BỘ GIÁO DỤC – ĐÀO TẠO
ĐẠI HỌC ĐÀ NẴNG

TRẦN THỊ BÍCH ĐÀO

ỨNG DỤNG KHAI PHÁ DỮ LIỆU ĐỂ TÌM LUẬT KẾT HỢP ĐÁNG TIN CẬY TRONG HỆ THỐNG BÁN HÀNG TẠI CÔNG TY DƯỢC TW3

Chuyên ngành: **KHOA HỌC MÁY TÍNH**
Mã số: **60.48.01**

TÓM TẮT LUẬN VĂN THẠC SĨ KỸ THUẬT

Đà Nẵng - Năm 2012

Công trình được hoàn thành tại
ĐẠI HỌC ĐÀ NẴNG

Người hướng dẫn khoa học: TS. Huỳnh Công Pháp

Phản biện 1: TS. Trương Ngọc Châu

Phản biện 2: TS. Trương Công Tuấn

Luận văn đã được bảo vệ trước hội đồng chấm Luận văn tốt nghiệp Thạc sĩ Kỹ thuật họp tại Đại học Đà Nẵng vào ngày 16 tháng 06 năm 2012.

Có thể tìm hiểu Luận văn tại:
- Trung tâm Thông tin – Học liệu, Đại học Đà Nẵng.
- Trung tâm Học liệu, Đại học Đà Nẵng.

MỞ ĐẦU

1. Lý do chọn đề tài

Ngày nay, công nghệ thông tin đang dần phổ biến trên hầu hết các lĩnh vực. Tỷ lệ thuận với sự phát triển đó là lượng dữ liệu được chúng ta lưu trữ cũng lớn theo. Chúng ta biết rằng trong lượng dữ liệu đó đang ẩn chứa những giá trị nhất định. Tuy nhiên theo thống kê, chỉ một lượng nhỏ những dữ liệu này (khoảng 5% - 10%) là được phân tích, số còn lại không biết để làm gì nhưng chúng ta vẫn luôn phải lưu trữ vì sợ sẽ bỏ qua những thông tin quan trọng nào đó hoặc một ngày nào đó sẽ dùng tới chúng. Do đó, các phương pháp quản trị và khai thác cơ sở dữ liệu truyền thống ngày càng không thể đáp ứng được thực tế đã làm phát sinh một khuynh hướng kỹ thuật mới: đó là phát hiện tri thức và khai phá dữ liệu KDD (Knowledge Discovery and Data Mining). Phát hiện tri thức và khai phá dữ liệu là quá trình phát hiện tri thức tiềm ẩn, tiềm năng, không biết trước và có lợi từ kho dữ liệu lớn. KDD là sự kế thừa và phát triển các thành tựu của nhiều lĩnh vực nghiên cứu ứng dụng tin học trước đó như: Hệ chuyên gia, Trí tuệ nhân tạo, lý thuyết nhận dạng, …

Thị trường về dược phẩm, thiết bị y tế ngày càng phát triển mạnh mẽ, các công ty kinh doanh về lĩnh vực này liên tục đưa ra các sản phẩm, các hình thức kinh doanh mới cạnh tranh với nhau nhằm thu hút người tiêu dùng. Để công ty có thể tồn tại, phát triển bền vững và cạnh tranh trên thị trường thì cần phải đưa ra những nhận định kịp thời, và người quản lý phải có cách nhìn tổng thể về cách thức đầu tư về mặt hàng nào nhằm đáp ứng đúng yêu cầu của khách hàng để có hướng đầu tư đúng đắn. Hiện tại, công ty Dược đang có một nguồn dữ liệu lớn thông tin về khách hàng, số lượng, doanh thu các mặt hàng được bán ra,… Dựa vào lưu lượng dữ liệu này, và do đây là một hướng đi tiềm năng, có nhiều khả năng phát triển trong tương lai, nên tôi đã chọn đề tài : *"Ứng dụng khai phá dữ liệu để tìm luật kết hợp tin cậy trong hệ thống bán hàng tại Công ty Dược TW3"*.

Đề tài chỉ mô tả và thực hiện một số chức năng của hệ thống bán hàng nhằm phục vụ cho mục đích chính của đề tài là ứng dụng khai phá dữ liệu, cụ thể là *ứng dụng thuật toán phân lớp với cây quyết định để tìm luật kết hợp* trong hệ thống bán hàng của Công ty Dược TW3, mang lại cho người quản lý có cách nhìn tổng quát về nhu cầu mua bán để đưa ra những nhận định đúng và hợp lý, mang lại hiệu quả cho hoạt động bán hàng tại công ty.

2. Đối tượng và phạm vi nghiên cứu

a. Đối tượng

Lý thuyết

- Kỹ thuật khai phá dữ liệu
- Nghiệp vụ quản lý bán hàng tại công ty dược TW3

Dữ liệu

- Cơ sở dữ liệu: khách hàng, loại hàng, mặt hàng…
- Các văn bản, qui định có liên quan…

b. Phạm vi

- Nghiên cứu các kiến thức cơ bản về phương pháp luật kết hợp.
- Tìm hiểu các qui trình tác nghiệp trong hệ thống.

3. Mục tiêu và nhiệm vụ

a. Mục tiêu

- Ứng dụng luật kết hợp vào công tác quản lý bán hàng.
- Đưa ra kết quả nhận định hay các dự đoán mang tính chiến lược cho nhà quản lý.

b. Nhiệm vụ

Nghiên cứu cơ sở lý thuyết

- Nghiên cứu kỹ thuật khai phá dữ liệu.
- Nghiên cứu và phát triển các thuật giải tìm luật kết hợp.

- Ứng dụng các thuật toán trên vào cơ sở dữ liệu quản lý bán hàng.

Triển khai xây dựng ứng dụng

- Xây dựng cơ sở dữ liệu mẫu.
- Xây dựng các ứng dụng.

4. Phương pháp nghiên cứu

- Tham khảo các tài liệu liên quan, các bài báo khoa học…
- Lập kế hoạch, lập qui trình, tiến độ thực hiện
- Nghiên cứu kỹ thuật khai phá dữ liệu bằng việc tìm luật kết hợp giữa các mặt hàng dựa trên loại hàng và doanh thu của các loại hàng đó.

5. Kết quả dự kiến

- Tổng hợp các kiến thức cơ bản của phương pháp khai phá luật kết hợp
- Xây dựng một số ứng dụng đơn giản, dễ sử dụng nhưng mang tính hiệu quả cao.

6. Ý nghĩa khoa học và thực tiễn của đề tài

a. Ý nghĩa khoa học

- Đây là phương pháp được nhiều nhà khoa học nghiên cứu và cũng có rất nhiều đóng góp vào thực tiễn.
- Ứng dụng tin học vào trong công tác quản lý.

b. Ý nghĩa thực tiễn

- Giải quyết được một số tác nghiệp trong công tác quản lý.
- Đánh giá kết quả nhận định, hỗ trợ đưa ra các quyết định hay các dự đoán mang tính chiến lược dựa trên loại hàng và doanh thu của các loại hàng đó.
- Giúp nhà quản lý nắm bắt kịp thời các nhu cầu mua bán trên thị trường và có một cách nhìn tổng quan hơn.

7. Cấu trúc luận văn

Luận văn gồm có 3 chương:

- Chương 1: Tổng quan về lý thuyết
  - Nghiên cứu, tìm hiểu lý thuyết khai phá dữ liệu.
  - Trình bày thuật toán được áp dụng trong luận văn: thuật toán phân lớp với cây quyết định.
- Chương 2: Phân tích thiết kế hệ thống quản lý bán hàng tại công ty dược TW3
  - Phát biểu bài toán: định nghĩa bài toán và qui trình bán hàng.
  - Phân tích thiết kế cơ sở dữ liệu và xác định các tác nhân liên quan đến hệ thống bán hàng.
- Chương 3: Xây dựng chương trình và thực nghiệm
  - Trình bày ngôn ngữ lập trình
  - Đưa ra các dữ liệu thực tế thu thập được
  - Thiết kế giao diện bao gồm 2 số chức năng chính: khai phá dữ liệu theo mã loại hàng và khai phá dữ liệu các loại hàng theo doanh thu. Bên cạnh đó còn có một số chức năng hỗ trợ thêm: danh mục khách hàng, cập nhật thông tin hóa đơn, quản lý doanh thu bán hàng…

CHƯƠNG 1: TỔNG QUAN VỀ LÝ THUYẾT

1.1. LÝ THUYẾT VỀ KHAI PHÁ DỮ LIỆU

1.1.1. Khai phá dữ liệu

1.1.1.1. Định nghĩa khai phá dữ liệu

Định nghĩa của Ferruzza: “Khai phá dữ liệu là tập hợp các phương pháp được dùng trong tiến trình khám phá tri thức để chỉ ra sự khác biệt các mối quan hệ và các mẫu chưa biết bên trong dữ liệu”.

Định nghĩa của Parsaye: “Khai phá dữ liệu là quá trình trợ giúp quyết định, trong đó chúng ta tìm kiếm các mẫu thông tin chưa biết và bất ngờ trong CSDL lớn”.

Định nghĩa của Fayyad: "Khai phá tri thức là một quá trình không tầm thường nhận ra những mẫu dữ liệu có giá trị, mới, hữu ích, tiềm năng và có thể hiểu được".

1.1.1.2. Đặc điểm của khai phá dữ liệu

Khai phá dữ liệu là giai đoạn chủ yếu của quá trình phát hiện tri thức.

Khai phá dữ liệu để tìm ra các mẫu (pattern) có ý nghĩa được tiến hành trên tập dữ liệu mà ta hy vọng là sẽ thích hợp với nhiệm vụ khai phá hiện thời.

Mẫu tìm được từ quá trình khai phá dữ liệu phải có tính mô tả (description) và dự đoán (prediction).

Khai phá dữ liệu là quá trình mà trong đó con người là trung tâm.

Khai phá dữ liệu là quá trình tìm kiếm tri thức chỉ từ dữ liệu.

Khai phá dữ liệu mang tính chất hướng nhiệm vụ.

1.1.1.3. Ý nghĩa thực tiễn và tình hình ứng dụng khai phá dữ liệu

a. Ý nghĩa thực tiễn

Cùng với sự tăng lên không ngừng của khối lượng dữ liệu , yêu cầu khai thác dữ liệu ngày càng cao hơn. Ngoài những đòi hỏi về tính linh hoạt, năng suất, sự chuyên môn hóa trong vấn đề khai thác, CSDL cần phải mang lại tri thức hơn là chính dữ liệu đó. Các quyết định cần phải hợp lý, nhanh chóng, chính xác và có khả năng dự đoán sự việc trong tương lai. Trước yêu cầu này, cách khai thác CSDL truyền thống cho thấy sự hạn chế của mình. Khai phá ra đời mở hướng cho sự khó khăn này.

Có thể kể một số ứng dụng của khai phá dữ liệu như sau: một công ty bảo hiểm muốn phát hiện từ CSDL của khách hàng bị nghi ngờ là gian lận, khi đó, người ta thực hiện khai phá dữ liệu trên CSDL chứa các thông tin liên quan đến giao dịch giữa khách hàng và công ty để tìm ra sự phân lớp, có thể là lớp "đáng tin" và lớp "không đáng tin" trong

khách hàng. Từ đó công ty sẽ có biện pháp hạn chế gian lận xảy ra. Hay công ty nhận đặt hàng từ khách hàng qua email có thể giảm bớt chi phí gửi email bằng cách dùng tri thức khám phá để chỉ gửi email liên lạc đến những khách hàng có khả năng mua thường xuyên. Bệnh viện cũng cần khám phá tri thức từ dữ liệu nhằm phục vụ cho mục đích nghiên cứu, chẩn đoán trong ngành y…

b. Tình hình ứng dụng

Ở Việt Nam, có nhiều đề tài nghiên cứu khoa học về khai phá dữ liệu và đạt được nhiều kết quả đáng khích lệ.

Khai phá dữ liệu là một lĩnh vực nghiên cứu mới dùng các kỹ thuật thông minh để khai phá tri thức tìm ẩn trong dữ liệu. Khả năng hỗ trợ công việc của khai phá dữ liệu làm cho việc ứng dụng kỹ thuật này vào thực tế ngày càng rộng rãi hơn. Mặc dù, các hệ thống khai phá dữ liệu khai phá dữ liệu trên thế giới ít nhiều còn hạn chế nhưng đã dần dần hoàn thiện hơn và thực sự trở thành một công cụ quan trọng không thể thiếu được trong hầu hết các lĩnh vực xã hội.

1.1.2. Các bước cơ bản của quá trình phát hiện tri thức

Nhìn chung, quá trình khai phá dữ liệu gồm các bước sau:

Bước 1: Tìm hiểu lĩnh vực ứng dụng và xác định mục đích khai phá dữ liệu.

Bước 2: Xác định dữ liệu liên quan và hình thức khai phá.

Bước 3: Tiền xử lý dữ liệu.

Bước 4: Chọn thuật toán khai phá và chuyển dữ liệu về dạng phù hợp.

Bước 5: Khai phá dữ liệu.

Bước 6: Trích lọc các mẫu thực sự có ý nghĩa.

Bước 7: Ứng dụng tri thức phát hiện được.

1.2. LUẬT KẾT HỢP TRONG KHAI PHÁ DỮ LIỆU

1.2.1. Vài nét về khai phá luật kết hợp

Mục đích chính của khai phá dữ liệu là trích rút tri thức một cách tự động, hiệu quả và "thông minh" từ kho dữ liệu.

Trong hoạt động sản xuất kinh doanh, ví dụ kinh doanh các mặt hàng tại siêu thị, các nhà quản lý rất thích có được những thông tin mang tính thống kê như: "90% phụ nữ có xe máy màu đỏ và đeo đồng hồ Thụy Sỹ thì dùng nước hoa hiệu Chanel" hoặc "70% khách hàng là công nhân thì mua TV thường mua loại 21 inches". Những thông tin như vậy rất hữu ích trong việc định hướng kinh doanh. Vậy vấn đề đặt ra là liệu có tìm được các luật như vậy bằng các công cụ khai phá dữ liệu hay không? Câu trả lời là hoàn toàn có thể. Đó chính là nhiệm vụ khai phá luật kết hợp.

1.2.2. Luật kết hợp

1.2.2.1. Định nghĩa về luật kết hợp

Định nghĩa 1: Cho $I=\{I_1, I_2, .., I_m\}$ là tập hợp của m tính chất riêng biệt. Giả sử D là cơ sở dữ liệu, với các bản ghi chứa một tập con T các tính chất (có thể coi như $T \subseteq I$), các bản ghi đều có chỉ số riêng. Một luật kết hợp là một mệnh đề kéo theo có dạng $X \rightarrow Y$, trong đó $X, Y \subseteq I$, thỏa mãn điều kiện $X \cap Y = \varnothing$. Các tập hợp X và Y được gọi là các tập hợp tính chất (itemset). Tập X gọi là nguyên nhân, tập Y gọi là hệ quả.

Có 2 độ đo quan trọng đối với luật kết hợp: Độ hỗ trợ (support) và độ tin cậy (confidence), được định nghĩa như phần dưới đây.

Định nghĩa 2: Độ hỗ trợ

Độ hỗ trợ của một tập hợp X trong cơ sở dữ liệu D là tỷ số giữa các bản ghi $T \subseteq D$ có chứa tập X và tổng số bản ghi trong D (hay là phần trăm của các bản ghi trong D có chứa tập hợp X), ký hiệu là Support(X) hay Supp(X).

Ký hiệu: Supp(X).

Ta có: $0 \le Supp(X) \le 1$ với mọi tập hợp X.

Độ hỗ trợ Supp(X) còn được hiểu là xác suất X được thỏa trong D.

Ký hiệu: P(X).

Độ hỗ trợ của một luật kết hợp $X \rightarrow Y$ là tỷ lệ giữa số lượng các bản ghi chứa tập hợp $X \cup Y$, so với tổng số các bản ghi trong D.

$$Supp(X \rightarrow Y) = Supp(X \cup Y) = \frac{|\{T \in D \mid X \cup Y \subseteq T\}|}{|D|}$$

Khi chúng ta nói rằng độ hỗ trợ của một luật là 70%, có nghĩa là có 70% tổng số bản ghi chứa $X \cup Y$. Như vậy, độ hỗ trợ mang ý nghĩa thống kê của luật.

Độ hỗ trợ của X là:

$$Supp(X) = \frac{\text{Số lượng giao dịch hỗ trợ (X)}}{\text{Tổng số giao dịch}}$$

Định nghĩa 3: Độ tin cậy

Độ tin cậy (*Confidence*) của luật kết hợp có dạng R: $X \rightarrow Y$ là tỷ lệ giữa số lượng các bản ghi trong D chứa $X \cup Y$ với số bản ghi trong D có chứa tập hợp X. Ký hiệu độ tin cậy của một luật là Conf(R).

$$Conf(X \rightarrow Y) = P(Y \mid X) = \frac{Supp(X \cup Y)}{Supp(X)}$$

Có thể định nghĩa độ tin cậy như sau:

Độ tin cậy của một luật kết hợp $X \rightarrow Y$ là tỷ lệ giữa số lượng các bản ghi của tập hợp chứa $X \cup Y$, so với tổng số các bản ghi chứa X.

Việc khai thác các luật kết hợp từ cơ sở dữ liệu chính là việc tìm tất cả các luật có độ hỗ trợ và độ tin cậy do người sử dụng xác định trước. Các ngưỡng của độ hỗ trợ và độ tin cậy được ký hiệu là *minsup, minconf* và do người dùng xác định.

Việc khai thác các luật kết hợp có thể được phân tích thành hai vấn đề:

1. Tìm tất cả các tập mục thường xuyên xảy ra mà có độ hỗ trợ lớn hơn hoặc bằng minsup.
2. Tạo ra các luật mong muốn sử dụng các tập mục lớn mà có độ tin cậy lớn hơn hoặc bằng minconf.

Định nghĩa 4: Độ quan trọng

Độ quan trọng (*importance*) của luật $X \rightarrow Y$, ký hiệu Imp($X \rightarrow Y$), được xác định bởi tỷ số giữa $Conf(X \rightarrow Y)$ và $Conf(\overline{X} \rightarrow Y)$.

$$\text{Imp}(X \to Y) = \lg\left(\frac{Conf(Y \to X)}{Conf(Y \to \overline{X})}\right) = \lg\left(\frac{P(X \mid Y)}{P(\overline{X} \mid Y)}\right)$$

Trong tính toán, ta thường đưa tỷ số này vào lôgarit để độ quan trọng có giá trị xung quanh 0.

1.2.2.2. Một số hướng tiếp cận trong khai phá luật kết hợp

1.2.2.3. Một số thuật toán phát hiện luật kết hợp

1.3. THUẬT TOÁN PHÂN LỚP VỚI CÂY QUYẾT ĐỊNH

1.3.1. Đặt vấn đề

Giả sử doanh nghiệp đã đưa ra một số tiêu chí để phân loại khách hàng là VIP hoặc không VIP: có khối lượng giao dịch trung bình mỗi tháng đạt từ 3,000,000 VND trở lên, có tần suất giao dịch trung bình 10 lần mỗi tháng.

Vấn đề đặt ra của doanh nghiệp là cần xác định các đặc trưng chung của nhóm khách hàng VIP, để từ đó làm cơ sở dự báo về một khách hàng (mới) có tiềm năng trở thành khách hàng VIP hay không. Trong bảng trên, các thuộc tính đã được rời rạc hóa theo cách:

Tuổi: Bằng 1 nếu tuổi nhỏ hơn 25, bằng 2 nếu tuổi từ 25 đến 40, bằng 3 nếu tuổi lớn hơn 40.

Giới tính: Bằng 1 nếu là nữ, bằng 0 nếu là nam,

Thu nhập: Bằng 1 nếu thu nhập ít hơn 30 triệu VND/năm, bằng 2 nếu từ 30 triệu VND đến 50 triệu VND/năm, bằng 3 nếu trên 50 triệu VND/năm,

Tình trạng hôn nhân: Bằng 0 nếu chưa lập gia đình, bằng 1 nếu ngược lại.

1.3.2. Một số định nghĩa

Cho bảng dữ liệu A gồm n dòng với các thuộc tính: (X_1, X_2, ..., X_N, Y), trong đó Y là thuộc tính output (thuộc tính cần dự báo) và X_1, X_2, ..., X_N là các thuộc tính input.

Giả sử Y đã được rời rạc hóa thành k giá trị là y_1, y_2, ..., y_k (nghĩa là giá trị tại Y của một dòng bất kỳ trong A phải là một trong các y_1, y_2, ..., y_k). Gọi n_{y_1} là số dòng trong bảng A thỏa điều kiện Y = y_1, ký hiệu tương tự cho n_{y_2}, ..., n_{y_k}. Đương nhiên ta có các n_{y_i} phải lớn hay bằng 0 và (n_{y_1} + n_{y_2} + ...+ n_{y_k})= n.

Định nghĩa 1: Độ phân tán thông tin của bảng A là một giá trị trong khoảng từ 0 đến 1, được tính bởi:

$$I(n_{y_1}, n_{y_2}, ..., n_{y_k}) =$$
$$-\frac{n_{y_1}}{n_{y_1}+n_{y_2}+...+n_{y_k}}\log_k\frac{n_{y_1}}{n_{y_1}+n_{y_2}+...+n_{y_k}}$$
$$-\frac{n_{y_2}}{n_{y_1}+n_{y_2}+...+n_{y_k}}\log_k\frac{n_{y_2}}{n_{y_1}+n_{y_2}+...+n_{y_k}}$$
$$...$$
$$-\frac{n_{y_k}}{n_{y_1}+n_{y_2}+...+n_{y_k}}\log_k\frac{n_{y_k}}{n_{y_1}+n_{y_2}+...+n_{y_k}}$$

Trong đó, ta qui ước $\log_k 0=0$.

Nhận xét:

- Hàm I không thay đổi giá trị khi ta hoán vị các n_{y_i}.

- Hàm I đạt giá trị lớn nhất (bằng 1) khi $n_{y_1} = n_{y_2} = ... = n_{y_k}$, nghĩa là các dòng trong bảng A được phân tán đều cho các trường hợp (rời rạc) của thuộc tính output Y.

- Hàm I đạt giá trị nhỏ nhất (bằng 0) khi có một n_{y_i} nào đó bằng n (tổng số dòng của bảng A), và đương nhiên là các n_{y_i} còn lại phải bằng 0. Khi đó, ta nói rằng bảng A không phân tán thông tin gì cả, và cũng có nghĩa là bằng A không có gì để dự báo.

Định nghĩa 2: Gọi n_{y_m} là một giá trị lớn nhất trong các n_{y_1}, n_{y_2}, ..., n_{y_k}, khi đó ta gọi y_m là giá trị trội của thuộc tính output Y; độ tin cậy của luật $1 \to (Y=y_m)$ được gọi là độ trội output của bảng A.

Nhận xét: $Conf(1 \to (Y=y_m)) = \frac{n_{y_m}}{n}$.

Định nghĩa 3: Gọi X là một thuộc tính input của bảng A, giả sử X đã được rời rạc hóa thành m giá trị x_1, x_2, ..., x_m. Phép tách A dựa vào thuộc tính X, ký hiệu là T_X, tạo thành m bảng con của A:

$T_X = \{A_1, A_2, ..., A_m\}$, trong đó:

- $A_1, A_2, ..., A_m$ tạo thành một phân hoạch trên A, nghĩa là $A_i \cap A_j = \emptyset$, $\forall i, j = 1,2,...,m, i \neq j$ và $\bigcup_{i}^{m} A_i = A$.
- A_i là tập hợp các dòng trong A có giá trị tại X là x_i, nghĩa là $A_i = \{t \in A | t.X = x_i\}$, $\forall i = 1, 2, ..., m$.

Định nghĩa 4: Gọi T_X là một phép tách như trong định nghĩa 2. Với mọi i từ 1 đến m, gọi $n_{y_1}^{A_i}$ là số dòng trong bảng A_i thỏa điều kiện $Y = y_1$, ký hiệu tượng tự cho $n_{y_2}^{A_i}, ..., n_{y_k}^{A_i}$.

Độ phân tán thông tin của phép tách T_X, ký hiệu $E(T_X)$, là một giá trị từ 0 đến 1, được tính bởi:

$$E(T_X) = \sum_{i=1}^{m} \left(\frac{\sum_{j=1}^{k} n_{y_j}^{A_i}}{\sum_{j=1}^{k} n_{y_j}} \times I(n_{y_1}^{A_i} n_{y_2}^{A_i}, ..., n_{y_k}^{A_i}) \right)$$

Trong đó:

- $n_{y_j}^{A_i}$ là số dòng trong bảng A_i thỏa điều kiện $Y=y_j$.
- $\sum_{j=1}^{k} n_{y_j}^{A_i}$ là số dòng của bảng A_i.
- $\sum_{j=1}^{k} n_{y_j}$ là số dòng của bảng A.
- $I(n_{y_1}^{A_i} n_{y_2}^{A_i}, ..., n_{y_k}^{A_i})$ là độ phân tán thông tin của bảng A.

Một phép tách T_X được gọi là "tốt" khi các bảng con A_i tạo thành có độ phân tán thông tin thấp, hay nói theo nghĩa của phương pháp gom cụm, các bảng con A_i là các cụm có đa số phần tử (dòng) có giá trị tại Y giống nhau. Từ đó, phép tách T_X là tốt khi $E(T_X)$ thấp, và ngược lại.

1.3.3. Thuật toán

Input:

- Bảng dữ liệu A gồm n dòng với các thuộc tính ($X_1, X_2, ..., X_N, Y$), trong đó Y là thuộc tính Output (thuộc tính cần dự báo) và $X_1, X_2, ..., X_N$ là các thuộc tính input. Tất cả thuộc tính của A đều có giá trị rời rạc.

- w: ngưỡng độ tin cậy chấp nhận được.

Output:

- Cây quyết định.

CHƯƠNG 2: PHÂN TÍCH THIẾT KẾ HỆ THỐNG QUẢN LÝ BÁN HÀNG TẠI CÔNG TY DƯỢC TW3

2.1. PHÁT BIỂU BÀI TOÁN

2.1.1. Định nghĩa bài toán

Thị trường cung cấp dược phẩm, các thiết bị y tế ngày càng phát triển mạnh mẽ, các công ty kinh doanh về lĩnh vực này liên tục đưa ra các sản phẩm, các hình thức kinh doanh mới nhằm thu hút người tiêu dùng. Để công ty có thể tồn tại, phát triển và cạnh tranh trên thị trường được thì cần phải đưa ra những nhận định kịp thời, người quản lý có cách nhìn tổng thể về cách thức đầu tư về mặt hàng nào nhằm đáp ứng đúng yêu cầu của khách hàng và có hướng đầu tư đúng đắn.

Với mục đích phát triển công ty thành một nhà cung cấp dược phẩm có quy mô lớn thì việc ứng dụng công nghệ thông tin vào công tác quản lý là sự lựa chọn hàng đầu của nhà quản lý. Phạm vi ứng dụng và vai trò của công nghệ thông tin trong công tác quản lý là rất lớn, nhưng vì thờ gian và điều kiện còn hạn chế nên tôi chọn một khía cạnh nhỏ trong công tác quản lý đó là xây dựng hệ hỗ trợ khai phá dữ liệu dựa trên các thông tin giao dịch trên hóa đơn, hỗ trợ cho người quản lý đưa ra những nhận định mang tính chất chiến lược trong kinh doanh. Bên cạnh đó, luận văn còn có thể đáp ứng một số chức năng giúp nhà quản lý có thể xem và đánh giá thông qua các danh mục khách hàng, các hóa đơn bán lẻ hàng ngày, các hàng hóa có trong kho, tính được doanh thu, lợi nhuận qua các tháng… Giải quyết được một số tác nghiệp và điều quan trọng là ứng dụng khai phá dữ liệu luật kết hợp để đưa ra các quyết định, nó bao gồm nhiều bảng thống kê mang tính chất nhận định, giúp ta có cách nhìn tổng quan về dữ liệu, dự đoán ra các quy luật để qua đó kiểm chứng lại những nhận định này.

Khai phá mối quan hệ về lợi nhuận của các loại hàng có trong hóa đơn, dự đoán kết quả ảnh hưởng của các loại hàng này như thế nào? Khách hàng liệu có thói quen mua hàng này hay không? Từ các quy luật đó, ta đánh giá và kiểm định lại độ tin cậy có chính xác không? Có được nhận định đúng sẽ dễ dàng giúp nhà kinh doanh tìm ra hướng đầu tư cho các loại mặt hàng được tốt nhất.

Bài toán cụ thể được nêu ra ở đây là: ứng dụng khai phá dữ liệu, cụ thể là dựa vào thuật toán phân lớp với cây quyết định để tìm luật kết hợp tin cậy dựa trên mã các loại hàng và dựa trên doanh thu của các loại hàng để đưa ra những đánh giá, những nhận định về sự ảnh hưởng của các loại hàng đến doanh thu và lợi nhuận của công ty.

2.1.2. Qui trình bán hàng

Hệ thống bán hàng được thực hiện theo một qui trình như sau:

- Bộ phận trình dược viên giới thiệu danh mục hàng hóa đến cho khách hàng.
- Khách hàng chọn các mặt hàng cần mua (hay còn gọi là đặt hàng).
- Bộ phận trình dược viên gởi yêu cầu đặt mua đến cho bộ phận quản lý bán hàng.
- Bộ phận quản lý bán hàng gửi yêu cầu đặt mua đến cho bộ phận quản lý vật tư (kho). Bộ phận vật tư hồi đáp cho biết danh mục mặt hàng khách hàng đặt mua có tồn kho hay không.
- Nếu kho vật tư còn hàng, bộ phận quản lý bán hàng yêu cầu bộ phận quản lý kho xuất kho (lập phiếu xuất kho) và yêu cầu bộ phận tài chính lập phiếu thu tiền khách hàng.
- Nếu khách hàng yêu cầu mua hàng trả chậm thì bộ phận quản lý bán hàng gửi yêu cầu công nợ đến bộ phận quản lý công nợ, nếu được bộ phận quản lý công nợ chấp nhận thì bộ phận quản lý bán hàng sẽ tra sổ công nợ khách hàng, thêm mục nợ mới đồng thời yêu cầu bộ phận quản lý vật tư đánh dấu chưa thanh toán vào phiếu xuất kho.

2.2. PHÂN TÍCH THIẾT KẾ

2.2.1. Cơ sở dữ liệu

Ký hiệu chữ viết :

P: Primary key (khoá chính)

U: Unique key, candidate key (khoá chỉ định)

M : Mandatory (không được rỗng)

L : Locked (không cho phép sửa đổi giá trị)

- Loại thực thể Người dùng (NguoiDung)

Thuộc tính	**Kiểu**	**Kích thước**	**P**	**U**	**M**	**L**
Tennguoidung	nvarchar	50	x	x	x	x
Matkhau	nvarchar	50			x	
Vaitro	int	4			x	

- Loại thực thể Khách hàng (KhachHang)

Thuộc tính	**Kiểu**	**Kích thước**	**P**	**U**	**M**	**L**
Makh	nvarchar	10	x	x	x	x
Tenkh	nvarchar	50			x	
Diachi	nvarchar	50			x	
DienThoai	nvarchar	50			x	

- Loại thực thể Hóa đơn (HoaDon)

Thuộc tính	**Kiểu**	**Kích thước**	**P**	**U**	**M**	**L**
Mahd	nvarchar	10	x	x	x	x
Makh	nvarchar	50			x	
Ngaylap	datetime	8			x	
Tonggiatri	float	10			x	

- Loại thực thể Hàng hóa (HangHoa)

Thuộc tính	**Kiểu**	**Kích thước**	**P**	**U**	**M**	**L**
Mahang	nvarchar	10	x	x	x	x
Tenhang	nvarchar	50			x	
Dongia	float	10			x	
Soluong	int	10			x	

Maloai	nvarchar	10			x	

- Loại thực thể Loại hàng (LoaiHang)

Thuộc tính	Kiểu	Kích thước	P	U	M	L
Maloai	nvarchar	10	x	x	x	x
Tenloai	nvarchar	50			x	

- Loại thực thể Các tháng (CacThang)

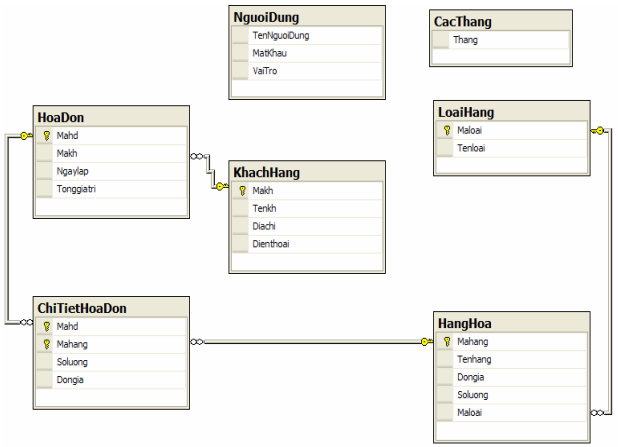
Thuộc tính	Kiểu	Kích thước	P	U	M	L
Thang	Int	4			x	

Dữ liệu Các tháng bao gồm 12 tháng trong năm.

- Sự kết hợp Chi tiết hóa đơn (ChiTietHoaDon)

Thuộc tính	Kiểu	Kích thước	P	U	M	L
Mahd	nvarchar	10	x	x	x	x
Mahang	nvarchar	10	x	x	x	x
Soluong	int	10			x	
Dongia	float	10			x	

⇨ Sơ đồ mối quan hệ của các thực thể



Hình 2.1. Mô hình cơ sở dữ liệu

2.2.2. Xác định các tác nhân

Dựa vào phần định nghĩa bài toán, ta có thể xác định được các tác nhân chính của hệ thống như sau:

TRÌNH DƯỢC VIÊN: là người giới thiệu các mặt hàng, thực hiện việc mua hàng và gởi các đơn đặt hàng cho người quản lý.

KHÁCH HÀNG: là người giao dịch với hệ thống thông qua các đơn đặt hàng, khách hàng có thể chọn lựa các mặt hàng mình muốn thông qua sự giới thiệu của trình dược viên.

NGƯỜI QUẢN LÝ: là người điều hành, quản lý và theo dõi mọi hoạt động của hệ thống.

NGƯỜI DÙNG: bao gồm người quản lý, trình dược viên và những khách hàng đã được cập nhật thông qua các đơn đặt hàng. Ứng với mỗi thành viên sẽ có những chức năng khác nhau nhằm phục vụ cho công việc cụ thể cho từng đối tượng.

2.2.3. Xác định các UC, các gói UC và xây dựng biểu đồ UC chi tiết

2.2.4. Đặc tả các Use Case

2.2.5. Xác định các lớp thực thể và các lớp biên

2.2.6. Biểu đồ hoạt động của các Use Case

2.2.7. Mô hình hóa tương tác trong các Use Case: Biểu đồ tuần tự

CHƯƠNG 3: XÂY DỰNG CHƯƠNG TRÌNH VÀ THỰC NGHIỆM

3.1. NGÔN NGỮ LẬP TRÌNH

Chọn lập trình trên Window Form C# để xây dựng chương trình (dùng công cụ Microsoft Visual Studio 2008).

Cơ sở dữ liệu chọn là SQL – dùng phiên bản SQL Server 2005 Developer Edition để tiện cho công việc khai phá dữ liệu.

3.2. DỮ LIỆU THỰC TẾ THU THẬP ĐƯỢC

- Dữ liệu được thu thập thực tế tại công ty dược dựa vào thông tin trên các hóa đơn. Thông tin trên các hóa đơn bao gồm thông tin khách hàng, loại hàng, mặt hàng, số lượng, đơn giá thuốc bán ra. Bên

cạnh đó, còn thu thập thêm thông tin số lượng hiện có trong kho, đơn giá gốc, dữ liệu này giúp người quản lý có thể nắm bắt được doanh thu bán hàng, lợi nhuận thu được từ các mặt hàng bán được.

- Khách hàng: Dữ liệu thông tin khách hàng bao gồm khoảng 2160 khách hàng thường xuyên giao dịch với công ty, bảng dữ liệu khách hàng bao gồm mã khách hàng, tên khách hàng, địa chỉ và số điện thoại của khách hàng

- Hóa đơn: Dữ liệu Hóa đơn bao gồm khoảng hơn 3.000 hóa đơn, dữ liệu hóa đơn bao gồm mã hóa đơn, mã khách hàng, ngày lập hóa đơn và tổng giá trị trên hóa đơn đó

- Chi tiết hóa đơn: Dữ liệu Chi tiết hóa đơn bao gồm khoảng hơn 12.000 bảng ghi, dữ liệu này bao gồm mã hóa đơn, mã loại hàng, số lượng và đơn giá mặt hàngHàng hóa: Dữ liệu Hàng hóa bao gồm khoảng hơn 189 mặt hàng, dữ liệu này bao gồm mã hàng, tên hàng, đơn giá, số lượng và mã loại hàng

- Loại hàng: bao gồm khoảng 82 loại hàng, dữ liệu này bao gồm mã loại và tên loại hàng được thể hiện ở màn hình bên dưới:

Table - dbo.LoaiHang | Summary

Maloai	Tenloai
1	Thuốc tránh thai
10	Viên nhiễm trong và sau phẫu thuật
11	Thuốc sát trùng
12	Thuốc trị ho khan ,ho có đờm,ho lâu ngày
13	Tiêu nhầy trong bệnh nhầy nhớt,bệnh hô hấp
14	Hạ sốt,giảm đau
15	Trị viêm khớp
16	Bs vitamin C, trị bệnh scorbut tăng sức đề kháng
17	Trị rối loạn tuần hoàn não và mạch máu ngoại biên
18	Chóng mặt,rối loạn thần kinh trung ương,thiếu máu
19	Thuốc điều trị cao huyết áp
2	Trị bệnh tâm thần
20	Điều trị viêm đại tràng
21	Tăng cường sức khỏe,yếu sinh lý,mãn dục nam
22	Lợi tiểu, lợi mật,trị sỏi thận,sỏi mật
23	Trị bệnh răng miệng
24	Thuốc an thần
25	Dạ dày,huyết áp,hạ cholesteron,đái tháo đường
26	Thuốc an thần gây ngủ
27	Trị đau bụng tiêu chảy,nhiễm khuẩn đường ruột
28	Trị thiếu Mg,yếu cơ,rối loạn thần kinh
29	Trị ngạt mũi,cảm cúm,ho,đau bụng,trặc gân,sưng...
3	Dịch truyền
30	Thuốc chống nôn, say tàu xe
31	Phòng, trị thiếu vitamin nhóm B,mệt mỏi tinh thần
32	Trị viêm mũi dị ứng,hắt hơi,chảy nước mũi

80 of 82

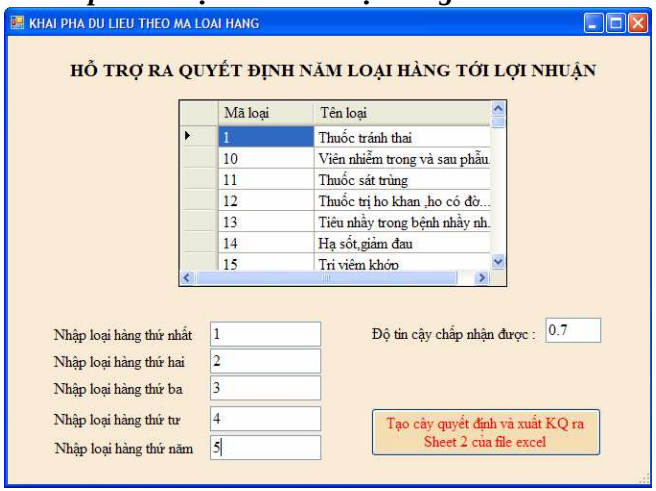
3.3. THIẾT KẾ GIAO DIỆN CHƯƠNG TRÌNH

3.3.1. Form Đăng nhập

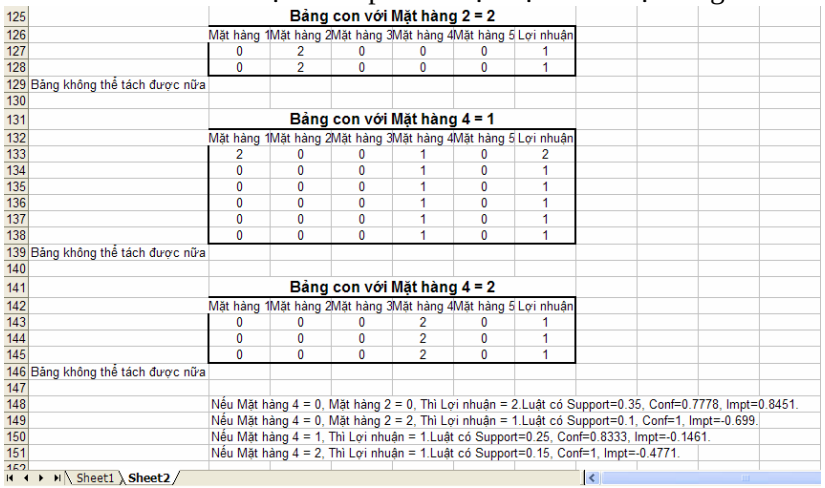
3.3.2. Giao diện chính

3.3.3. Chức năng chính

3.3.3.1. Khai phá dữ liệu theo mã loại hàng

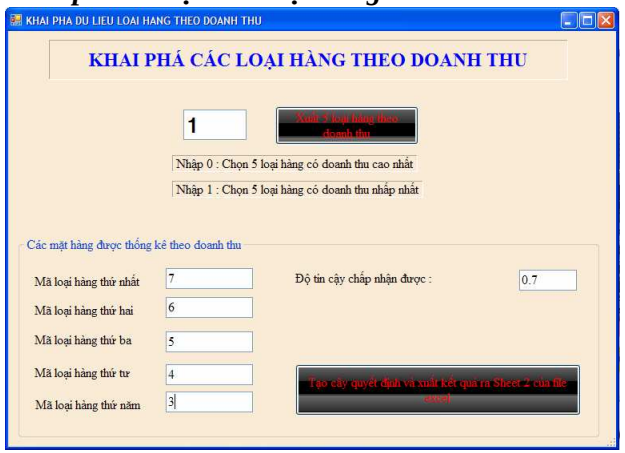


Hình 3.25. Giao diện Khai phá dữ liệu dựa theo loại hàng hóa

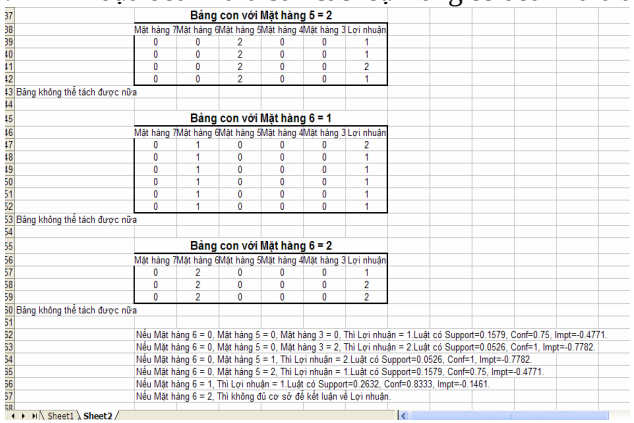


Hình 3.26. Kết quả sau khi khai phá dữ liệu dựa theo loại hàng hóa

3.3.3.2. Khai phá dữ liệu các loại hàng theo doanh thu



Hình 3.27. KPDL dựa doanh thu bán các loại hàng có doanh thu thấp nhất



Hình 3.28. Kết quả sau khi khai phá dữ liệu mã loại hàng có doanh thu thấp nhất

3.3.4. Chức năng hỗ trợ

3.3.4.1. Danh mục khách hàng

Cho ta biết được các thông tin về khách hàng như Mã khách hàng , Tên khách hàng, Địa chỉ , Điện Thoại..

3.3.4.2. Danh mục hóa đơn

Dùng để quản lý các hóa đơn và chi tiết hóa đơn của tất cả các khách hàng , ta có thể thêm xóa sửa một hóa đơn cho một khách hàng cũng như các chi tiết hóa đơn của một hóa đơn.

3.3.4.3. Danh mục sản phẩm

Là mục để quản lý tất cả các loại hàng hóa và các hàng hóa có trong loại hàng đó , cũng có chức năng thêm xóa sửa một loại hàng mới hoặc 1 sản phẩm mới.

3.3.4.4. Thống kê giao dịch

3.3.4.5. Quản lý doanh thu

Mục Quản Lý Doanh Thu này hiển thị cho ta thấy được tất cả các Hóa Đơn và Danh sách các Chi Tiết Hóa Đơn của Hóa Đơn đó trong khoảng thời gian cụ thể .Từ đó tính được các Doanh Thu & Lợi Nhuận trong khoảng thời gian các tháng hoặc năm .Từng mặt hàng , loại hàng đã bán đem lại lợi nhuận và dannh thu ra sao trong khoảng thời gian đó hoặc năm đó.

3.3.4.6. Tình hình biến động giá

Mỗi loại hàng hóa đem lại cho ta doanh thu khác nhau , thống kê được các biến động về giá của chúng ta sẽ có cái nhìn tổng quan hơn các mặt hàng trong siêu thị cần được đầu tư : Ta thống kê các loại hàng có độ lệch chuẩn theo doanh thu trong năm.

3.4. KẾT QUẢ THỬ NGHIỆM VÀ NHẬN XÉT

Kết quả khai phá luật kết hợp bằng kỹ thuật phân lớp với cây quyết định trên bảng doanh thu gồm 352 giao dịch, mỗi giao dịch gồm có 6 thuộc tính.

Kết quả đạt được ứng với 5 mã loại hàng lần lượt 1, 2, 3, 4, 5 như sau :

STT	Ngưỡng tin cậy cho trước	Số giao dịch	Số luật thu được
1	0.6	352	12
2	0.7	352	47
3	0.8	352	59
4	0.9	352	67

Rời rạc các thuộc tính trong bảng trên theo phương thức sau :

- Các loại hàng : loại hàng 1, loại hàng 2, loại hàng 3,... được rời rạc theo trung bình doanh thu
- Nếu là 0: doanh thu bằng 0
- Nếu là 1: có doanh thu thấp hơn mức trung bình doanh thu
- Nếu là 2: có doanh thu cao hơn mức trung bình doanh thu
- Lợi nhuận:
- Nếu là 1: lợi nhuận thấp hơn mức trung bình lợi nhuận.
- Nếu là 2: lợi nhuận cao hơn mức trung bình lợi nhuận.
- Bảng kết quả sau khi đã rời rạc các thuộc tính sẽ được xuất ra file excel tại Sheet1.
- Giả sử ta cho một ngưỡng tin cậy cho trước là 0.6, và thử nghiệm với 5 loại mặt hàng lần lượt như sau: 1, 2, 3, 4, 5 với số giao dịch là 340 ta sẽ có những tập luật như sau:
- Nếu mã loại hàng 5 có doanh thu bằng 0 và mã loại hàng 1 có doanh thu thấp hơn mức trung bình doanh thu thì khi đó lợi nhuận thu được có thể sẽ cao hơn mức trung bình lợi nhuận.
- Nếu mã loại hàng 5 có doanh thu bằng 0 và mã loại hàng 1 có doanh thu cao hơn mức trung bình doanh thu thì khi đó lợi nhuận thu được sẽ thấp hơn mức trung bình trung bình lợi nhuận.
- Nếu mã loại hàng 1, mã loại hàng 3, mã loại hàng 5 có doanh thu đồng thời bằng 0 thì khi đó lợi nhuận thu được sẽ cao hơn mức trung bình lợi nhuận.
- Nếu mã loại hàng 1, mã loại hàng 5 có doanh thu đồng thời bằng 0 và mã loại hàng 3 có doanh thu thấp hơn mức trung bình doanh thu thì khi đó lợi nhuận thu được sẽ thấp hơn mức trung bình lợi nhuận.
- Nếu mã loại hàng 5 có doanh thu thấp hơn mức trung bình doanh thu thì lợi nhuận thu được sẽ có thể thấp hơn mức trung bình lợi nhuận.

- Nếu mã loại hàng 5 có doanh thu cao hơn mức trung bình doanh thu và mã loại hàng 4 có doanh thu có doanh thu bằng 0 thì lợi nhuận thu được sẽ có thể cao hơn mức trung bình lợi nhuận.
- Nếu mã loại hàng 5 có doanh thu cao hơn mức trung bình doanh thu và mã loại hàng 4 có doanh thu thấp hơn mức trung bình doanh thu thì khi đó lợi nhuận thu được sẽ có thể thấp hơn mức trung bình lợi nhuận.
- Nếu mã loại hàng 4 và mã loại hàng 5 đồng thời có doanh thu cao hơn mức trung bình doanh thu và mã loại hàng 2 có doanh thu bằng 0 thì lợi nhuận thu được khi đó có thể sẽ cao hơn mức trung bình lợi nhuận.
- Nếu mã loại hàng 4, mã loại hàng 5 đồng thời có doanh thu cao hơn mức trung bình doanh thu và mã loại hàng 1, mã loại hàng 2 đồng thời có doanh thu thấp hơn mức trung bình doanh thu thì khi đó lợi nhuận thu được sẽ có thể thấp hơn mức trung bình lợi nhuận.
- Nếu mã loại hàng 2, mã loại hàng 4, mã loại hàng 5 đồng thời có doanh thu cao hơn mức trung bình doanh thu và mã loại hàng 3 có doanh thu thấp hơn mức trung bình doanh thu thì khi đó lợi nhuận thu được có thể sẽ cao hơn mức trung bình lợi nhuận.
- Nếu mã loại hàng 2, mã loại hàng 3, mã loại hàng 4, mã loại hàng 5 đồng thời có doanh thu cao hơn mức trung bình doanh thu và mã loại hàng 1 có mức doanh thu thấp hơn mức trung bình doanh thu thì khi đó lợi nhuận thu được sẽ thấp hơn mức trung bình lợi nhuận.
- Nếu mã loại hàng 1, mã loại hàng 2, mã loại hàng 3, mã loại hàng 4, mã loại hàng 5 đồng thời có doanh thu cao hơn mức trung bình doanh thu thì khi đó lợi nhuận thu được sẽ cao hơn mức trung bình lợi nhuận.

NHẬN XÉT

- Qua các lần chạy thử mô hình, ta thấy Conf = 1 ở bất cứ mọi giá trị, chứng tỏ độ tin cậy của các luật là tốt. Bên cạnh đó độ

phổ biến cũng chênh lệch trong một khoảng các giá trị xác định cho thấy mức độ dao động giữa các luật không cao, có thể chấp nhận được nhiều luật cùng một lúc.

- Càng tăng chỉ số độ tin cậy thì số luật cũng thay đổi không đáng kể nên chứng tỏ rằng các luật đều đã mang tính chất liên kết nhau cao. Ta thấy có một số luật luôn xuất hiện ở các mô hình chạy thử mà ta có thể tin tưởng được.

KẾT LUẬN

1. Đánh giá kết quả

- Về mặt lý thuyết: Nghiên cứu kiến thức về khai phá tri thức và khai phá dữ liệu, các thuật toán tìm luật kết hợp như: Apriori, Apriori-TIP, Apriori-Hybrid, FP-Growth, phân lớp với cây quyết định. Cài đặt thuật toán tìm luật kết hợp bằng phương pháp phân lớp với cây quyết định.

- Về mặt ứng dụng: Xây dựng được hệ thống hỗ trợ đưa ra các quyết định phục vụ cho công tác quản lý thông qua việc khai phá dữ liệu dựa trên loại hàng và doanh thu loại hàng có ở công ty.

2. Hạn chế

- Chỉ mới minh họa hệ thống trên cơ sở dữ liệu của công ty TNHH MTV Dược TW3, chưa minh họa trên nhiều cơ sở dữ liệu khác.

- Hệ thống còn đơn giản, chưa có nhiều chức năng thiết thực giúp phục vụ hiệu quả công tác quản lý của công ty.

3. Hướng phát triển

- Tiếp tục hoàn thiện đề tài, xây dựng hệ thống nhiều chức năng hơn, dùng thuật toán phân lớp với cây quyết định thử nghiệm và đánh giá kỹ hơn trên cơ sở dữ liệu lớn hơn và cơ sở dữ liệu khác.

- Đưa thêm các phương pháp khai phá dữ liệu khác vào việc phân tích mô hình, như gom cụm để phân lớp dữ liệu để từ đó có thể phân tích dữ liệu chính xác hơn và đưa ra những luật có độ tin cậy cao hơn.

- Khai phá dữ liệu trên kho dữ liệu với các luật kết hợp đa chiều, nhiều mức.

- Tìm hiểu công cụ hỗ trợ hiển thị kết quả thuật toán ở dạng đồ họa như đồ thị, biểu đồ, …