BỘ GIÁO DỤC VÀ ĐÀO TẠO

ĐẠI HỌC ĐÀ NẴNG

LÊ VĂN ĐÔNG

NGHIÊN CỨU LUẬT KẾT HỢP VÀ ỨNG DỤNG TRONG CÔNG TÁC QUẢN LÝ KHO HÀNG TẠI SIÊU THỊ METRO

Chuyên ngành : **KHOA HỌC MÁY TÍNH**

Mã số : **60.48.01**

TÓM TẮT LUẬN VĂN THẠC SĨ KỸ THUẬT

Đà Nẵng - Năm 2011

Công trình được hoàn thành tại

ĐẠI HỌC ĐÀ NẴNG

Người hướng dẫn khoa học: **PGS.TSKH TRẦN QUỐC CHIẾN**

Phản biện 1 : **TS. HUỲNH CÔNG PHÁP**

Phản biện 2 : **TS. TRƯƠNG CÔNG TUẤN**

Luận văn được bảo vệ tại Hội đồng chấm Luận văn tốt nghiệp thạc sĩ kỹ thuật họp tại Đại học Đà Nẵng vào ngày 10 tháng 09 năm 2011.

Có thể tìm hiểu luận văn tại:

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

MỞ ĐẦU

1. Lý do chọn đề tài

Trong những năm gần đây, sự phát triển mạnh mẽ của công nghệ thông tin đã làm cho khả năng thu thập và lưu trữ thông tin của hệ thống thông tin tăng một cách nhanh chóng. Bên cạnh đó, việc tin học hóa một cách ồ ạt và nhanh chóng các hoạt động sản xuất, kinh doanh cũng như nhiều lĩnh vực hoạt động khác đã tạo cho chúng ta một lượng dữ liệu cần lưu trữ và xử lý khổng lồ.

Trong bối cảnh đó, việc nghiên cứu đề ra các phương pháp, công cụ mới hỗ trợ con người khám phá, phân tích, tổng hợp thông tin nhằm để tìm và rút ra các tri thức hữu ích, các qui luật tiềm ẩn hỗ trợ tiến trình ra quyết định là một nhu cầu bức thiết. Từ đó giúp cho nhà quản lý có cái nhìn tổng quan hơn về dữ liệu, có thể đưa ra những nhận định, quyết định, những dự đoán mang tính chiến lược nhất.

Hiện nay vấn đề khai phá luật kết hợp chỉ mới được đề cập và đang trở thành một khuynh hướng quan trọng của khai phá dữ liệu. Luật kết hợp là luật ngầm định một số quan hệ kết hợp giữa một tập các đối tượng mà các đối tượng có thể độc lập hoàn toàn với nhau. Do đây là một hướng đi tiềm năng, có nhiều khả năng phát triển trong tương lai, nên em đã chọn đề tài : ***"Nghiên cứu luật kết hợp và ứng dụng trong công tác quản lý kho hàng tại siêu thị Metro"*** trong đợt thực hiện Luận văn tốt nghiệp này.

2. Đối tượng và phạm vi nghiên cứu

Đối tượng

- Lý thuyết
  - Kỹ thuật khai phá dữ liệu
  - Nghiệp vụ quản lý kho hàng trong Siêu thị
- Dữ liệu
  - Cơ sở dữ liệu: các mặt hàng, khách hàng . . .
  - Các văn bản, qui định liên quan đến công tác quản lý trong siêu thị.
- Công nghệ
  - Công cụ lập trình: Visual Studio C#.
  - Cơ sở dữ liệu: Microsoft SQL Server 2005

Phạm vi

- Nghiên cứu các kiến thức cơ bản về phương pháp phát hiện luật kết hợp
- Nghiên cứu các quá trình tác nghiệp trong hệ thống
- Xây dựng Hệ hỗ trợ ra quyết định phục vụ cho công tác quản lý.

3. Mục tiêu và nhiệm vụ

Mục tiêu

- Ứng dụng luật kết hợp vào công tác quản lý kho hàng.
- Giúp cho nhà quản lý có thể đưa ra những nhận định, những dự đoán mang tính chiến lược.

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 tập mục phổ biến, luật kết hợp, luật phân lớp, luật gom cụm dữ liệu.
- Ứng dụng các thuật toán trên vào cơ sở dữ liệu quản lý kho 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 cáo khoa học. . .

❖ Lập kế hoạch, lên quy trình, tiến độ thực hiện

❖ Nghiên cứu kỹ thuật khai phá dữ liệu bằng luật kết hợp vào việc quản lý kho hàng tại siêu thị.

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

Ý nghĩa khoa học

❖ Ứng dụng tin học trong công tác quản lý.

Ý nghĩa thực tiễn

❖ Giải quyết được các công việc tác nghiệp

❖ Hỗ trợ đưa ra các quyết định, các dự đoán mang tính chiến lược cho người quản lý.

❖ Giúp nhà quản lý có cái nhìn tổng quan về dữ liệu.

6. Tên đề tài

“NGHIÊN CỨU LUẬT KẾT HỢP VÀ ỨNG DỤNG TRONG CÔNG TÁC QUẢN LÝ KHO HÀNG TẠI SIÊU THỊ METRO”

7. Cấu trúc luận văn

Nội dung chính của luận văn được chia thành 2 chương như sau:

- *Chương 1:* Cơ sở lý thuyết về khai phá dữ liệu và luật kết hợp.
- *Chương 2:* Ứng dụng khai phá luật kết hợp trong công tác quản lý kho hàng tại siêu thị .

CHƯƠNG 1
CƠ SỞ LÝ THUYẾT VỀ KHAI PHÁ DỮ LIỆU VÀ LUẬT KẾT HỢP

1.1. TỔNG QUAN VỀ KHAI PHÁ DỮ LIỆU

1.1.1. Định nghĩa khai phá dữ liệu

Khai phá dữ liệu là tiến trình khám phá tri thức tiềm ẩn trong các CSDL, cụ thể hơn, đó là tiến trình lọc, sản sinh những tri thức hoặc các mẫu tiềm ẩn, chưa biết, những thông tin hữu ích từ các CSDL lớn.

1.1.2. Các ứng dụng của khai phá dữ liệu

Phát hiện tri thức và khai phá dữ liệu liên quan đến nhiều ngành, nhiều lĩnh vực: thống kê, trí tuệ nhân tạo, CSDL, thuật toán, tính toán song song… Đặc biệt phát hiện tri thức và khai phá dữ liệu rất gần gũi với lĩnh vực thống kê, sử dụng các phương pháp thống kê để mô hình hóa dữ liệu và phát hiện các mẫu. Khai phá dữ liệu có nhiều ứng dụng trong thực tế, ví dụ như: Bảo hiểm, tài chính và thị trường chứng khoán; Thống kê, phân tích dữ liệu và hỗ trợ ra quyết định; Điều trị y học và chăm sóc y tế; Sản xuất và chế biến; Text mining và Web mining; Lĩnh vực khoa học. . .

1.1.3. Các bước của quy trình khai phá dữ liệu

Quy trình khai phá dữ liệu thường tuân theo các bước sau:
Bước thứ nhất: Hình thành, xác định và định nghĩa bài toán
Bước thứ hai: Thu thập và tiền xử lý dữ liệu
Bước thứ ba: Khai phá dữ liệu, rút ra các tri thức
Bước thứ tư: Phân tích và kiểm định kết quả
Bước thứ năm: Sử dụng các tri thức phát hiện được

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

1.1.4. Nhiệm vụ chính trong khai phá dữ liệu

Quá trình khai phá dữ liệu là quá trình phát hiện ra mẫu thông tin. Trong đó giải thuật khai phá tìm kiếm các mẫu đáng quan tâm theo dạng xác định như các luật, phân lớp, hồi quy, cây quyết định, …

1.1.4.1. Phân lớp (phân loại – classification)

1.1.4.2. Hồi quy (regression)

1.1.4.3. Phân nhóm (clustering)

1.1.4.4. Tổng hợp (summarization)

1.1.4.5. Mô hình hóa sự phục thuộc (dependency modeling)

1.1.4.6. Phát hiện sự biến đổi và độ lệch (change and deviation dectection)

1.1.5. Các phương pháp khai phá dữ liệu

1.1.5.1. Các thành phần của giải thuật khai phá dữ liệu

1.1.5.2. Phương pháp suy diễn/ quy nạp

1.1.5.3. Phương pháp ứng dụng K – láng giềng gần

1.1.5.4. Phương pháp sử dụng cây quyết định và luật

1.1.5.5. Phương pháp phát hiện luật kết hợp

1.1.6. Lợi thế của khai phá dữ liệu so với các phương pháp cơ bản

1.1.6.1. Học máy (Machine Learning)

1.1.6.2. Phương pháp hệ chuyên gia

1.1.6.3. Phát kiến khoa học

1.1.6.4. Phương pháp thống kê

1.1.7. Lựa chọn phương pháp

1.1.8. Thách thức trong ứng dụng và nghiên cứu kỹ thuật khai phá dữ liệu

Ở đây, ta đưa ra một số khó khăn trong việc nghiên cứu và ứng dụng kỹ thuật khai phá dữ liệu. Tuy nhiên, có khó khăn không có nghĩa là việc giải quyết là hoàn toàn bế tắc mà chỉ muốn nêu lên rằng để khai phá được dữ liệu không phải là đơn giản, mà phải xem xét cũng như tìm cách giải quyết những vấn đề này. Ta có thể liệt kê một số khó khăn sau:

1.1.8.1. Các vấn đề về CSDL

Đầu vào chủ yếu của một hệ thống khám phá tri thức là các dữ liệu thô cơ sở, phát sinh trong khai phá dữ liệu chính là từ đây. Do các dữ liệu trong thực tế thường động, không đầy đủ, lớn và bị nhiễu. Trong những trường hợp khác, người ta không biết CSDL có chứa các thông tin cần thiết cho việc khai phá hay không và làm thế nào để giải quyết với sự dư thừa những thông tin không thích hợp.

1.1.8.2. Một số vấn đề khác

- *"Quá phù hợp"*
- *Đánh giá tầm quan trọng thống kê*
- *Khả năng biểu đạt các mẫu*
- *Sự tương tác giữa người sử dụng và các tri thức sẵn có*

1.2. LUẬT KẾT HỢP TRONG KHAI PHÁ DỮ LIỆU

1.2.1. Vài nét về khai phá luật kết hợp

1.2.2. Một số định nghĩa cơ bản

Định nghĩa 1.1: Luật kết hợp

Hạng mục *(item)* là mặt hàng trong giỏ hàng hay một thuộc tính.

Tập các hạng mục *(itemset)* là tập các mặt hàng trong giỏ hàng hay tập các thuộc tính, $I = \{i_1, i_2, .., i_m\}$

Ví dụ : tập I = { sữa, bánh mì, ngũ cốc, sữa chua}

Giao dịch *(Transation)* là tập các hạng mục được mua trong một giỏ hàng *(có TID là mã giao dịch)*. Giao dịch t là tập các hạng mục sao cho $t \subseteq I$.

Ví dụ: t = {bánh mì, sữa chua, ngũ cốc}

Cơ sở dữ liệu giao dịch là tập các giao dịch, ví dụ cơ sở dữ liệu giao dịch $D = \{t_1, t_{2,} \ldots t_n\}$.

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 $X \cap Y = \emptyset$. Các tập X và Y được gọi là tập các hạng mục *(itemset)*. Tập X gọi là nguyên nhân, tập Y gọi là hệ quả.

Định nghĩa 1.2: Độ hỗ trợ

Độ hỗ trợ của tập các hạng mục X trong cơ sở dữ liệu giao dịch D là tỷ lệ giữa số các giao dịch chứa X trên tổng số các giao dịch trong D, ký hiệu là Support(X) hay Supp(X).

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

Độ hỗ trợ của một luật kết hợp $X \rightarrow Y$ sẽ là:

$Supp(X \rightarrow Y) = Supp(X \cup Y)$

Định nghĩa 1.3: Độ tin cậy

Độ tin cậy *(Confidence)* của luật kết hợp có dạng: X → Y là tỷ lệ giữa số lượng các giao dịch trong D chứa X ∪ Y với số giao dịch trong D có chứa tập X. Ký hiệu độ tin cậy của một luật là *Conf (X →Y)*.

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

- Việc khai thác các luật kết hợp có thể được phân tích thành hai giai đoạn sau:

1. Tìm tất cả các tập mục phổ biến từ CSDL D tức là tìm tất cả các tập mục có độ hỗ trợ lớn hơn hoặc bằng *minsupp*.
2. Sinh ra các luật từ các tập mục phổ biến *(large itemsets)* sao cho độ tin cậy của luật lớn hơn hoặc bằng *minconf*.

1.2.3. Ví dụ về bài toán phát hiện luật kết hợp

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

- Luật kết hợp nhị phân
- Luật kết hợp có thuộc tính số và thuộc tính hạng mục
- Luật kết hợp tiếp cận theo hướng tập thô
- Luật kết hợp nhiều mức
- Luật kết hợp mờ
- Luật kết hợp với thuộc tính được đánh trọng số
- Khai phá luật kết hợp song song

1.2.5. Một số thuật toán phát hiện luật kết hợp

1.2.5.1. Thuật toán AIS

1.2.5.2. Thuật toán SETM

1.2.5.3. Thuật toán Apriori

1.2.5.4. Thuật toán Apriori -TID

1.2.5.5. Thuật toán Apriori –Hybrid

1.2.5.6. Thuật toán FP-Growth

1.2.5.7. Thuật toán tìm luật kết hợp với cây quyết định

❖ ***Một số định nghĩa***

Định nghĩa 1.4 : Cây quyết định là một cấu trúc phân cấp của các nút và các nhánh. Trong đó có 3 loại nút trên cây :

- Nút gốc
- Nút nội bộ : mang tên thuộc tính của CSDL
- Nút lá : mang tên lớp

Một cây quyết định biểu diễn một phép tuyển của các kết hợp, của các ràng buộc đối với các giá trị thuộc tính.

Mỗi đường đi từ nút gốc đến nút lá sẽ tương ứng với một kết hợp của các kiểm tra giá trị thuộc tính.

** Phát biểu vấn đề :*

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 $(n_{y_1} + n_{y_2} + ... + n_{y_k}) = n$. Khi đó ta có các định nghĩa sau:

Định nghĩa 1.5 : 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 bảng A không có gì để dự báo.

Định nghĩa 1.6 : Gọi n_{y_m} là một giá trị lớn nhất trong các $n_{y_1}, n_{y_2}, ..., n_{y_k}$, và gọi y_m là giá trị trội của thuộc tính output Y, thì khi đó ta có độ trội output của bảng A sẽ là $\frac{n_{y_m}}{n}$

Định nghĩa 1.7 : 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 $\bigcup_{i=1}^{m} A_i = A$.
- 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 1.8 : Gọi T_X là một phép tách như trong *định nghĩa 1.7*. 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 con A_i.

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.

❖ ***Giải thuật xây dựng cây quyết định***

** Phát biểu bài toán:* 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. Tất cả thuộc tính của A đều có giá trị rời rạc và w là ngưỡng độ tin cậy chấp nhận được.

** 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*:

- Các luật sinh ra từ cây quyết định.

** Các bước thực hiện*:

Bước 1: Xác định thuộc tính X_m trong các X_1, X_2, ..., X_N thỏa E(T_{X_m}) là bé nhất.

Bước 2: Thực hiện phép tách T(X_m) trên bảng A, ta có tầng thứ nhất của cây quyết định với nút gốc là X_m.

Bước 3: Với mỗi bảng con A_i (tạo thành từ phép tách ở bước 2).

- Nếu bảng con có độ trội output lớn hơn hay bằng w thì bảng này chính là một nút lá của cây quyết định. Giá trị trội chính là kết luận tại nút lá, và độ trội output chính là độ tin cậy của kết luận.

- Nếu bảng con có độ trội output bé hơn w và mọi cột (mọi thuộc tính) đều chỉ có một giá trị hoặc bảng không có dòng nào (nghĩa là bảng không thể tách được nữa) thì bảng này cũng chính là một nút lá, và kết luận tại nút này là "Không đủ cơ sở để kết luận gì về output".

- Nếu bảng con này có độ trội output bé hơn w thì thực hiện lại thao tác tương tự như đã làm với bảng A ở bước 1, bước 2 và bước 3.

❖ ***Ưu điểm của cây quyết định***

❖ ***Chuyển đổi từ cây quyết định sang luật***

Tri thức trên cây quyết định có thể được rút trích và biểu diễn thành một dạng luật **IF – THEN *(NẾU – THÌ)***. Khi đã xây dựng được cây quyết định, ta có thể dễ dàng chuyển cây quyết định này thành một tập các luật tương đương, một luật tương đương với một đường đi từ gốc đến nút lá. Giai đoạn chuyển đổi từ cây quyết định sang luật thường bao gồm 4 bước sau :

- *Cắt tỉa*
- *Lựa chọn*
- *Sắp xếp*
- *Ước lượng, đánh giá*

❖ ***Ví dụ minh họa***

** Phát biểu bài toán :* 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? Giả sử doanh nghiệp dựa vào các thuộc tính (của khách hàng) để chọn đặc trưng gồm: *Tuổi, giới tính, khoảng thu nhập, TT Hôn nhân.* Khảo sát giá trị tại các thuộc tính này trên nhóm khách hàng đã được phân loại theo tiêu chí trên, ta có bảng dữ liệu sau khi đã rời rạc các thuộc tính như sau:

Bảng 1.5: Bảng sau khi rời rạc các thuộc tính của khách hàng

STT	Tuổi	Giới tính	Thu nhập	TT Hôn nhân	Là KH VIP
1	2	1	3	0	1
2	1	1	3	0	0
3	2	1	3	1	0
4	3	1	1	1	1
5	2	0	3	1	0
6	2	1	3	1	1
7	2	1	1	1	0
8	1	1	2	1	0
9	2	1	3	0	1
10	3	1	2	1	1
11	2	0	3	1	0
12	3	0	1	1	1
13	2	1	3	0	1
14	3	1	2	1	0
15	3	0	2	1	0
16	3	0	3	1	0
17	1	1	3	0	0
18	1	0	3	0	0
19	1	1	2	1	1
20	3	0	2	1	0

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,
- *TT HN:* Bằng 0 nếu chưa lập gia đình, bằng 1 nếu ngược lại.
- *Là KH VIP:* Bằng 0 nếu không thuộc loại khách hàng VIP, bằng 1 nếu ngược lại.

Khi đó, các đặc trưng chung mà doanh nghiệp cần tìm chính là một sự *phân lớp* hay *gom cụm có định hướng* (trên bảng dữ liệu đã có ở trên) mà các kết quả có thể được biểu diễn ở dạng luật kết hợp E(X)→E(Y).

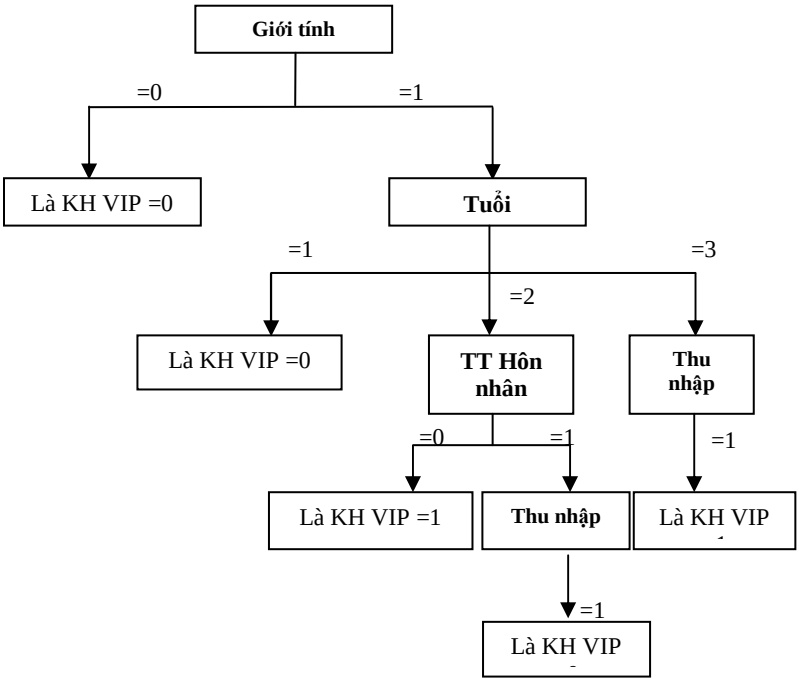
Trong đó: Y chính là thuộc tính "Là KH VIP" và E(Y) là điều kiện "Y=1" (hoặc thậm chí là Y =0), nghĩa là mọi dòng t trong bảng trên được gọi là thỏa E(Y) khi giá trị tại cột Y là 1, X là tập (hoặc tập con của) các thuộc tính còn lại (Tuổi, Giới tính, Thu nhập, TT Hôn nhân), và E(X) là một điều kiện mô tả đặc trưng chung trên X. Đương nhiên rằng luật kết hợp được chọn phải có độ phổ biến, độ tin cậy và độ quan trọng đủ tốt.

Áp dụng thuật toán cho bảng dữ liệu ở trên *(mục bảng 1.5)*, với ngưỡng độ tin cậy cho trước w là 0.7

**** Kết quả tập luật ta thu được ở ví dụ trên là :***

- Luật 1. (Giới tính =0) → (là KH VIP =0)
- Luật 2. (Giới tính = 1,Tuổi = 1) → (Là KH VIP = 0)
- Luật 3. (Giới tính = 1,Tuổi = 2,TT Hôn nhân= 0) → (Là KHVIP = 1)
- Luật 4. (Giới tính = 1,Tuổi = 2,TT Hôn nhân = 1, Thu nhập = 1) → (Là KH VIP = 0)
- Luật 5. (Giới tính = 1,Tuổi = 3,Thu nhập = 1) → (Là KH VIP = 1).

Các luật 1, 2, ..., 5 tìm được từ ví dụ trên có thể được biểu diễn lại ở dạng cây quyết định như sau:



Hình 1.3 : Sơ đồ cây quyết định

CHƯƠNG 2
ỨNG DỤNG KHAI PHÁ LUẬT KẾT HỢP TRONG CÔNG TÁC QUẢN LÝ KHO HÀNG TẠI SIÊU THỊ

2.1. Phát biểu vấn đề

Đề tài nghiên cứu lý thuyết khai phá dữ liệu, tìm hiểu về luật kết hợp và áp dụng thuật toán cây quyết định để khai phá trên cơ sở dữ liệu quản lý kho hàng tại siêu thị đã có với mong muốn tìm ra những kết quả khai phá thú vị, hữu ích nhằm giúp cho nhà quản lý có cái nhìn tổng quan hơn, nắm bắt được những mã loại hàng nào mang lại lợi nhuận cho doanh nghiệp mình. Những kết quả đạt được trong phạm vi của luận văn có thể chưa có ý nghĩa thiết thực vào công việc quản lý nhưng nó cũng góp một phần nhỏ hỗ trợ giúp cho người quản lý đưa ra được những nhận định đúng đắn hơn, mang tính chiến lược hơn.

Bài toán cụ thể đặt ra ở đây là : Xây dựng Hệ hỗ trợ ra quyết định dựa trên mã các loại hàng để đưa ra những đánh giá, những nhận định về việc doanh thu của những mã loại hàng đó có ảnh hưởng như thế nào đến lợi nhuận của doanh nghiệp.

2.2. Cơ sở dữ liệu quản lý kho hàng siêu thị

- **Xác định các thực thể :**
  - Thực thể Khách hàng : ***dbo.Khachhang***
  - Thực thể Hóa đơn : ***dbo.Hoadon***
  - Thực thể Hàng hóa : ***dbo.Hanghoa***
  - Thực thể Loại hàng : ***dbo.Loaihang***
  - Thực thể Chi tiết hóa đơn : ***dbo.Chitiethoadon***
- **Sơ đồ quan hệ các thực thể :**

- **Bảng mô tả chi tiết các ràng buộc toàn vẹn dữ liệu của các thực thể**

và dữ liệu mẫu cho các thực thể:

- ***Sơ đồ quan hệ giữa các thực thể***

2.3. Rời rạc các thuộc tính

Bảng doanh thu trước khi rời rạc các thuộc tính của 5 mã loại hàng đã chọn (loại hàng 1, loại hàng 2, loại hàng 3, loại hàng 4, loại hàng 5) và lợi nhuận thu được tương ứng. Trong bảng này ta có 347 giao dịch (dựa trên bảng chi tiết hóa đơn), mỗi giao dịch có 6 thuộc tính.

Bảng 2.6 : Bảng doanh thu trước khi rời rạc

Loaihang1	Loaihang2	Loaihang3	Loaihang4	Loaihang5	LoiNhuan
0	5.6	0	4.86	23.22	2923.74000000..
96.39	51.71	92.57	200.470000000...	15.2999999999...	8951.44
143.9	50.96	65.93	42.1499999999...	78.8000000000...	6526.77999999..
120.19	118.839999999...	74.56	156.22	0	8968.36000000..
167.510000000...	91.98	63.05	176.740000000...	15.2999999999...	11826.0600000..
182.19	27.9900000000...	33.67	87.78	32.4000000000...	5764.27999999..
205.510000000...	149.37	134.57	97.52	94.32	8370.15999999..
152.969999999...	71.1	99.79	230.630000000...	11.16	7994.49999999..
588.86	203.7	135.05	346.27	64.1400000000...	19446.8400000..

Từ bảng doanh thu ở trên, ta tiến hành 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 được xuất ra file Excel tại Sheet1 như sau:

Bảng 2.7 : Bảng kết quả sau khi đã rời rạc các thuộc tính

	A	B	C	D	E	F
1	Loại hàng 1	Loại hàng 2	Loại hàng 3	Loại hàng 4	Loại hàng 5	Lợi nhuận
2	0	1	0	1	1	2
3	1	1	1	2	1	2
4	1	1	1	1	2	1
5	1	2	1	2	0	2
6	1	2	1	2	1	2
7	2	1	1	1	1	2
8	2	2	2	1	2	1
9	1	2	2	2	1	2
10	2	2	2	2	2	2
11	1	1	2	2	2	1
12	2	2	2	2	2	1
13	1	1	1	1	1	2
14	1	1	1	1	0	2
15	2	1	2	2	2	2
16	2	1	2	2	2	1
17	2	2	2	2	2	1
18	1	2	2	2	0	2
19	1	2	1	1	2	1
20	1	1	1	1	1	1
21	1	2	2	1	1	2
22	2	1	1	2	0	1
23	1	1	1	1	1	1
24	2	2	2	2	2	1
25	1	1	1	1	1	1
26	1	1	1	1	2	1
27	1	1	1	1	0	1

Sheet1 Sheet2

2.4. Chương trình Demo minh họa

2.5. Kết quả thử nghiệm và nhận xét đánh giá

- ***Kết quả thử nghiệm:***

Kết quả khai thác luật kết hợp bằng phương pháp phân lớp với cây quyết định trên bảng doanh thu gồm 347 giao dịch, mỗi giao dịch gồm 6 thuộc tính.

Kết quả thử nghiệm đạt được cho 5 mã loại hàng lần lượt là: 1, 2, 3, 4, 5

Bảng 2.8 : Bảng kết quả thử nghiệm

STT	Ngưỡng tin cậy cho trước	Số giao dịch	Số luật thu được
1	0.6	347	12
2	0.7	347	47
3	0.8	347	59
4	0.9	347	67

- ***Nhận xét và đánh giá kết quả :***

- Từ bảng kết quả thử nghiệm ở trên ta nhận thấy rằng trong cùng một số lượng giao dịch như nhau thì giá trị của ngưỡng tin cậy sẽ tỷ lệ thuận với số luật thu được, nghĩa là khi giá trị của ngưỡng tin cậy thấp thì số luật thu được cũng sẽ ít, còn khi giá trị của ngưỡng tin cậy tăng lên thì số luật thu được cũng tăng theo.

- Thông thường người ta thường chọn ra những luật có độ tin cậy đủ tốt (độ tin cậy cao) để đánh giá, còn những luật có độ tin cậy thấp có thể chỉ để tham khảo hoặc có thể bỏ qua.

KẾT LUẬN

a) Đánh giá kết quả

1. Kết quả đạt được

✡ *Về mặt lý thuyết:*

- Nắm được kiến thức về khám phá tri thức và khai phá dữ liệu.
- Nắm được các thuật toán tìm luật kết hợp như: Apriori, Apriori-*TID*, 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.
- Hiểu rõ hơn về lập trình trên C#, và truy vấn dữ liệu trên SQL

✡ *Về mặt ứng dụng:*

- Xây dựng được hệ hỗ trợ ra quyết định phục vụ cho công tác quản lý.

2. Những hạn chế

- Chỉ mới minh họa hệ thống trên cơ sở dữ liệu của siêu thị Walmart, chưa minh họa trên nhiều cơ sở dữ liệu khác.
- Hệ thống còn đơn giản, chưa có nhiều chức năng.

b) Hướng phát triển

- Tiếp tục hoàn thiện đề tài, xây dựng hệ thống nhiều chức năng hơn, thử nghiệm và đánh giá kỹ hơn các thuật toán trên dữ liệu lớn.
- Đưa thêm các phương pháp khác của khai phá dữ liệu vào việc phân tích mô hình, như gom cụm để phân lớp dữ liệu từ đó có thể phân tích dữ liệu chính xác hơn đưa ra nhưng luật có xác suất lớn 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, ...