viết trong một bài blog gần đây, tổng kết thống kê của ông.

Muenchen đánh giá tính phổ biến của các gói phần mềm dữ liệu bằng cách theo dõi tần suất người sử dụng đăng tải các nghiên cứu khoa học được công bố và số lượng người nhận xét gói phần mềm trong các thảo luận trên forum, blog, danh sách công việc và các nguồn khác.

Trong khảo sát này của ông Muenchen, các nhà nghiên cứu tiếp tục sử dụng các gói phần mềm truyền thống trong hầu hết công việc của họ, có thể kể đến như các gói của SAS và MATLAB, cũng như gói SPSS của IBM. SPSS dẫn đầu với hơn 75.000 trích dẫn trong các bài báo khoa học được liệt kê trong kết quả tìm kiếm của Google Scholar. SAS theo sau, đứng thứ 2 với 40.000 trích dẫn. R được sử dụng hơn 20.000 dự án nghiên cứu.

Ngoài ra, khi Muenchen tìm con số trích dẫn từ năm 1995, ông nhận ra rằng SPSS đã giảm kể từ năm 2007. SAS cũng theo chiều hướng của SPSS, đạt đỉnh hồi năm 2008. Ngược lại, R tăng rất nhanh, nhanh hơn cả các gói như Statistica và Stada. Ông Muenchen đề cập: “Xu hướng SPSS giảm và R tăng trong suốt quãng mùa hè vừa qua cho thấy R đang trở thành ngôn ngữ hàng đầu trong các gói phân tích dữ liệu được sử dụng trong các bài báo khoa học. Vì quá trình đăng tải các bài báo in xuất bản trước một thời gian trước khi đưa lên mạng, tạo chỉ mục tìm kiếm… nên chúng tôi chưa thể xác định chính xác điều gì sẽ xảy ra trong năm tới.”

R là ngôn ngữ lập trình chức năng, nguồn mở, được thiết kế chuyên cho điện toán dữ liệu và đồ hoạ. Muenchen là nhà thống kê, quản lý mảng hỗ trợ bộ phận điện toán tại đại học Tennessee, Mỹ, ngoài việc xác nhận tính phổ biến của R, ông cũng là giáo viên giảng dạy R trên danh nghĩa Revolution Analytics. Ông còn được cộng đồng công nhận là chuyên gia về phân tích điện toán, viết mã cho SAS, SPSS và nhiều gói R khác. Ông cũng từng làm việc trong ban cố vấn của SAS, SPSS trước khi IBM mua lại hồi năm 2009.

Theo IDC, ngôn ngữ R trở nên phổ biến một phần là vì nó là nguồn mở, miễn phí và các nhà nghiên cứu có thể tải nó về để bắt đầu một dự án nào đó mà không phải tốn tiền.Trong nghiên cứu của Muenchen, ông không phân biệt giữa các phiên bản khác nhau của R, có thể đó là phiên bản nguồn mở hoặc phiên bản dành cho doanh nghiệp của Revolution Analytics, hoặc là bản nguồn mở của R Project.

Cũng có một số dấu hiệu khác cho thấy tính phổ biến của R. Nhiều đăng tải tìm việc trên Indeed.com yêu cầu thành thạo R nhiều hơn so với SPSS, mặc dù vẫn có vài nhà tuyển dụng cần đến SAS. Số lượng sách và forum thảo luận về R cũng nhiều hơn SAS và SPSS.

3.2.3 SQL Datamining

3.2.3.1 Giới thiệu

Giới thiệu công cụ xây dựng mô hình khai phá dữ liệu Business Intelligence Development Studio (BIDS) của Microsoft Sql Server 2008

Nhằm xác lập chỗ đứng trong thị trường giải pháp thông tin doanh nghiệp (Business Intelligence - BI), Microsoft SQL Server 2008 cung cấp các công cụ có khả năng quản lý báo cáo và phân tích, khai phá dữ liệu đủ mọi cấp độ, tích hợp chặt chẽ với Microsoft Office cùng với cơ sở hạ tầng mạnh, linh hoạt và có thể mở rộng, cho phép đưa thông tin doanh nghiệp đến tất cả nhân viên, giúp ra quyết định nhanh hơn và tốt hơn. Giải pháp BI của Microsoft được xây dựng trên nền tảng dữ liệu, đồng thời cung cấp các công cụ mạnh mẽ cho phép người dùng cuối truy cập và phân tích thông tin doanh nghiệp. Trung tâm của giải pháp này là một nền tảng dịch vụ dữ liệu hoàn chỉnh có khả năng.

- Hợp nhất việc lưu trữ và truy cập cho tất cả dữ liệu
- Xây dựng và quản lý các giải pháp BI phức tạp
- Mở rộng phạm vi giải pháp BI đến tất cả nhân viên

Một số giải phá kỹ thuật khai phá dữ liệu:

SQL Server Analysis Services cung cấp công cụ phân tích và khai thác dữ liệu dựa trên cơ sở 5 giải thuật Data Mining sau:

+Thuật toán kết hợp (Assocication Algorithm) <dùng trong luận văn>

+Thuật toán phân loại (Microsoft Decision Trees) <dùng trong luận văn>

+Thuật toán phân đoạn (Segmentation Algorithm)

+Thuật toán phân tích chuỗi (Sequence Analysis Allgorithm)

+Thuật toán hồi quy (Regression Algorithm)

3.2.3.2 Thuật toán kết hợp trong công cụ (Assocication Algorithm)

The Microsoft Association cũng thuộc về họ các thuật toán tìm luật kết hợp theo thuật toán Apriori tức là việc tìm các luật kết hợp sẽ gồm hai pha chính là tìm tập các mục chọn thường xuyên sau đó dùng tập các mục chọn thường xuyên để sinh ra các luật kết hợp. Ngoài ra còn có một khái niệm quan trọng khác liên quan trực tiếp đến việc sử dụng thuật toán luật kết hợp.

Độ quan trọng (I): Độ quan trọng của một tập các mục chọn được định nghĩa như sau: $I(\{A,B\}) = P(A,B)/(P(A)*P(B))$

Nếu I = 1 thì A và B là hai mục chọn độc lập. Từ việc mua sản phẩm A và việc mua sản phẩm B là hai sự kiện độc lập.

Nếu I<1 thì A và B có mối liên quan với nhau một cách tiêu cực. Tức là khi khách hàng mua sản phẩm A thì không có khả năng anh ta sẽ mua sản phẩm B.

Nếu I>1 thì A và B có mối liên quan với nhau một cách tích cực. Tức là khi khách hàng mua sản phẩm A thì khả năng anh ta sẽ mua sản phẩm B.

Trong thuật toán kết hợp Microsoft còn sử dụng khái niệm xác suất (Probability) thay cho độ tin cậy (Confidence). Ngoài ra còn có một số danh sách tham số:

+ Minimum_Support: là một tham số giới hạn. Nó xác định tần suất tối thiểu cho tập các mục chọn, nếu tập các mục chọn có tần suất lớn hơn hoặc bằng Minimum_Support thì tập đó là thường xuyên. Minimum_Support có miền giá trị từ 0 đến 1, giá trị mặc định của nó là 0.03. Nếu Minimum_Support được thiết lập với giá trị lơn hơn 1 lúc đo ta hiểu Minimum_Support chính là số lần xuất hiện của tập các mục chọn

+ Maximum_Support: là một tham số giới hạn. Nó xác định tần suất tối đa cho các mục chọn thường xuyên. Maximum_Support có miền giá trị từ 0 đến 1, giá trị mặc định là 0,03. Nếu Maximum_Support được thiết lập giá trị lớn hơn 1 lúc đó ta hiểu Maximum_Support chính là số lần xuất hiện của tập các mục chọn.

+ Minimum_Probability: là một tham số giới hạn. Nó xác định xác suất tối thiểu cho một luật kết hợp. Miền giá trị của nó từ 0 đến 1, giá trị mặc định là 0,04.

+ Minimum_Importance: là tham số giới hạn cho các luật kết hợp. Các luật với độ quan trọng nhỏ hơn Minimum_Importance sẽ bị loại.

+ Maximum_Itemset_Size: xác định kích thước tối đa của tập các mục chọn. Giá trị mặc định là 0, tức không có giới hạn về kích thước của tập các mục chọn

+ Minimum_Itemset_Size: xác định kích thước tối thiểu của tập các mục chọn. Giá trị mặc định là 0.

+ Maximum_Itemset_Count: xác định số lượng tối đa của tập các mục chọn. Nếu không được xác định giá trị, thuật toán sẽ sinh ra tất cả tập các mục chọn dựa vào tham số Minimum_Support.

+ Optimized_Prediction_Count: được sử dụng để số lượng các mục chọn để nghị cho việc dự báo được yêu cầu bởi các truy vấn. Giá trị mặc định là 2.

3.2.3.3 Thuật toán phân loại trong công cụ (Classification Algorithm)

Dự đoán ra một hoặc nhiều giá trị biến rời rạc, dựa trên các thuộc tính khác của tập dữ liệu. Điển hình là thuật toán cây quyết định – *Microsoft Decision Trees Algorithm***.**

Thuật toán Microsoft Decision Tree hỗ trợ cả việc phân loại và hồi quy. Sử dụng thuật toán này có thể dự đoán cả các thuộc tính rời rạc và liên tục. Trong việc xây dựng mô hình, thuật toán này sẽ khảo sát sự ảnh hưởng của mỗi thuộc tính trong tập dữ liệu và kết quả của thuộc tính dự đoán.

Sau đó sẽ sử dụng các thuộc tính input để tạo thành 1 nhóm phân hoá gọi là các node. Khi các 1 node mới được thêm vào mô hình thì 1 cấu trúc cây sẽ được thiết lập. Node đỉnh của cây miêu tả sự phân tích của các thuộc tính dự đoán thông qua các mẫu. Mỗi node thêm vào sẽ được tạo ra dựa trên sự sắp xếp các trường của thuộc tính dự đoán, để so sánh với các dữ liệu input. Nếu 1 thuộc tính input được coi là nguyên nhân của thuộc tính dự đoán thì 1 node mới sẽ thêm vào mô hình. Mô hình tiếp tục phát triển cho đến lúc không còn thuộc tính nào, tạo thành 1 sự phân tách (split) để cung cấp 1 dự báo hoàn chỉnh thông qua các node đã tồn tại. Mô hình đòi hỏi tìm kiếm 1 sự kết hợp giữa các thuộc tính, nhằm thiết lập 1

sự phân phối không cân xứng giữa các trường trong thuộc tính dự đoán. Vì vậy, nó cho phép dự đoán kết quả của thuộc tính dự đoán 1 cách tôt nhất.

Đối với thuộc tính rời rạc, thuật toán đưa ra các dự đoán dựa trên các mối quan hệ giữa các cột nhập vào trong dataset. Nó sử dụng các giá trị, trạng thái, các cột của chúng để dự đoán trạng thái cột mà bạn chỉ định hay dự đoán. Đặc biệt, thuật toán nhận biết các cột nhập vào tương quan với cột dự đoán. Ví dụ, trong một kịch bản, để dự đoán những khách hàng nào có khả năng mua xe đạp, nếu có 9 trong số 10 khách hàng trẻ hơn mua xe đạp, trong khi có 2 trong số 10 khách hàng lớn tuổi hơn mua, thuật toán sẽ suy luận ra tuổi dự đoán tốt cho việc mua xe đạp. Cây quyết định tạo ra các dự đoán dựa trên xu hướng đi tới kết quả cụ thể.

3.3 Quy trình thực hiện khai phá dữ liệu sinh viên và phát hiện tri thức với bài toán cố vấn học tập tại Trường Đại học Kinh tế Quốc dân.

Quy trình cho khai phá dữ liệu với bài toán thực hiện theo các bước sau: hiểu về lĩnh vực đang khai phá (lĩnh vực giáo dục hệ đại học), hiểu về dữ liệu liên quan lĩnh vực đó (quản lý đào tạo sinh viên đại học chính quy theo tín chỉ), chuẩn bị dữ liệu cần thiết liên quan đến đối tượng sinh viên, thiết lập mô hình, đánh giá mô hình, triển khai áp dụng tri thức tìm được.

a. Tìm kiếm thông tin và hiểu về hệ đào tạo đại học chính quy theo hình thức tín chỉ (xác định mục tiêu). Sự hiểu biết về quy chế đào tạo đại học chính quy, quy định về công tác cố vấn học tập, xác định mục đích thực hiện, phát biểu được bài toán, và ý nghĩa kết quả cuối cùng đạt được, chuyển đổi mục đích này vào nhiệm vụ khai thác dữ liệu và xây dựng một kế hoạch triển khai thực hiện sơ bộ để đạt được những mục tiêu đã đề ra.

b. Tìm hiểu về bộ dữ liệu quản lý đào tạo sinh viên thuộc hệ đào tạo đại học chính quy tại đại học kinh tế quốc dân, gia đoạn này bao gồm việc thu thập, quan sát, mô tả và khám phá dữ liệu, xem xét đánh giá chất lượng của dữ liệu, lựa chọn thuật toán và phương pháp giải bài toán.

c. Giai đoạn chuẩn bị dữ liệu liên quan đến phương pháp giải bài toán, việc lựa chọn, dọn dẹp, xây dựng dữ liệu, tránh việc trùng lặp, khuyết thiếu dữ liệu. Toàn bộ dữ liệu được thu thập và xử lý đều lấy từ hệ thống quản lý đào tạo trường kinh tế quốc dân.

d. Thiết lập mô hình và thực hiện, giai đoạn này lựa chọn một công cụ kỹ thuật, trong luận văn sử dụng công cụ BIDS để xây dựng 2 mô hình tìm luật kết

hợp và phân loại bằng cây quyết định, hoặc kết hợp giữa các kỹ thuật sao cho phù hợp. Chạy chương trình cho ra kết quả, sắp xếp và thu gom kết quả.

e. Sắp xếp, mô tả kết quả tìm được, để đảm bảo rằng kết quả từ mô hình đạt được đúng các mục tiêu, ý nghĩa của bài toán, đưa kết quả đạt được ứng dụng trong thực tế.

Chú ý : quy trình trên được thực hiện lặp đi lặp lại để tìm ra tri thức có ích và ý nghĩa.

3.4 Khai phá dữ liệu bằng luật kết hợp giải bài toán 1

Mô tả dữ liệu sử dụng để giải bài toán bằng khai phá luật kết hợp, mô hình và các bước thực hiện khai phá để giải quyết bài toán 1 tìm ra các luật có ích nhằm cố vấn đăng ký học phần tự chọn cho sinh viên như sau.

3.4.1 Từ dữ liệu thô thu thập được

Dữ liệu dùng để xây dựng mô hình là dữ liệu thô được thu thập từ phòng quản lý đào tạo trường Đại học kinh tế quốc dân.

Results | Messages

	TENLOP	TenMon	MALOP	MaMon	HOTEN	MASV	LANHOC	LANTHI	DIEMD...	DiemLo...	SEMESTER	MaCN
1	ENGLISH 1 - L...	ENGLISH 1 - LE...	NNKC1101B(...	NNKC1101B	Lê Thanh Tâm	11123419	1	1	8.2	8.2	2012_2013_2	BDS
2	Những nguyên...	Những nguyên lý...	LLNL1103(21...	LLNL1103	Lê Cúc Phương	11123077	1	1	6.9	6.9	2012_2013_2	BDS
3	Những nguyên...	Những nguyên lý...	LLNL1103(21...	LLNL1103	Trương Thị Mai...	11124790	1	1	7.2	7.2	2012_2013_2	BDS
4	Những nguyên...	Những nguyên lý...	LLNL1103(21...	LLNL1103	Trần Ngọc Anh	11120195	1	1	5.9	5.9	2012_2013_2	BDS
5	Những nguyên...	Những nguyên lý...	LLNL1103(21...	LLNL1103	La Thanh Ngọc	11122820	1	1	6.6	6.6	2012_2013_2	BDS
6	Những nguyên...	Những nguyên lý...	LLNL1103(21...	LLNL1103	Phí Mạnh Hùng	11121630	1	1	6.6	6.6	2012_2013_2	BDS
7	Những nguyên...	Những nguyên lý...	LLNL1103(21...	LLNL1103	Đào Đình Hiếu	11121405	1	1	6.3	6.3	2012_2013_2	BDS
8	Những nguyên...	Những nguyên lý...	LLNL1103(21...	LLNL1103	Trần Thị Huệ	11121616	1	1	6	6	2012_2013_2	BDS
9	Những nguyên...	Những nguyên lý...	LLNL1103(21...	LLNL1103	Lê Thị Phương...	11123574	1	1	6.2	6.2	2012_2013_2	BDS
10	Những nguyên...	Những nguyên lý...	LLNL1103(21...	LLNL1103	Nguyễn Duy Ki...	11121999	1	1	5.9	5.9	2012_2013_2	BDS

Hình 3.1 Dữ liệu thu thập

- Dữ liệu dưới dạng file excel của khóa học đã ra trường, có 4000 sinh viên, 12 thuộc tính, hàng chục nghìn bản ghi.
- Trên mỗi file có 5 sheet, mỗi sheet có hàng nghìn bản ghi là thể hiện của 1 năm học (2014_2015), mỗi năm có 2 kỳ học *(ví dụ kỳ 1 năm 2013_2014 ký hiệu là "2013_2014_1", kỳ 2 là: " 3013_2014_2")*
- Mỗi kỳ học thể hiện thông tin số học phần của một sinh viên đăng ký gồm có học phần tự chọn và bắt buộc trên tất cả các chuyên ngành *(mã sinh viên,*

tên học phần được lặp đi lặp lại trên các dòng, có thể coi đây như là một bộ các giao dịch đăng ký môn học của sinh viên).

- Toàn bộ dữ liệu thể hiện được quá trình đăng ký tất cả các môn học phần của sinh viên trong tất cả các chuyên ngành trong khóa học đó, tách theo từng kỳ học. Mỗi sinh viên học 4 năm, mỗi năm 2 kỳ.

3.4.2 Tiến hành biến đổi dữ liệu theo bài toán 1

- Theo như bài toán 1 đã phát biểu: tìm ra mối quan hệ kết hợp giữa các môn học phần, để cố vấn cho sinh viên lựa chọn các học phần tự chọn thì các thuộc tính sẽ được chọn cho mô hình là: tên học phần, mã sinh viên, tên sinh viên, mã chuyên ngành. Như vậy các thuộc tính còn lại được loại bỏ vì không sử dụng cho bài toán.

- Lọc bỏ những bản ghi không có điểm và mã chuyên ngành *(do sinh viên hủy học phần hoặc chuyển trường)*.

- Lọc bỏ những bản ghi bị trùng lặp *(do lỗi xuất dữ liệu từ hệ thống)*

- Cuối cùng Bộ dữ liệu con thu được dùng trong mô hình khai phá gồm có 2 view như sau: (DanhSachSV, và SV_DangKy_MonHoc)

Results Messages

	MASV	HOTEN	MaCN
1	11120001	Đào Bằng An	DOTHI
2	11120002	Nguyễn Diệu An	KTTN
3	11120003	Bùi Duy An	KDTM
4	11120005	Mạc Đình An	TTMAR
5	11120008	Nguyễn Hữu An	TCQT
6	11120010	Võ Thị Mai An	Chương trì
7	11120011	Nguyễn Mạnh An	KDQT
8	11120012	Nguyễn Phúc An	HTTT

Results Messages

	MASV	HOTEN	TenMon	MaCN
1	11123419	Lê Thanh Tâm	ENGLISH 1 - LEVEL 2	BDS
2	11123077	Lê Cúc Phương	Những nguyên lý cơ bản của chủ nghĩa Mác-Lênin 1	BDS
3	11124790	Trương Thị Mai Hương	Những nguyên lý cơ bản của chủ nghĩa Mác-Lênin 1	BDS
4	11120195	Trần Ngọc Anh	Những nguyên lý cơ bản của chủ nghĩa Mác-Lênin 1	BDS
5	11122820	La Thanh Ngọc	Những nguyên lý cơ bản của chủ nghĩa Mác-Lênin 1	BDS
6	11121630	Phí Mạnh Hùng	Những nguyên lý cơ bản của chủ nghĩa Mác-Lênin 1	BDS
7	11121405	Đào Đình Hiếu	Những nguyên lý cơ bản của chủ nghĩa Mác-Lênin 1	BDS
8	11121616	Trần Thị Huệ	Những nguyên lý cơ bản của chủ nghĩa Mác-Lênin 1	BDS

Hình 3.2 Dữ liệu cho khai phá luật kết hợp

3.4.3 Thực hiện thử nghiệm trên công cụ BIDS

Cũng như qui trình xây dựng các Data Mining Model khác, qui trình xây dựng mô hình khai phá luật kết hợp với BIDS theo các bước sau:

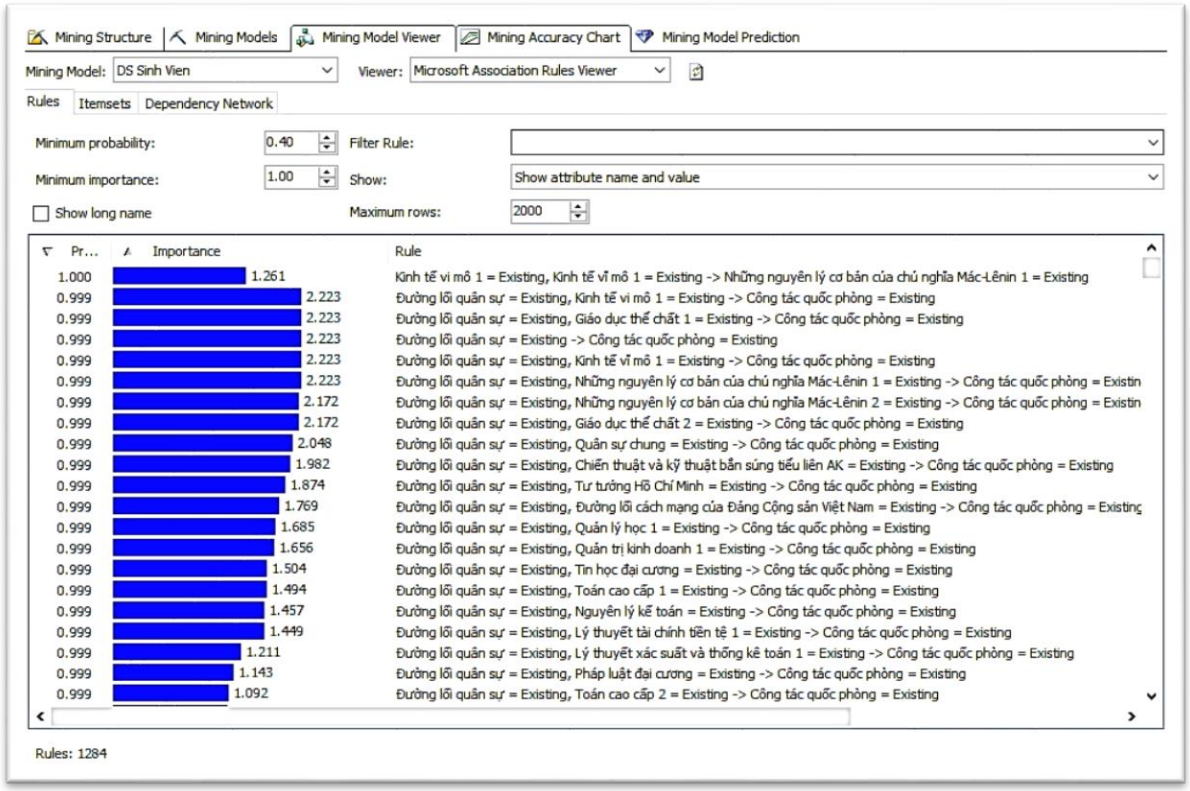
- Tạo kết nối dữ liệu nguồn (Data sources)
- Tạo các khung nhìn dữ liệu cho mô hình (Data source views)
- Tạo một cấu trúc mô hình khai phá (Mining Model structure)
- Hiệu chỉnh model
- Khai phá dữ liệu từ Model và View kết quả

Kết quả khai phá dữ liệu giải bài toán 1 sử dụng kỹ thuật khai phá luật kết hợp trong 3 lần chính như sau:

Lần 1: Với tất cả dữ liệu đăng ký môn học phần *(cả tự chọn và bắt buộc)* của khóa học với tất cả các chuyên ngành và đưa 2 view vào hệ quản trị cơ sở dữ liệu SQL.

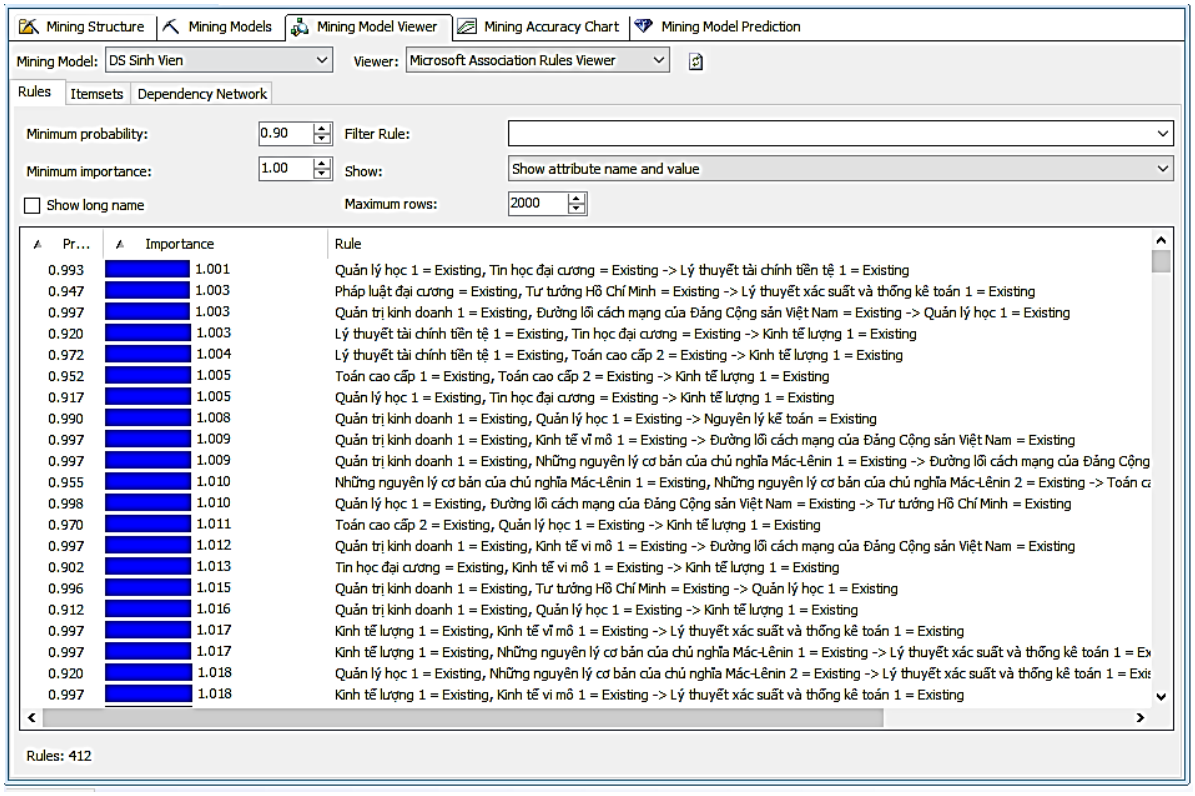
- Thực hiện khai phá, chọn thuộc tính đầu vào (input) và thuộc tính dự đoán (predict) là thuộc tính tên môn học phần trên hai bảng lồng *nhau (DanhSachSV chọn là case, và SV_DangKy_MonHoc chọn là nested).*

- Tiến hành chạy với minsupport=0.4 và minprobability = 0.4, thì thu được kết quả gồm 1284 luật kết hợp với tất cả các môn học của 45 chuyên ngành và 1 năm học.



Hình 3.3 L1.1: minsupport=0.4 và minprobability = 0.4

- Sau đó điều chỉnh tăng minprobability = 0.9, minsupport=0.4 thu được 421 luật.



Hình 3.4 L1.2: minsupport=0.4 và minprobability = 0.9

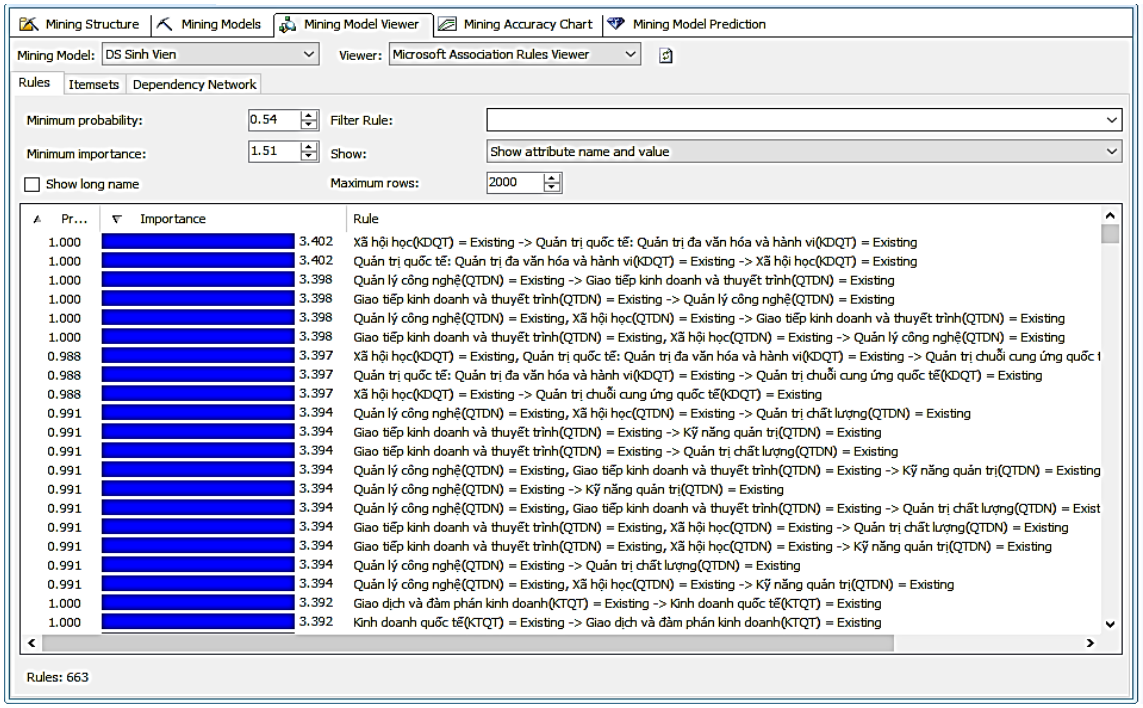
Nhận xét: Công cụ đã cho ra các luật như mong muốn, nhưng vì trên các luật không có thông tin chuyên ngành *(vì dữ liệu gồm tất cả các môn của tất cả các chuyên ngành)*, nên muốn tư vấn cho từng chuyên ngành lại phải tìm xem môn đó thuộc chuyên ngành nào việc tư vấn cho từng chuyên nghành là khó khăn. Có quá nhiều luật và luật lại kết hợp cả học phần tự chọn và học phần bắt buột nên lần 1 chạy là không khả thi, tiến hành thử nghiệm lần 2.

Lần 2: Vẫn dữ liệu như lần 1 và có thay đổi sau:

- Loại bỏ các học phần bắt buộc ra khỏi dữ liệu *(còn lại các học phần tự chọn)*.

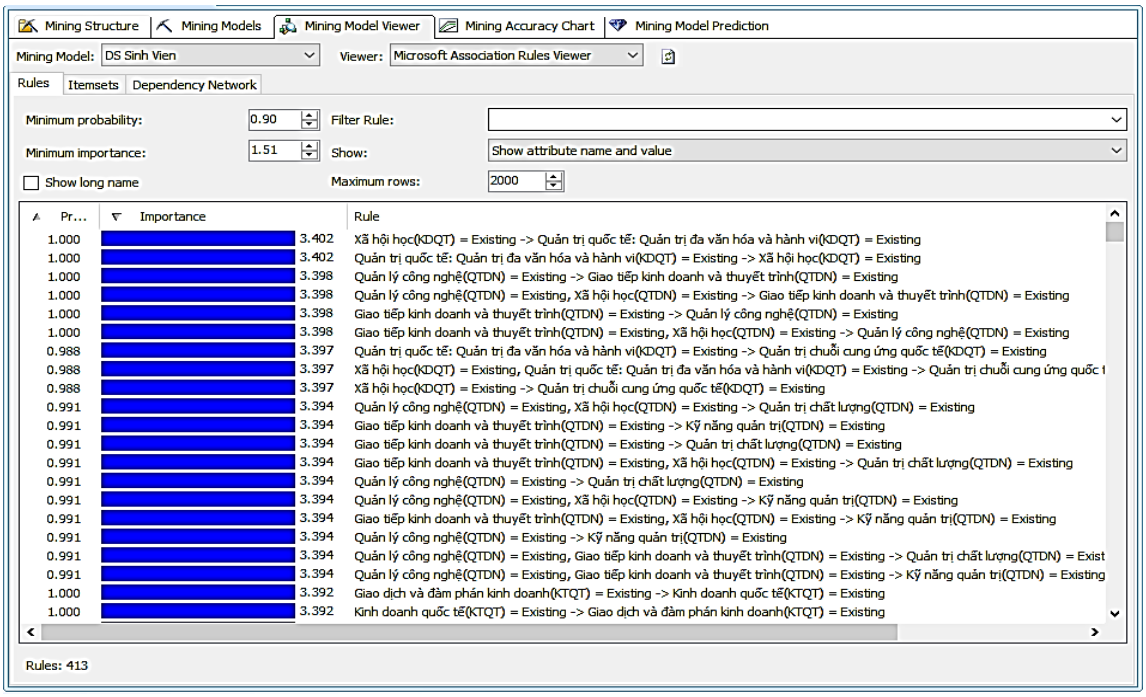
- Đưa thêm mã chuyên ngành vào sau các môn học phần tự chọn (*ví dụ; xã hộ học(KDQT), quản lý công nghệ(QTDN)*).

Kết quả: Chạy với minsupp= 0.03, minprobability= 0.54, thu được 663 luật



Hình 3.5 L2.1: minsupp= 0.03, minprobability= 0.54

- Sau đó thay đổi: minsupp= 0.03, minprobability= 0.9, thu được 413 luật



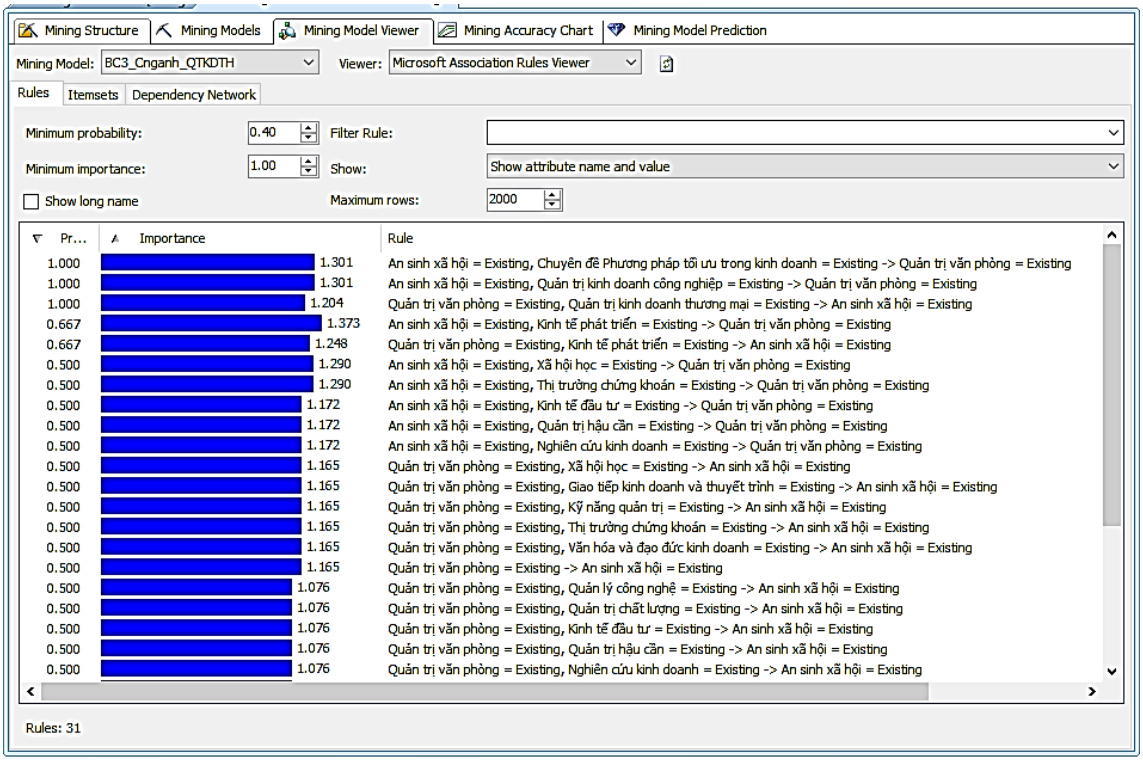
Hình 3.6 L2.2: minsupp= 0.03, minprobability= 0.9

Nhận xét: Tất cả luật thu được ở lần chạy 2 đều như mong muốn, tăng minprobability= 0.9, cũng thu được 413 luật cho nhiều chuyên ngành với xác suất cao , nhưng không đủ cho tất cả các chuyên ngành, hơn nữa muốn tư vấn theo chuyên ngành thì phải dùng công cụ lọc *(Filter Rule)* theo mã chuyên ngành, không có ý nghĩa với bài toán, Lần chạy 2 không khả thi, tiến hành thử nghiệm lần 3.

Lần 3: Vẫn là dữ liệu đã được loại bỏ học phần bắt buộc như lần chạy thứ hai và có một vài thay đổi như sau:

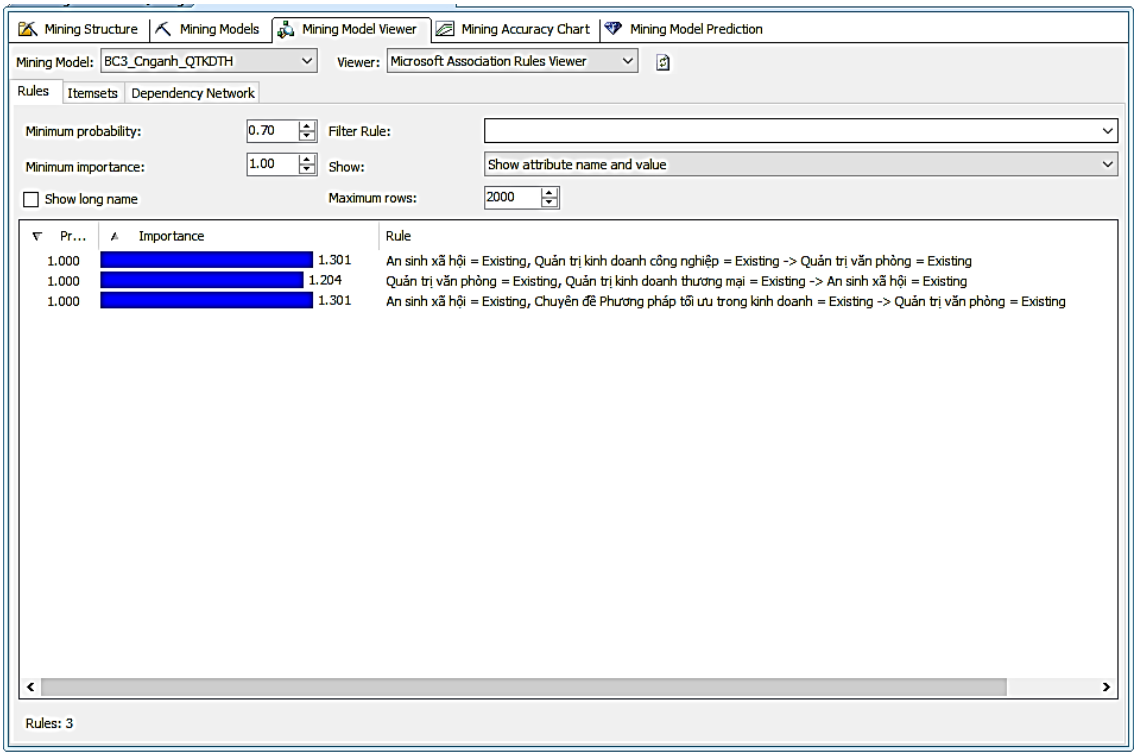
- Dữ liệu mới lúc này được tách ra mỗi chuyên ngành một bộ 2 view thể hiện sinh viên, môn học của chuyên ngành đó và quá trình đăng ký các học phần tự chọn. Tiến hành chạy thử với chuyên ngành Quản trị kinh doanh tổng hợp.

- Kết quả: chạy với minsupport=0.01, minprobability= 0.4, thu được 31



Hình 3.7 L3.1: minsupport=0.01, minprobability= 0.4

- Sau đó tăng minprobability= 0.7, giữ nguyên minsupport=0.01, thu được 3 luật với xác suất xảy ra là 100%.



Hình 3.8 L3.2: minsupport=0.01, minprobability= 0.7

Nhận xét: Dễ dành nhận thấy kết quả các luật thu được trong lần 3 này là phù hợp với ý nghĩa bài toán đã phát biểu nhất, có giá trị để tư vấn cho sinh viên đăng ký các học phần tự chọn của chuyên ngành quản trị kinh doanh tổng hợp. Có ý nghĩa rất phù hợp với yêu cầu bài toán 1, làm cơ sở thông tin cho cố vấn học tập tư vấn đăng ký môn học cho sinh viên.

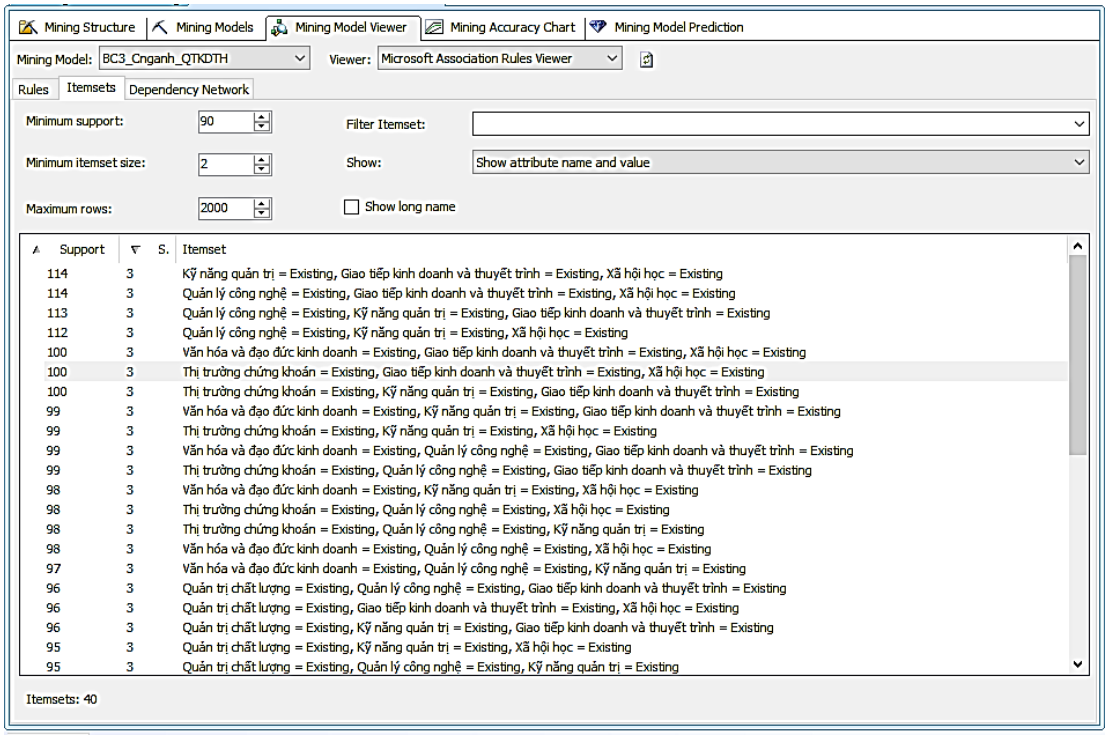
Ví dụ Luật : An sinh xã hội, Quản trị kinh doanh công nghiệp → Quản trị kinh doanh văn phòng (xảy ra với xác suất 100%)

- *Phát biểu luật tư vấn:* Sinh viên khi đăng ký môn *An sinh xã hội* trong tổ hợp 1 kiến thức lựa chọn của ngành, và đăng ký môn Quản trị kinh doanh công nghiệp trong tổ hợp 5 kiến thức lựa chọn của ngành, thì thường sẽ đăng ký môn Quản trị kinh doanh văn phòng trong tổ hợp 6 kiến thức lựa chọn của ngành với xác suất là 100%.

- Do đó với các lần thử nghiệm tiếp theo chỉ chạy với bộ dữ liệu như lần thử nghiệm thứ 3 riêng cho các chuyên ngành và ngành khác nhau. Luận văn đã chọn những ngành, chuyên ngành có số lượng sinh viên lớn nhất để chạy thử nghiệm, kết quả thu được rất hữu ích cho cán bộ cố vấn tư vấn đăng ký môn học cho sinh viên. Phần kết quả cho các chuyên ngành khác và phát biểu luật tư vấn được trình bày trong phần phụ lục.

Ngoài ra Trong Tab Itemsets: Itemsets là tập mục phổ biến, cho biết các thông tin quan trọng của luật kết hợp như Support (độ hỗ trợ của luật kết hợp), Size (Số items trong Itemsets).

- Thể hiện trên Tab Itemsets: tập mục phổ biến có Support là 114 gồm 3 môn học (items) đó là Kỹ năng quản trị, Giao tiếp kinh doanh và thuyết trình, Xã hội học, có ý nghĩa là trong tất cả các lượt đăng ký môn học thì xuất hiện 114 *(nhiều nhất trong tập 3 items)* lần trong đó sinh viên đăng ký 3 học phần tự chọn đó cùng với nhau hay nói cách khác đây cũng là tập 3 môn học phần tự chọn hay được sinh viên đăng ký cùng nhau nhất.



Hình 3.9 L3.3: thể hiện tập mục phổ biến (Itemsets)

Kết luận thực hiện: Thử nghiệm được tiến hành nhiều lần với nhiều chuyên ngành khác nhau và kết quả thu được có ý nghĩa với bài toán 1 giúp cán bộ cố vấn học tập có thêm cơ sở tư vấn lựa chọn môn học cho sinh viên, tập luật kết hợp giúp sinh viên nên đăng ký môn này cùng môn kia, và tập môn phổ biến chính là các môn học hay được đăng ký cùng nhau.

3.5 Khai phá dữ liệu bằng cây quyết định giải bài toán 2

3.5.1 Từ dữ liệu thô thu thập được

Dựa vào yêu cầu bài toán 2 để phân loại dự báo sinh viên có ra trường đúng hạn hay không và dựa vào quy chế đào tạo về số lượng tín chỉ, xếp loại học lực mỗi kỳ, mô hình cây quyết định được xây dựng để giải bài toán 2 sử dụng thông tin

đầu vào liên quan đến thuộc tính số tín chỉ, xếp loại học lực từ điểm trung bình chung cuối mỗi kỳ.

Lớp chuyên ngành KTQT 54A, KY 2010_2011_1

TT	Mã SV	Họ tên	GT	0	1	2	3	4	5	6	7	8	9	10	11	12	13	TC K5	TC TLK5	TBC K5	TBC TLK5
1	11120001	Nguyễn Ngọc Anh	Nữ										7.3	5				9	85	2.66	1.98
2	11120002	Nguyễn Tuấn Anh	Nam		5								3	5				6	84	0.75	2.17
3	11120003	Trần Văn Dũng	Nam											5			6.8	5	84	1.9	2.07
4	11120004	Đặng Ngọc Dương	Nam													5		6	83	1.83	2.18
5	11120005	Nguyễn Tài Hạnh	Nữ													0	0		73	0	2.24
6	11120006	Lê Quang Hiệp	Nam																65	0	2.12
7	11120007	Trịnh Tiến Hiếu	Nam			6.3	5.3					6.2		5				15	77	2.14	2.38
8	11120008	Lê Ngọc Hùng	Nam						8					5				5	89	1.64	2.01
9	11120009	Bùi Văn Huy	Nam							7.2					6.9			5	80	2.8	2.16
10	11120010	Nguyễn Thanh Hưng	Nam	0			0									0			85	0	1.98
11	11120011	Nguyễn Cảnh Linh	Nam									4.8			7.6			4	84	2	2.16
12	11120012	Nguyễn Hữu Nguyên	Nam							7.4	6.2							7	92	2.43	2.58
13	11120013	Lưu Vũ Phi	Nam			0										4.1		2	40	0.4	2.11
14	11120014	Phạm Văn Sự	Nam												7.3	0		2	65	1.5	1.98
15	11120015	Nguyễn Thưởng	Nam												8			2	77	3.5	2.07

Hình 3.10 Bảng điểm từng chuyên ngành theo kỳ sau khi biến đổi

3.5.2 Tiến hành biến đổi dữ liệu theo bài toán 2

Dữ liệu được biến đổi sang dạng bảng điểm từng lớp chuyên ngành theo kỳ (các môn học được quay lên các cột (ký hiệu bằng số), mỗi sinh viên cùng với điểm là một bản ghi), mục đích là để có điểm các học phần đã học trong kỳ đó, từ đó tính ra được 4 cột thông tin về tín chỉ và học lực cho từng kỳ, đây là những thông tin ảnh hưởng đến khả năng ra trường của sinh viên.

- Từ bảng điểm cho mỗi lớp chuyên ngành theo từng kỳ như hình trên, tiến hành loại bỏ thuộc tính các môn học phần, giữ lại 4 thuộc tính cuối là; Tín chỉ tích lũy trong kỳ đó, tín chỉ tích lũy từ kỳ đầu cho đến kỳ đó, điểm chung bình chung tại kỳ đó, điểm chung bình chung tích lũy từ kỳ đầu đến kỳ đó, bảng dữ liệu như sau:

- Làm tương tự với các kỳ còn lại với các chuyên ngành khác, cuối cùng gộp tất cả các chuyên ngành theo kỳ, dữ liệu tổng kết tương ứng với từng kỳ thu được như sau:

	A	B	C	D	AA	AB	AC	AD
1	QTKDTH 54A, KỲ 2010_2011_1							
2								
3	TT	Mã SV	Họ tên	GT	TC K5	TC TLK5	TBC K5	TBC TLK5
4	1	11120001	Nguyễn Ngọc Anh	Nữ	9	85	2.66	1.98
5	2	11120002	Nguyễn Tuấn Anh	Nam	6	84	0.75	2.17
6	3	11120003	Trần Văn Dũng	Nam	5	90	1.9	2.07
7	4	11120004	Đặng Ngọc Dương	Nam	6	83	1.83	2.18
8	5	11120005	Nguyễn Tài Hạnh	Nữ		73	0	2.24
9	6	11120006	Lê Quang Hiệp	Nam		65	0	2.12
10	7	11120007	Trịnh Tiến Hiếu	Nam	18	77	2.14	2.38
11	8	11120008	Lê Ngọc Hùng	Nam	5	89	1.64	2.01
12	9	11120009	Bùi Văn Huy	Nam	5	80	2.8	2.16
13	10	11120010	Nguyễn Thanh Hưng	Nam		85	0	1.98
14	11	11120011	Nguyễn Cảnh Linh	Nam	4	84	2	2.16
15	12	11120012	Nguyễn Hữu Nguyên	Nam	7	92	2.43	2.58
16	13	11120013	Lưu Vũ Phi	Nam	2	90	0.4	2.11
17	14	11120014	Phạm Văn Sự	Nam	2	65	1.5	1.98
18	15	11120015	Nguyễn Thưởng	Nam	2	77	3.5	2.07
19	16	11120016	Lê Bá Toán	Nam		89	0	2.04
20	17	11120017	Đặng Anh Tuấn	Nam	2	80	1.5	2.00

Hình 3.11 Bảng điểm tổng kết của một kỳ, tất cả chuyên ngành (ví dụ kỳ 5)

- Theo bài toán phân lớp dự đoán sinh viên có ra trường đúng hạn hay không, thực tế năm thứ 3 trở đi sinh viên thường hay đi làm thêm và có nhiều nguyên nhân dẫn đến lười học. Cán bộ cố vấn thường xuyên phải theo dõi 3 kỳ cuối. Do đó 3 mô hình khai phá được đề xuất cho 3 kỳ cuối là kỳ 5, kỳ 6, kỳ 7 được xây dựng dựa trên cây quyết định. Dự báo kịp thời vào 3 kỳ cuối có ý nghĩa với bài toán 2 đã phát biểu ở trên.

- Bảng điểm của 3 kỳ 5, 6, 7 *(dạng số)* sau khi thu được sẽ được biết đổi về dạng rời rạc *(các giá trị rời rạc)*, 4 thuộc tính đầu vào *(input)* được thay bằng giá trị rời rạc như sau:

+ TCKy5, TCKy6, TCKy7 mà lớn hơn 15 tín chỉ là giá trị "đủ", nhỏ hơn 15 tín chỉ là giá trị "không đủ"

+ TCTichLuyK5 mà lớn hơn 75 tín chỉ thì nhận giá trị "đủ", nhỏ hơn 75 tín chỉ nhận giá trị "không đủ"

+ TCTichLuyK6 mà lớn hơn 90 tín chỉ thì nhận giá trị "đủ", nhỏ hơn 90 tín chỉ nhận giá trị "không đủ"

+ TCTichLuyK7 mà lớn hơn 105 tín chỉ thì nhận giá trị "đủ", nhỏ hơn 112 tín chỉ là "không đủ"

+ HlucKy5,6,7 và HLucDenKy5,6,7: nằm trong các khoảng sau: 3.6< xuất sắc< 4 | 3.2< Giỏi< 3.6 | 2.5< Khá< 3,5 | 2 < tb <2.5 | <2 là Yếu.

- Thêm cột thuộc tính dự báo được lấy từ dữ liệu là cột: " tình trạng sinh viên" với 2 giá trị phân lớp (đúng hạn, không đúng hạn), dữ liệu sau khi biến đổi cuối cùng

để đưa vào công cụ khai phá có dạng sau (bốn cột thuộc tính đầu vào, một cột dự báo).

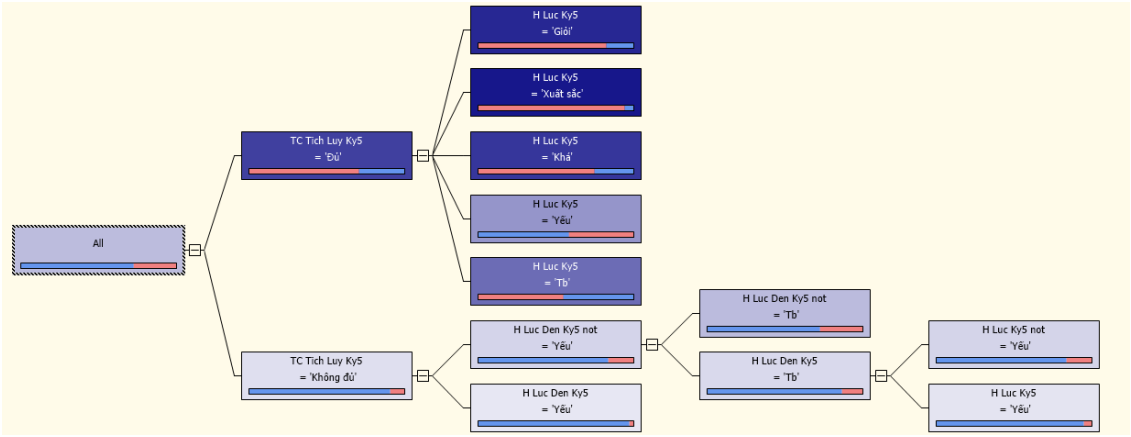
	A	B	C	D	E	F
1	MaSV	TCKy1	HLucKy1	TCTichLuyKy1	HLucDenKy1	Dự Báo
2	11120001	Không đủ	Yếu	Không đủ	Yếu	Không đúng hạn
3	11120002	Không đủ	Khá	Không đủ	Khá	Không đúng hạn
4	11120003	Đủ	Tb	Đủ	Tb	Không đúng hạn
5	11120004	Không đủ	Yếu	Không đủ	Yếu	Không đúng hạn
6	11120005	Không đủ	Khá	Không đủ	Khá	Không đúng hạn
7	11120006	Không đủ	Tb	Đủ	Khá	Đúng hạn
8	11120007	Không đủ	Tb	Không đủ	Tb	Không đúng hạn
9	11120008	Không đủ	Yếu	Không đủ	Yếu	Không đúng hạn
10	11120009	Không đủ	Tb	Không đủ	Tb	Đúng hạn
11	11120010	Không đủ	Yếu	Không đủ	Yếu	Không đúng hạn
12	11120011	Không đủ	Khá	Không đủ	Khá	Không đúng hạn
13	11120012	Không đủ	Tb	Không đủ	Tb	Đúng hạn
14	11120013	Không đủ	Yếu	Không đủ	Yếu	Không đúng hạn
15	11120014	Không đủ	Tb	Không đủ	Tb	Không đúng hạn

Hình 3.12 Bảng dữ liệu đưa vào khai phá

3.5.3 Thực hiện thử nghiệm trên công cụ BIDS

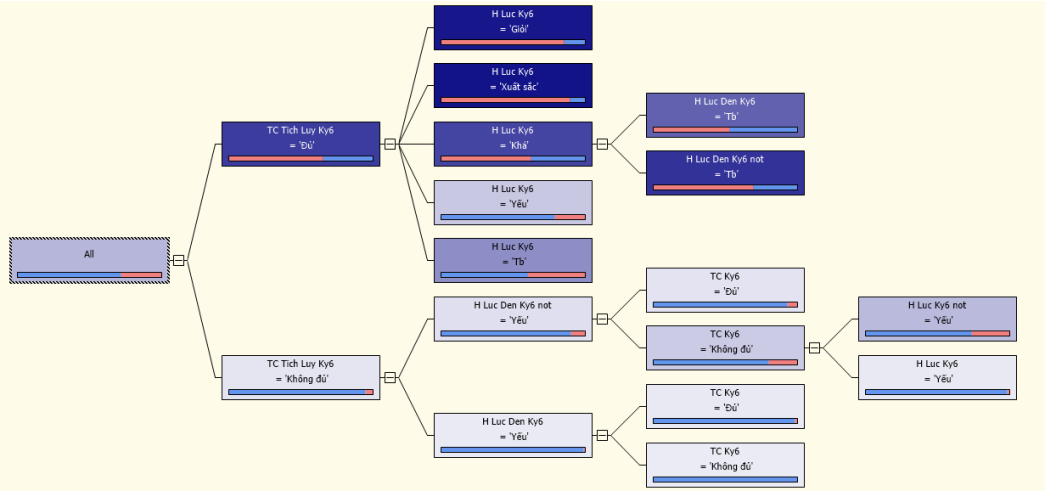
- Tiến hành giống 5 bước khai phá luật kết hợp giải bài toán 1, ở bài toán 2 chỉ khác là chọn kỹ thuật cây quyết định. Như trên đã đề xuất 3 mô hình dự báo phân lớp cho 3 kỳ. Kết quả chạy thử nghiệm 3 mô hình như sau:

Lần 1: Mô hình phân lớp dự báo cho kỳ 5. Trên cây quyết định phân lớp thu được, phần xanh đậm là lớp dự đoán cho khả năng sinh viên ra trường đúng hạn.



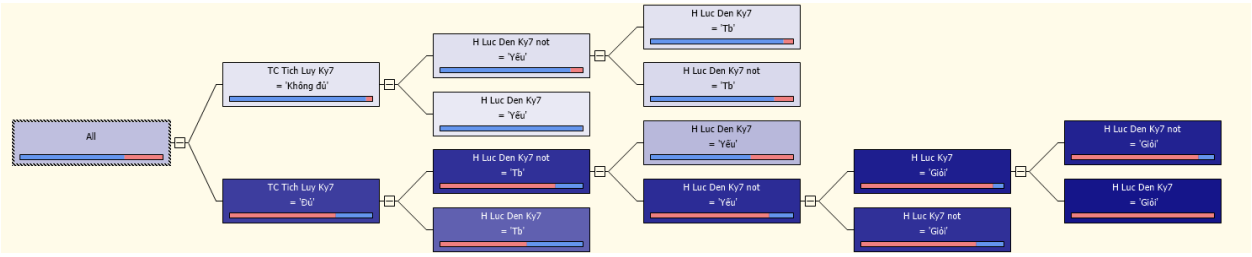
Hình 3.13 Cây quyết định phân lớp kỳ 5

Lần 2: Mô hình phân lớp dự báo cho kỳ 6



Hình 3.14 Cây quyết định phân lớp kỳ 6

Lần 3: Mô hình phân lớp dự báo cho kỳ 7



Hình 3.15 Cây quyết định phân lớp kỳ 7

Nhận xét: Kết quả thu được sau khi thử nghiệm là 3 cây quyết định dùng để phân lớp dự đoán. Dựa vào 3 cây quyết định của kỳ 5, kỳ 6, kỳ 7, cán bộ cố vấn học tập sẽ phân loại được nhóm ra trường đúng hạn hay không đúng hạn. Có ý nghĩa làm cơ sở để cảnh báo kịp thời cho sinh viên, phù hợp với bài toán 2 đã phát biểu.

- Từ cây quyết định có thể hiểu thành các luật như sau:

Ví dụ: Nếu TC Tich Luy Ky7 = 'Đủ' and H Luc Den Ky7 not = 'Tb' and H Luc Den Ky7 not = 'Yếu' and H Luc Den Ky7 not = 'Khá' thì ra trường Đúng hạn.

Kết quả bảng luật cụ thể có ý nghĩa góp phần giúp cán bộ cố vấn học tập ra quyết định cảnh báo học tập.

3.6 Một số đề xuất, kiến nghị

- Qua thực nghiệm và thu được kết quả ý nghĩa với bài toán cố vấn học tập trong thực tiễn, có một số đề xuất kiến nghị như sau:

- Quy chế và chương trình đào tạo tại trường Đại học kinh tế quốc dân nên được thống nhất giữa các khóa để dễ dàng theo dõi thống kê, quản lý dữ liệu, phục vụ cho việc phân tích khai phá tốt hơn, nhằm tìm ra các mẫu có ích với thực tiễn.

- Từ kết quả đạt được và có ý nghĩa thực tiễn qua thực nghiệm trong luận văn, tác giả đề xuất Trường đại học kinh tế quốc dân nên xây dựng một hệ thống cố vấn học tập hoàn chỉnh cho sinh viên chính quy. Giúp bộ phận cố vấn có thêm công cụ trực quan tư vấn cho sinh viên kế hoạch học tập sao cho phù hợp nhất.

- Hệ thống hoàn chỉnh gồm có cơ sở dữ liệu lớn và các lớp bài toán cố vấn nhằm giải quyết tất cả những vấn đề khúc mắc về học tập trong 4 năm của sinh viên.

- Triển khai được hệ thống cũng là góp phần nâng cao việc ứng dụng công nghệ thông tin vào công tác quản lý chung của nhà trường. Nâng cao chất lượng phục vụ, đào tạo và học tập của sinh viên cũng là góp phần thực hiện mục tiêu đổi mới, hội nhập và phát triển của Nhà trường.

3.7 Tổng kết chương 3

Qua chương 3 tác giả đã hiểu được cơ bản các bước thực hiện khai phá dữ liệu trên một vài công cụ, nhấn mạng chủ yếu vào công cụ BIDS. Hiểu và cài đặt được công cụ, nắm được 5 bước chính để tiến hành khai phá dữ liệu.

Từ bài toán đề xuất trong chương 2, từ tiền đề cơ sở lý thuyết về kỹ thuật khai phá luật kết hợp và cây quyết định trong chương 1. Tác giả đã biết đổi dữ liệu thô thu thập được phù hợp tương ứng với từng phương pháp và bài toán. Đề xuất mô hình khai phá dữ liệu cho 2 bài toán.

Bằng việc thực nghiệm trên công cụ BIDS và đã thu được kết quả như mong muốn. Tác giả đã nhận xét và kết luận các kết quả đạt được là đúng mục tiêu và ý nghĩa đối với bài toán trong chương 2. Ý nghĩa quan trọng nhất là đóng góp nhiều thông tin cho đội ngũ cán bộ cố vấn hoàn thành tốt nhiệm vụ của mình. Góp phần nâng cao chất lượng phục vụ và đào tạo tại Trường kinh tế. Cuối cùng tác giả đề xuất một vài ý kiến về xây dựng hệ thống cố vấn học tập hoàn chỉnh cho trường Đại học Kinh tế quốc dân.

KẾT LUẬN

Sau một thời gian nghiên cứu và thực hiện đề tài dưới sự hướng dẫn của thầy TS. Nguyễn Trung Tuấn, luận văn đã đạt được mục tiêu đã đề ra, thu được những kết quả ý nghĩa với thực tiễn.

Đã tóm tắt được lý thuyết liên quan đến phát hiện tri thức và khai phá dữ liệu, đặt biệt nhấn mạnh vào hai phương pháp khai phá dữ liệu cơ bản là luật kết hợp và cây quyết định.

Đã hiểu được quy định chung trong đào tạo theo học chế tín chỉ, những vấn đề còn tồn tại trong công tác cố vấn học tập, thu thập và tìm hiểu về dữ liệu quản lý đào tạo sinh viên đại học.

Đã đề xuất được bài toán mà mục tiêu là trợ giúp cho các hoạt động cố vấn học tập. Có thêm cơ sở thông tin cho cán bộ cố vấn học tập hoàn thành nhiệm vụ.

Sau khi áp dụng thử nghiệm trên công cụ BIDS để khai thác dữ liệu giải bài toán dựa vào kỹ thuật thuật cây quyết định và luật kết hợp đã thu được các kết quả có ý nghĩa với mục tiêu bài toán đã phát biểu.

Hạn chế:

Do thời gian có hạn nên luận văn không tránh khỏi những thiếu sót, dữ liệu thực nghiệm cần thu thập nhiều hơn nữa.

Hướng phát triển:

- Nghiên cứu thêm các kỹ thuật khai phá dữ liệu và các công cụ khác nữa.

- Phân tích sâu hơn về các phương phát KPDL để lựa chọn phương phát tối ưu nhất cho các bài toán cố vấn học tập.

- Thu thập và xử lý thêm dữ liệu của các khóa khác để tăng độ chính xác.

- Phát biểu thêm các bài toán cố vấn học tập khác nữa, nhằm có thêm nhiều cơ sở giúp ích cho hoạt động cố vấn học tập thêm ý nghĩa.

- Xây dựng một hệ thống hoàn chỉnh gồm nhiều bài toán cố vấn học tập, hỗ trợ tốt cho đội ngũ cố vấn, giúp ích cho nâng cao chất lượng đào tạo chung của Trường Đại học Kinh tế Quốc dân.

DANH MỤC TÀI LIỆU THAM KHẢO TIẾNG VIỆT

[1] Bài giảng "Kho dữ liệu và khai phá dữ liệu", Hà Quang Thụy, Đại học Công Nghệ, 2015.

[2] Phần V Mục 13 Quy định về cố vấn học tập *(Trích Quyết định số: 1808/QĐ-KTQD-TTr&KT ngày 25/11/2010 của Hiệu trưởng Trường Đại học Kinh tế Quốc dân).*

[3] Khóa luận tốt nhiệp, Nghiên cứu các thuật toán phân lớp dữ liệu dựa trên cây quyết định, Nguyễn Thị Thùy Linh, Đại học Công nghệ, 2005.

DANH MỤC TÀI LIỆU THAM KHẢO TIẾNG ANH

[4] Bao H.T, *Introduction to Knowledge Discovery and Data Mining*, Lecture note, Institute of Information Technology, VietNam, 2008.

[5] Dasarathy B.V., *Data mining tasks and methods: Classification: nearest-neighbor approaches*, Oxford University Press, Inc., New York, NY, USA, 2002.

[6] Fayyad U., Piatetsky-Shapiro G., Smyth P., *From data mining to Knowledge Discovery in Databases*, American Association for Artificial Intelligence, 1996.

[7] Han J. and Kamber M., *Data mining: concepts and techniques*, 2nd ed., Morgan Kaufmann, 2006.

DANH MỤC WEBSITE THAM KHẢO

[8] https://www.mssqltips.com/sqlservertip/3184/sql-server-2012-analysis-services-association-rules-data-mining-example/

[9] https://www.mssqltips.com/sqlservertip/2965/classic-machine-learning-example-in-sql-server-analysis-services/

[10] https://docs.microsoft.com/en-us/sql/analysis-services/data-mining/data-mining-ssas

[11] http://bis.net.vn/forums/p/378/661.aspx