HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

Nguyễn Dương Hùng

ỨNG DỤNG CÂY QUYẾT ĐỊNH ĐỂ PHÂN LOẠI KHÁCH HÀNG VAY VỐN CỦA NGÂN HÀNG THƯƠNG MẠI

Chuyên ngành: Hệ thống thông tin

Mã số: 60.48.01.04

TÓM TẮT LUẬN VĂN THẠC SĨ

HÀ NỘI – NĂM 2013

Luận văn được hoàn thành tại:

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

Người hướng dẫn khoa học: PGS. TS. TRẦN ĐÌNH QUẾ

Phản biện 1: ……

Phản biện 2: ……..

Luận văn sẽ được bảo vệ trước Hội đồng chấm luận văn thạc sĩ tại Học viện Công nghệ Bưu chính Viễn thông

Vào lúc: ..... giờ ..... ngày ..... tháng .... năm .....

Có thể tìm hiểu luận văn tại:

- Thư viện của Học viện Công nghệ Bưu chính Viễn thông

MỞ ĐẦU

Lý do chọn đề tài

Ngành công nghiệp ngân hàng trên thế giới đã trải qua một sự thay đổi to lớn trong cách thức kinh doanh được thực hiện. Ngành ngân hàng đã bắt đầu nhận ra sự cần thiết của các kỹ thuật như khai phá dữ liệu, các kỹ thuật đó có thể giúp họ cạnh tranh trên thị trường. Các ngân hàng hàng đầu đã và đang sử dụng các công cụ khai phá dữ liệu (DM: Data Mining) cho việc phân khúc khách hàng và lợi nhuận,chấm điểm tín dụng và phê duyệt, quảng bá và bán sản phẩm, phát hiện các giao dịch gian lận, vv…

Có nhiều phương pháp phân lớp được đề xuất, tuy nhiên không có phương pháp tiếp cận phân loại nào là tối ưu và chính xác hơn hẳn những phương pháp khác. Dù sao với mỗi phương pháp có một lợi thế và bất lợi riêng khi sử dụng. Một trong những công cụ khai phá tri thức hiệu quả hiện nay là sử dụng cây quyết định để tìm ra các luật phân lớp. Với mong muốn nghiên cứu về việc ứng dụng cây quyết định để phân loại khách hàng của Ngân hàng thương mại, tôi đã chọn đề tài "***Ứng dụng cây quyết định để phân loại khách hàng vay vốn của Ngân hàng thương mại***" làm luận văn tốt nghiệp.

Mục tiêu nghiên cứu

Nghiên cứu các vấn đề cơ bản của thuật toán xây dựng cây quyết định ID3, cài đặt và đánh giá thuật toán đó; bước đầu áp dụng mô hình cây quyết định (ID3: Decision Tree) đã xây dựng vào việc phân loại khách hàng vay vốn của Ngân hàng thương mại.

Đối tượng, phạm vi nghiên cứu

- Tìm hiểu thuật toán khai phá dữ liệu ID3 để phân loại khách hàng dựa trên dữ liệu ngân hàng đã có.
- Cài đặt và thử nghiệm với dữ liệu là các tập tin Excel.

Phương pháp nghiên cứu

- Phương pháp nghiên cứu tài liệu: Phân tích và tổng hợp các tài liệu về khai phá dữ liệu sử dụng thuật toán về Decision Tree có thuật toán ID3, phân loại dữ liệu, mô hình dự báo.
- Phương pháp thực nghiệm: Ứng dụng kết hợp kỹ thuật phân loại và mô hình cây quyết định để phân loại khách hàng vay vốn của Ngân hàng thương mại.

Bố cục luận văn:

Chương 1 Tổng quan về khai phá dữ liệu

1.1 Giới thiệu về khai phá dữ liệu

1.2 Một số phương pháp khai phá dữ liệu hiện đại

1.3 Một số phương pháp khai phá dữ liệu thông dụng

1.4 Ứng dụng khai phá dữ liệu trong lĩnh vực khách hàng.

Chương 2 Ứng dụng cây quyết định trong quy trình tín dụng

2.1 Quy trình tín dụng

2.2 Sử dụng cây quyết định để phân loại khách hàng

2.3 Thuật toán xây dựng cây quyết định dựa vào Entropy

Chương 3 Xây dựng chương trình thử nghiệm và đánh giá

3.1 Giới thiệu bài toán

3.2 Cơ sở dữ liệu

3.3 Cài đặt ứng dụng

3.5 Kết luận

Chương 1 - TỔNG QUAN VỀ KHAI PHÁ DỮ LIỆU

1.1 Giới thiệu về khai phá dữ liệu

1.1.1 Khám phá tri thức

Quá trình khám phá dữ liệu gồm các bước cơ bản sau đây [1]:

Bước 1: ***Xác định vấn đề và lựa chọn nguồn dữ liệu*** *(Problem Understanding anh Data Understanding)*

Bước 2: ***Chuẩn bị dữ liệu*** **(Data preparation)**

Quá trình này gồm các quá trình sau:

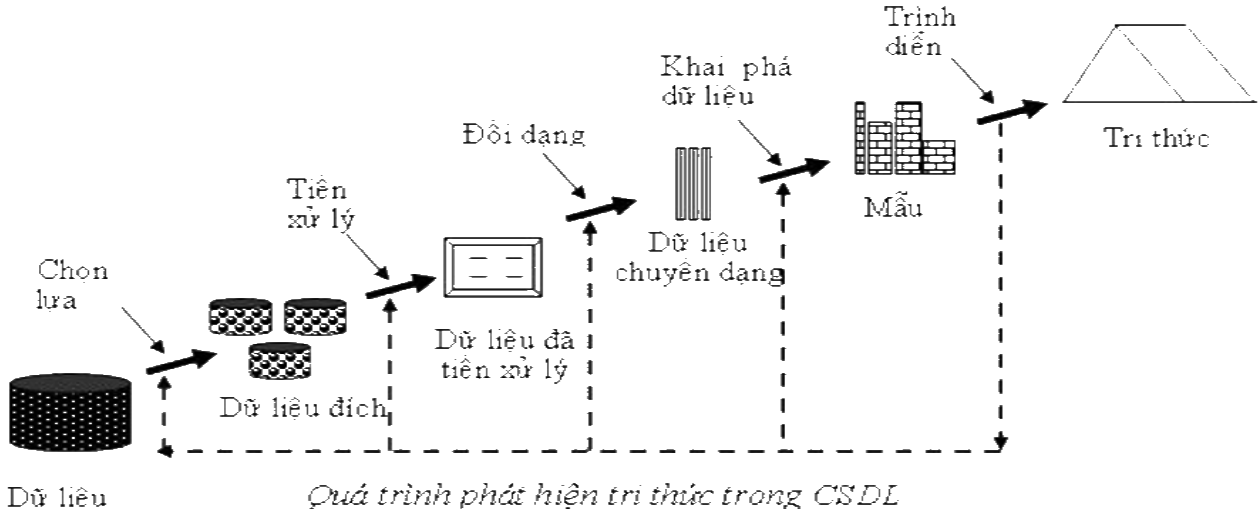
- Thu thập dữ liệu (Data gathering)
- Làm sạch dữ liệu (Data cleaning)
- Tích hợp dữ liệu (Data integeration)
- Chọn dữ liệu (Data selection)
- Biến đổi dữ liệu (Data transformation)

Bước 3: ***Khai phá dữ liệu*** *(Data Mining)*

Bước 4: ***Đánh giá mẫu*** *(Partern Evalution)*

Bước 5: ***Biểu diễn tri thức và triển khai (****Knowlegde presentation and Deployment)*

Tóm lại: KDD 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 [2], [5].



Hình 1: Quá trình phát hiện tri thức trong CSDL

1.1.2 Khai phá dữ liệu

Khai phá dữ liệu được dùng để mô tả quá trình phát hiện ra tri thức trong CSDL. Quá trình khai phá dữ liệu bao gồm các giai đoạn [2]:

Giai đoạn 1: ***Gom dữ liệu***

Giai đoạn 2: ***Trích lọc dữ liệu***

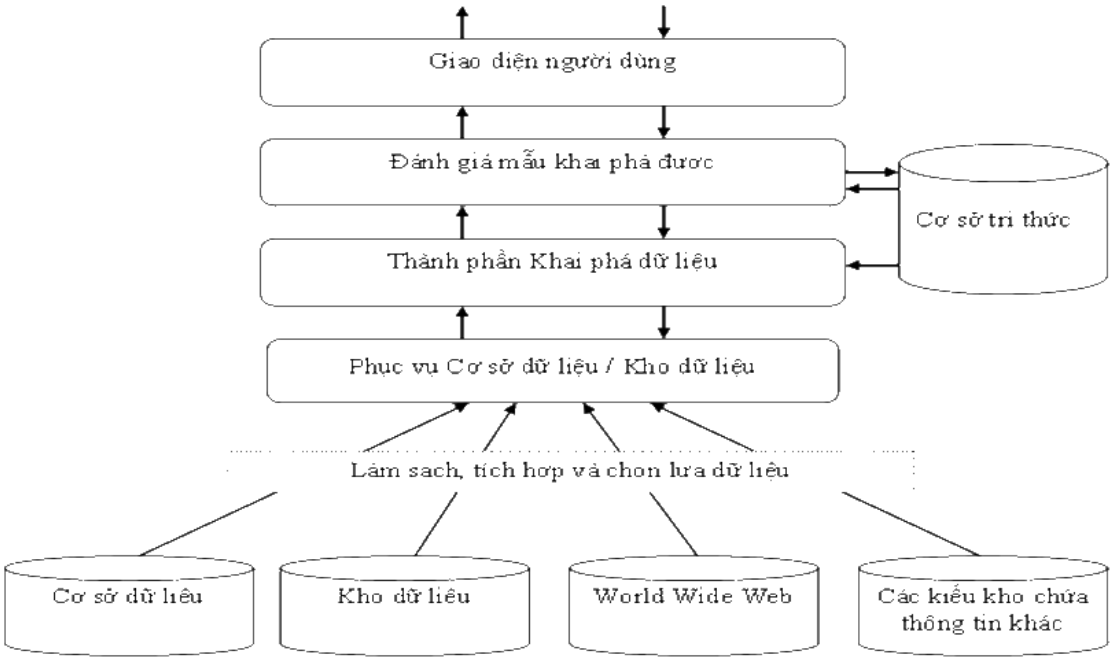
Giai đoạn 3: ***Làm sạch, tiền xử lý và chuẩn bị trước dữ***

Giai đoạn 4: ***Chuyển đổi dữ liệu***

Giai đoạn 5: ***Phát hiện và trích mẫu dữ***

Giai đoạn 6: ***Đánh giá kết quả mẫu***

Quá khai phá dữ liệu được mô hình hóa một cách tổng quát như hình vẽ dưới đây [2]:



Hình 2: Kiến trúc điển hình của hệ thống khai phá dữ liệu

1.2 Ứng dụng của khai phá dữ liệu

Hiện nay, các kĩ thuật khai phá dữ liệu đang được áp dụng một cách rộng rãi trong rất nhiều lĩnh vực kinh doanh và đời sống khác nhau như marketing, tài chính, ngân hàng và bảo hiểm, khoa học, giáo dục, y tế, an ninh, internet:

1.3 Một số phương pháp khai phá dữ liệu thông dụng

1.3.1 Phân lớp (Classification)

Quá trình phân lớp dữ liệu thường gồm 2 pha:

Pha 1: **Xây dựng mô hình**

Trong bước này, một mô hình sẽ được xây dựng dựa trên việc phân tích các mẫu dữ liệu sẵn có. Đầu vào của quá trình này là một tập dữ liệu có cấu trúc được mô tả bằng các thuộc tính và được tạo ra từ tập các bộ giá trị của các thuộc tính đó. Mỗi bộ giá trị được gọi chung là một mẫu (sample). Trong tập dữ liệu này, mỗi mẫu được giả sử thuộc về một lớp định trước, lớp ở đây là giá trị của một thuộc tính được chọn làm thuộc tính gán nhãn lớp hay thuộc tính quyết định. Đầu ra của bước này thường là các quy tắc phân lớp dưới dạng luật dạng **if-then** (nếu-thì), cây quyết định, công thức logic, hay mạng nơron.

Pha 2: Sử dụng mô hình đã xây dựng để phân lớp dữ liệu

Trong bước này việc đầu tiên là phải làm là tính độ chính xác của mô hình. Nếu độ chính xác là chấp nhận được mô hình sẽ được sử dụng để dự đoán nhãn lớp cho các mẫu dữ liệu khác trong tương lai.

1.3.2 Phân cụm (Clustering)

Phân cụm là việc mô tả chung để tìm ra các tập hay các nhóm, loại mô tả dữ liệu. Các nhóm có thể tách nhau hoặc phân cấp hay gối lên nhau. Có nghĩa là dữ liệu có thể vừa thuộc nhóm này lại vừa thuộc nhóm khác. Các ứng dụng khai phá dữ liệu có nhiệm vụ phân nhóm như phát hiện tập các khách hàng có phản ứng giống nhau trong CSDL tiếp thị; xác định các quang phổ từ các phương pháp đo tia hồng ngoại...

1.3.3 Luật kết hợp (Association Rules)

Khai phá luật kết hợp được thực hiện qua 2 bước:

- Bước 1: Tìm tất cả các tập mục phổ biến, một văn bản phổ biến được xác định qua độ hỗ trợ và thỏa mãn độ hỗ trợ cực tiểu.
- Bước 2: Sinh ra các luật kết hợp mạnh từ tập mục phổ biến, các luật phải thỏa mãn độ hỗ trợ cực tiểu và độ tin cậy cực tiểu.

1.4 Ứng dụng khai phá dữ liệu trong lĩnh vực ngân hàng

1.4.1 Marketing

Một trong những lĩnh vực được ứng dụng rộng rãi nhất cho ngành ngân hàng của kỹ thuật khai phá dữ liệu đó là lĩnh vực quảng bá sản phẩm. Bộ phận tiếp thị và bán hàng của các Ngân hàng có thể sử dụng kỹ thuật khai phá dữ liệu để phân tích cơ sở dữ liệu về khách hàng. Kỹ thuật khai thác dữ liệu cũng giúp xác định khách hàng nào sẽ mang lại lợi nhuận và khách hàng nào không mang lại lợi nhuận.

1.4.2 Quản lý rủi ro

Khai phá dữ liệu được sử dụng rộng rãi để quản lý rủi ro trong ngành công nghiệp ngân hàng [4]. Giám đốc điều hành ngân hàng cần phải biết rằng các khách hàng mà họ đang có liệu đáng tin cậy hay không.

1.4.3 Phát hiện gian lận

Một lĩnh vực khác trong khai phá dữ liệu có thể được sử dụng trong ngành công nghiệp ngân hàng là việc phát hiện gian lận. Phát hiện các hành động gian lận là một mối quan tâm ngày càng tăng cho nhiều doanh nghiệp, và với sự giúp đỡ của kỹ thuật khai phá dữ liệu các hành động gian lận ngày càng được phát hiên nhiều hơn.

1.4.4 Quản trị quan hệ khách hàng

Trong thời đại cạnh tranh khốc liệt ngày nay nói chung, đặc biệt là trong ngành ngân hàng, khách hàng luôn luôn là nhân tố quan trọng nhất quyết định sự tồn tại và phát triển của họ. Khai phá dữ liệu rất hữu ích trong tất cả ba giai đoạn trong một chu kỳ mối quan hệ khách hàng: Tìm kiếm khách hàng, tăng giá trị của khách hàng và duy trì khách hàng.

1.5 Kết luận

Trong chương này, luận văn đã giới thiệu tổng quan về khai phá dữ liệu, ứng dụng của khai phá dữ liệu, một số phương pháp khai phá dữ liệu thông dụng. Trong chương sau, luận văn sẽ trình bày nội dung lý thuyết và ứng dụng của thuật toán khai phá dữ liệu thông dụng : Thuật toán cây quyết định ID3. Đó là một thuật toán được ứng dụng để khai phá dữ liệu trong các lĩnh vực khác nhau, đặc biệt trong lĩnh vực ngân hàng.

Chương 2 - ỨNG DỤNG CÂY QUYẾT ĐỊNH TRONG QUY TRÌNH TÍN DỤNG

2.1 Quy trình tín dụng

2.1.1 Khái niệm quy trình tín dụng

Để chuẩn hoá quá trình tiếp xúc, phân tích, cho vay và thu nợ đối với khách hàng, các Ngân hàng thường đặt ra quy trình phân tích tín dụng [4]. Đó chính là các bước (hoặc nội dung công việc) mà cán bộ tín dụng, các phòng ban có liên quan trong Ngân hàng phải thực hiện khi làm việc cho khách hàng.

2.1.2 Ý nghĩa của quy trình tín dụng

Việc thiết lập một quy trình tín dụng và không ngừng hoàn thiện nó đặc biệt quan trọng đối với một ngân hàng thương mại. Về mặt hiệu quả, một quy trình tín dụng hợp lý sẽ giúp cho ngân hàng nâng cao chất lượng tín dụng và giảm thiểu rủi ro tín dụng.

2.1.3 Quy trình tín dụng căn bản

Bước 1: Lập hồ sơ vay vốn

Bước này do cán bộ tín dụng thực hiện ngay sau khi tiếp xúc khách hàng. Nhìn chung một bộ hồ sơ vay vốn cần phải thu thập các thông tin như:

- Năng lực pháp lý, năng lực hành vi dân sự của khách hàng
- Khả năng sử dụng vốn vay
- Khả năng hoàn trả nợ vay (vốn vay và lãi)

Bước 2: Phân tích tín dụng

Phân tích tín dụng là xác định khả năng hiện tại và tương lai của khách hàng trong việc sử dụng vốn vay và hoàn trả nợ vay với mục tiêu:

- Tìm kiếm những tình huống có thể xảy ra dẫn đến rủi ro cho ngân hàng, dự đoán khả năng khắc phục những rủi ro đó, dự kiến những biện pháp giảm thiểu rủi ro và hạn chế tổn thất cho ngân hàng.
- Phân tích tính chân thật của những thông tin đã thu thập được từ phía khách hàng trong bước 1, từ đó nhận xét thái độ, thiện chí của khách hàng làm cơ sở cho việc ra quyết định cho vay.

Bước 3: Ra quyết định tín dụng

Trong khâu này, ngân hàng sẽ ra quyết định đồng ý hoặc từ chối cho vay đối với một hồ sơ vay vốn của khách hàng.

Bước 4: Giải ngân

Nguyên tắc giải ngân: phải gắn liền sự vận động tiền tệ với sự vận động hàng hóa hoặc dịch vụ có liên quan, nhằm kiểm tra mục đích sử dụng vốn vay của khách hàng và đảm bảo khả năng thu nợ.

Bước 5: Giám sát tín dụng

Nhân viên tín dụng thường xuyên kiểm tra việc sử dụng vốn vay thực tế của khách hàng, hiện trạng tài sản đảm bảo, tình hình tài chính của khách hàng... để đảm bảo khả năng thu nợ.

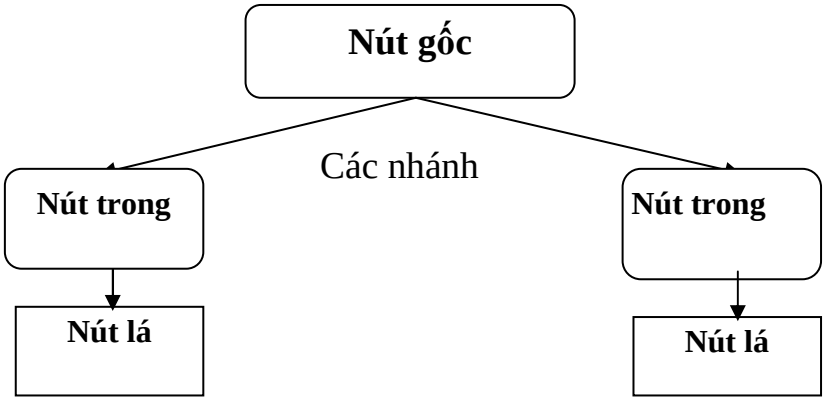
Bước 6: Thanh lý hợp đồng tín dụng

2.2 Sử dụng cây quyết định (DT) để phân loại khách hàng

2.2.1 Tổng quan về thuật toán cây quyết định

Chúng ta có thể định nghĩa cây quyết định có các tính chất sau:

- Mỗi nút trong (internal node) biểu diễn một thuộc tính cần kiểm tra giá trị (an attribute to be tested) đối với các các tập thuộc tính.
- Nút lá (leaf node) hay còn gọi là nút trả lời biểu thị cho một lớp các trường hợp mà nhãn của nó là tên của lớp, nó biểu diễn một lớp (a classififcation)
- Nút nhánh (branch) từ một nút sẽ tương ứng với một giá trị có thể của thuộc tính gắn với nút đó.
- Nhãn (lable) của nút này là tên của thuộc tính và có một nhánh nối nút này đến các cây con ứng với mỗi kết quả có thể có phép thử. Nhãn của nhánh này là các giá trị của thuộc tính đó. Nút trên cùng gọi là nút gốc.



Hình 3: Mô tả chung về cây quyết định

2.2.2 Thiết kế cây quyết định

2.2.2.1 Xử lý dữ liệu

Công việc cụ thể của bước tiền xử lý dữ liệu gồm các công việc:

- ***Lọc thuộc tính*** *(Filtering Attributes)*
- ***Lọc các mẫu*** *(Filtering samples)*
- Lọc các mẫu (instances, patterns)
- ***Chuyển đổi dữ liệu*** (*Transformation)*
- ***Rời rạc hóa dữ liệu*** *(Discretization)*

2.2.2.2 Tạo cây

Cây quyết định được tạo thành bằng cách lần lượt chia (theo phương pháp đệ quy) một tập dữ liệu thành các tập dữ liệu con, mỗi tập con được tạo thành từ các phần tử của cùng một lớp. Các nút (không phải là nút lá) là các điểm phân nhánh của cây. Việc phân nhánh tại các nút có thể dựa trên việc kiểm tra một hay nhiều thuộc tính để xác định việc phân chia dữ liệu.

2.2.2.3 Tiêu chuẩn tách

Chúng ta mong muốn chọn thuộc tính sao cho việc phân lớp tập mẫu là tốt nhất. Như vậy chúng ta cần phải có một tiêu chuẩn để đánh giá vấn đề này. Có rất nhiều tiêu chuẩn được đánh giá được sử dụng đó là: Lượng thông tin thu thêm IG (Information Gain), thuật toán ID3 của John Ross Quilan.

2.2.2.4 Tiêu chuẩn dừng

Chúng ta tập trung một số tiêu chuẩn dừng chung nhất được sử dụng trong cây quyết định. Tiêu chuẩn dừng truyền thống sử dụng các tập kiểm tra. Chúng ta có thể thay ngưỡng như là giảm nhiễu, số các mẫu trong một nút, tỉ lệ các mẫu trong nút, hay chiều sâu của cây.

2.2.2.5 Tỉa cây

Sau giai đoạn tạo cây chúng ta có thể dùng phương pháp "Độ dài mô tả ngắn nhất" (Minimum Description Length) hay giá trị tối thiểu của IG để tỉa cây (chúng ta có thể chọn giá trị tối thiểu của IG trong giai đoạn tạo cây đủ nhỏ để cho cây phát triển tương đối sâu, sau đó lại nâng giá trị này lên để tỉa cây).

2.2.3 Các bước tổng quát để xây dựng cây quyết định

Quá trình xây dựng một cây quyết định cụ thể bắt đầu bằng một nút rỗng bao gồm toàn bộ các đối tượng huấn luyện và làm như sau :

1. Nếu tại nút hiện thời, tất cả các đối tượng huấn luyện đều thuộc vào một lớp nào đó thì nút này chính là nút lá có tên là nhãn lớp chung của các đối tượng.
2. Trường hợp ngược lại, sử dụng một độ đo, chọn thuộc tính điều kiện phân chia tốt nhất tập mẫu huấn luyện có tại nút.
3. Tạo một lượng nút con của nút hiện thời bằng số các giá trị khác nhau của thuộc tính được chọn. Gán cho mỗi nhánh từ nút cha đến nút con một giá trị của thuộc tính rồi phân chia các các đối tượng huấn luyện vào các nút con tương ứng.
4. Nút con K được gọi là thuần nhất, trở thành lá, nếu tất cả các đối tượng mẫu tại đó đều thuộc vào cùng một lớp.
5. Lặp lại các bước 1 - 3 đối với mỗi nút chưa thuần nhất.

2.2.4 Nghiên cứu cây quyết định trong khai phá dữ liệu

2.2.4.1 Xác định lớp của các mẫu mới

Trên cơ sở đã biết giá trị của các thuộc tính của các mẫu X_1, X_2, …, X_n ta xác định thuộc tính quyết định (hay phân lớp) Y của đối tượng đó (có thể dùng kỹ thuật này để nhận dạng mẫu, dự báo …)

2.2.4.2 Rút ra các tri thức hay luật từ cây

Với mục đích và nhiệm vụ chính của việc khai phá dữ liệu là phát hiện ra các quy luật, các mô hình từ trong CSDL. Từ mô hình thu được ta rút ra các tri thức hay các quy luật

dưới dạng cây hoặc các luật dưới dạng "**If … Then**…". Hai mô hình trên là tương đương, chúng có thể được chuyển đổi qua lại giữa các mô hình đó với nhau.

Ví dụ :

Các luật rút ra từ cây trong ví dụ trên:

+ Luật 1: IF(**Nhiệt độ**: cao) AND (**Ngoài trời**: mưa) THEN (=> Quyết định: Không)

+ Luật 2: IF(**Độ ẩm**: cao) AND (**Ngoài trời**: nắng) THEN (=> Quyết định:Không)

+ Luật3: IF(**Độ ẩm**:Cao) AND (**Ngoài trời**:Bình thường)THEN (=> Quyết định: Có)

Sau đó, ta sử dụng các luật này để hỗ trợ quá trình ra các quyết định, dự đoán.

2.3 Thuật toán xây dựng cây quyết định dựa vào Entropy

2.3.1 Tiêu chí chọn thuộc tính phân lớp

Tiêu chí để đánh giá tìm điểm chia là rất quan trọng, chúng được xem là một tiêu chuẩn "heuristic" để phân chia dữ liệu. Ý tưởng chính trong việc đưa ra các tiêu chí trên là làm sao cho các tập con được phân chia càng trở nên "trong suốt" (tất cả các bộ thuộc về cùng một nhãn) càng tốt. Thuật toán dùng độ đo lượng thông tin thu thêm (Information Gain - IG) để xác định điểm chia [2]. Độ đo này dựa trên cơ sở lý thuyết thông tin của nhà toán học Claude Shannon, độ đo này được xác như sau:

Xét bảng quyết định DT = (U, C ∪ {d}), số giá trị (nhãn lớp) có thể của d là k. Khi đó Entropy của tập các đối tượng trong DT được định nghĩa bởi:

$$Entropy(U) = -\sum_{i=1}^{k} p_i \log_2 p_i$$

trong đó p_i là tỉ lệ các đối tượng trong DT mang nhãn lớp i. Ý nghĩa của đại lượng Entropy trong lĩnh vực lý thuyết công nghệ thông tin: Entropy của tập U chỉ ra sô lượng bít cần thiết để mã hóa lớp của một phần tử được lấy ra ngẫu nhiên từ tập U. Lượng thông tin thu thêm (***Information Gain*** - IG) là lượng Entropy còn lại khi tập các đối tượng trong DT được phân hoạch theo một thuộc tính điều kiện c nào đó. IG xác định theo công thức sau [6]:

$$IG(U,c) = Entropy(U) - \sum_{v \in V_c} \frac{|U_v|}{|U|} Entropy(U_v)$$

trong đó V_c là tập các giá trị của thuộc tính c, U_v là tập các đối tượng trong DT có giá trị thuộc tính c bằng v. Giá trị IG(U, c) được sử dụng làm độ đo lựa chọn thuộc tính phân chia dữ liệu tại mỗi nút trong thuật toán xây dựng cây quyết định ID3. Thuộc tính

được chọn là thuộc tính cho lượng thông tin thu thêm lớn nhất. Ý nghĩa của đại lượng IG trong lĩnh vực lý thuyết công nghệ thông tin: IG của tập S chỉ ra số lượng bít giảm đối với việc mã hóa lớp của một phần tử c được lấy ra ngẫu nhiên từ tập U.

2.3.2 Thuật toán ID3

Ý tưởng của thuật toán ID3:

- Thực hiện giải thuật tìm kiếm thám lam (greedy search) đối với không gian các cây quyết định có thể.
- Xây dựng nút (node) theo chiến lược Top-Down, bắt đầu từ nút gốc.
- Ở mỗi nút, thuôc tính kiểm tra (test attribute) là thuộc tính có khả năng phân loại tốt nhất.
- Tạo mới một cây con (sub-tree) của nút hiện tại cho mỗi giá trị có thể của thuộc tính kiểm tra, và tập dữ liệu đầu vào sẽ được tách ra thành các tập con tương ứng với các cây con vừa tạo.
- Mỗi thuộc tính chỉ được phép xuất hiện tối đa 1 lần đối với bất kỳ đường đi nào trong cây.
- Quá trình phát triển cây sẽ tiếp tục cho tới khi:
  - Cây quyết định phân loại hoàn toàn (perfectly classifies) các dữ liệu đầu vào.
  - Tất cả các thuộc tính được sử dụng.

Giả mã của thuật toán ID3 như sau:

```
Dữ liệu vào: Bảng quyết định DT = (U, C ∪ {d})
Dữ liệu ra: Mô hình cây quyết định
Function Create_tree (U, C, {d})
Begin
If tất cả các mẫu thuộc cùng nhãn lớp di then
return một nút lá được gán nhãn di
else     if  C = null  then
return nút lá có nhãn dj là lớp phổ biến nhất trong DT
else
    begin
            bestAttribute:= getBestAttribute(U, C);
            // Chọn thuộc tính tốt nhất để chia
```

```
            C := C- {bestAttribute};
        //xóa bestAttribute khỏi tập thuộc tính
        với mỗi v in bestAttribute
        Begin
            Uv := [U]v ;
            //Uv là phân hoạch của U
            ChildNode:=Create_tree(Uv, C, {d});
            //Tạo 1 nút con
            end
        end
End
```

Giả mã của hàm getBestAttribute như sau:

```
Dữ liệu vào: Bảng quyết định DT = (U, C∪{d})
Dữ liệu ra: Thuộc tính điều kiện tốt nhất
Function getBestAttribute (U, C);
Begin
    maxIG := 0;
    Với mỗi c in C
    begin
        tg : = IG(U, c);
        // Tính lượng thông tin thu thêm IG(U,c)
        If  (tg > max IG) then
            begin
                maxIG := tg;
                kq := c;
            end
    end
    return kq;
        //Hàm trả về thuộc tính có lượng thông tin thu
    thêm IG là lớn nhất
End
```

2.3.3 Ví dụ về thuật toán ID3

Xét bảng quyết định DT = {U, C ∪ {d}} sau đây:

Bảng 1: Dữ liệu huấn luyện

Ngày	Quang cảnh	Gió	Nhiệt độ	Độ ẩm	Quyết định
Ngày 1	Âm u	Có	Mát mẻ	Cao	Có
Ngày 2	Nắng	Không	Ấm áp	Cao	Không
Ngày 3	Nắng	Không	Nóng	Cao	Không
Ngày 4	Âm u	Không	Nóng	Trung bình	Không
Ngày 5	Nắng	Có	Nóng	Thấp	Có
Ngày 6	Mưa	Không	Ấm áp	Cao	Không
Ngày 7	Mưa	Không	Nóng	Cao	Không
Ngày 8	Mưa	Không	Nóng	Trung bình	Không
Ngày 9	Âm u	Có	Nóng	Thấp	Có
Ngày 10	Mưa	Không	Ấm áp	Trung bình	Có
Ngày 11	Mưa	Có	Nóng	Trung bình	Không
Ngày 12	Mưa	Không	Nóng	Cao	Không

Thuật toán xây dựng cây quyết định với dữ liệu ở bảng trên như sau:

- ***Trước tiên nút lá được khởi tạo gồm các mẫu từ 1 đến 12***

Đầu tiên sẽ tính Entropy cho toàn bộ tập huấn luyện U gồm: bốn bộ {1, 5, 9, 10} có giá trị thuộc tính nhãn là "CÓ" và tám bộ {2, 3, 4, 6, 7, 8, 11, 12} có thuộc tính nhãn là "KHÔNG", do đó:

$$Entropy\,(U) = -\frac{4}{12}\log_2\frac{4}{12} - \frac{8}{12}\log_2\frac{8}{12} = 0.918$$

Tính IG cho từng thuộc tính:

Thuộc tính "Quang cảnh". Thuộc tính này có ba giá trị là "Âm u", "Nắng" và "Mưa". Căn cứ vào bảng dữ liệu ta thấy:

- Với giá trị của "Âm u" có ba bộ {1, 9} có giá trị thuộc tính nhãn là "CÓ" và có một bộ {4} có nhãn lớp là "KHÔNG".
- Tương tự giá trị của "Nắng" có một bộ {5} có nhãn lớp là "CÓ" và có hai bộ {2, 3} có nhãn lớp là "KHÔNG";
- Với giá trị "Mưa" có một bộ {10} có nhãn lớp "CÓ" và năm bộ {6, 7, 8, 11, 12} có nhãn lớp "KHÔNG".

Theo công thức trên, độ đo lượng thông tin thu thêm của thuộc tính “Quang cảnh” xét trên U là:

$$IG(U,\text{Outlook}) = Entropy(U) - \sum_{v \in V_{\text{Outlook}}} \frac{|U_v|}{|U|} Entropy(U_v)$$

$$= 0.918 - [\frac{3}{12}(-\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3}) + \frac{3}{12}(-\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3}) + \frac{6}{12}(-\frac{1}{6}\log_2\frac{1}{6} - \frac{5}{6}\log_2\frac{5}{6})] = 0.134$$

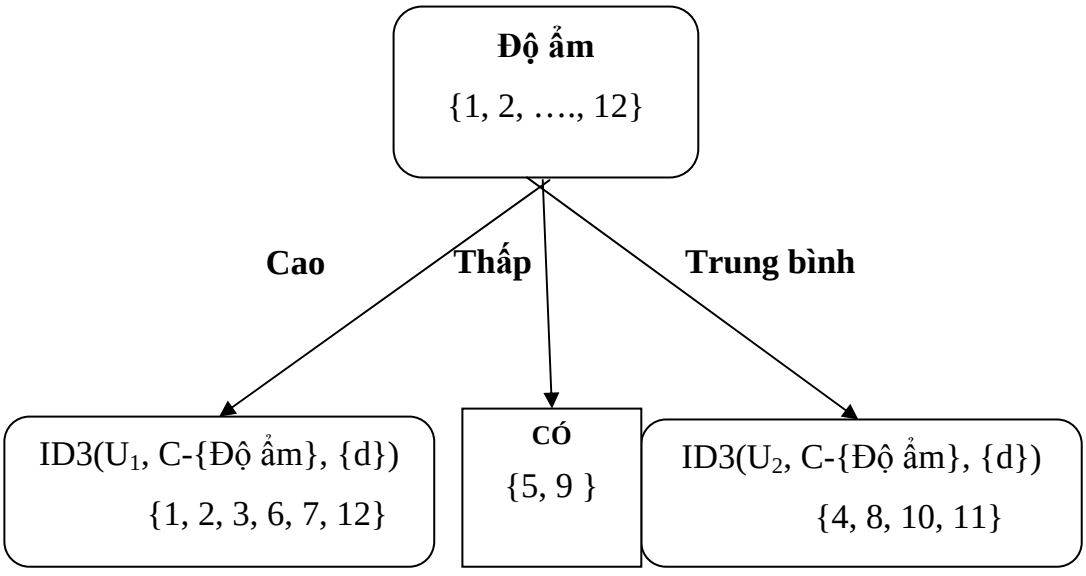
Theo cách tính tương tự như trên, ta tính được:

$$\text{IG(U,Gió)} = 0.918 - [\frac{4}{12}(-\frac{3}{4}\log_2\frac{3}{4} - \frac{1}{4}\log_2\frac{1}{4}) + \frac{8}{12}(-\frac{1}{8}\log_2\frac{1}{8} - \frac{7}{8}\log_2\frac{7}{8})] = 0.285$$

$$\text{IG(U,Nhiệt độ)} = 0.918 - [\frac{3}{12}(-\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3}) + \frac{8}{12}(-\frac{2}{8}\log_2\frac{2}{8} - \frac{6}{8}\log_2\frac{6}{8})] = 0.148$$

$$\text{IG(U,Độ ẩm)} = 0.918 - [\frac{6}{12}(-\frac{1}{6}\log_2\frac{1}{6} - \frac{5}{6}\log_2\frac{5}{6}) + \frac{4}{12}(-\frac{1}{4}\log_2\frac{1}{4} - \frac{3}{4}\log_2\frac{3}{4})] = 0.323$$

Như vậy, thuộc tính “Độ ẩm” là thuộc tính có chỉ số IG lớn nhất nên sẽ được chọn là thuộc tính phân chia. Vì thế thuộc tính “Độ ẩm” được chọn làm nhãn cho nút gốc, ba nhánh được tạo ra lần lượt với tên là: “Cao”, “Trung bình”, “Thấp”.Hơn nữa nhánh “Thấp” có các mẫu {5, 9} cùng thuộc một lớp “CÓ ” nên nút lá được tạo ra với nhãn là “CÓ ”.Kết quả phân chia sẽ là cây quyết định như sau:



Hình 4: Cây sau khi chọn thuộc tính Độ ẩm (ID3)

Bước tiếp theo gọi thuật toán đệ quy: ID3(U_1, C-{Độ ẩm}, {d})

Tương tự để tìm điểm chia tốt nhất tại thuật toán này, phải tính toán chỉ số IG của các thuộc tính "Quang cảnh", "Gió", "Nhiệt độ".

- Đầu tiên ta cũng tính Entropy cho toàn bộ tập huấn luyện trong U_1 gồm một bộ {1} có thuộc tính nhãn là "CÓ " và năm bộ {2, 3, 6, 7, 12} có thuộc tính nhãn là "KHÔNG":

$$Entropy\ (U_1) = -\frac{1}{6}\log_2\frac{1}{6} - \frac{5}{6}\log_2\frac{5}{6} = 0.65$$

- Tiếp theo tính IG cho thuộc tính "Quang cảnh", thuộc tính này có ba giá trị là "Âm u", "Nắng" và "Mưa". Nhìn vào bảng dữ liệu:
  - Với giá trị "Âm u" chỉ có một bộ {1} có giá trị thuộc tính nhãn là "CÓ ".
  - Tương tự giá trị "Nắng" chỉ có hai bộ {2, 3} đều có nhãn lớp là "KHÔNG";
  - Với giá trị "Mưa" chỉ có ba bộ {6, 7, 12} đều có nhãn lớp "KHÔNG".

  Do đó, độ đo lượng thông tin thu thêm của thuộc tính "Quang cảnh" xét trên U_1 là:

$$IG(U_1, \text{Quang cảnh}) = 0.65 - [\frac{1}{6}(-\frac{1}{1}\log_2\frac{1}{1}) + \frac{2}{6}(-\frac{2}{2}\log_2\frac{2}{2}) + \frac{3}{6}(-\frac{3}{3}\log_2\frac{3}{3})] = 0.65$$

- Tính tương tự ta cũng có:

$$IG(U_1, \text{Gió}) = 0.65 - [\frac{1}{6}(-\frac{1}{1}\log_2\frac{1}{1}) + \frac{5}{6}(-\frac{5}{5}\log_2\frac{5}{5})] = 0.65$$

$$IG(U_1, \text{Nhiệt độ}) = 0.65 - [\frac{1}{6}(-\frac{1}{1}\log_2\frac{1}{1}) + \frac{5}{6}(-\frac{5}{5}\log_2\frac{5}{5})] = 0.65$$

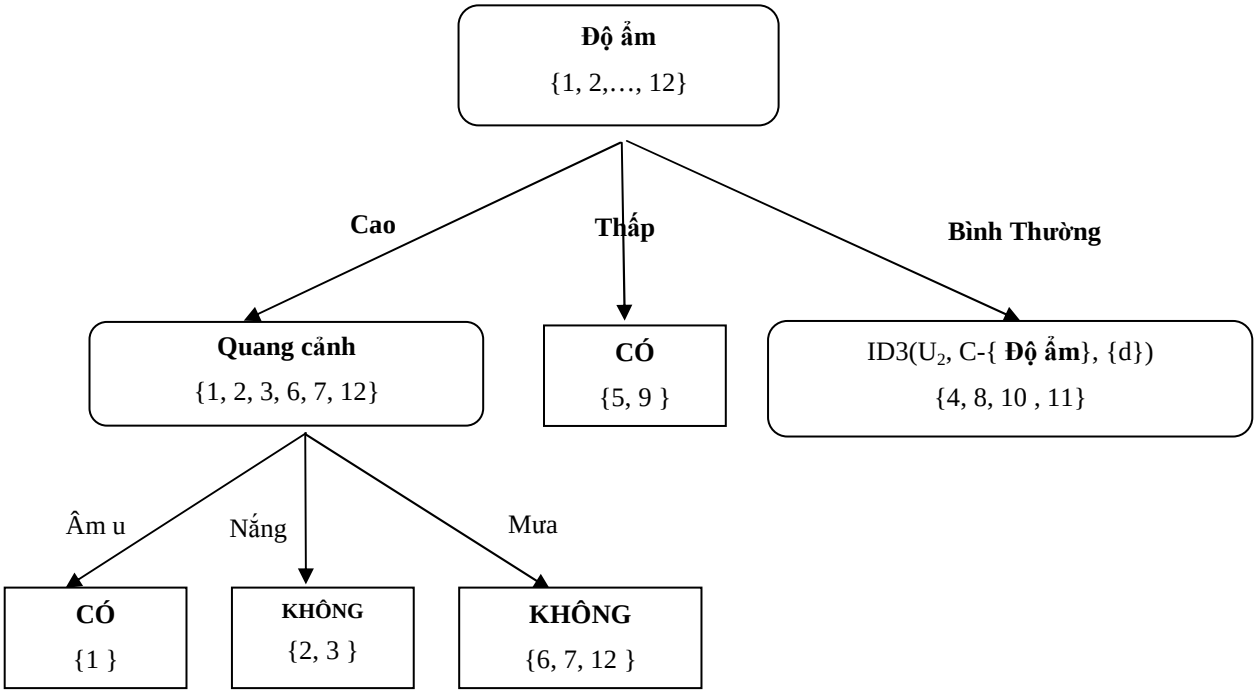
Ta thấy chỉ số IG của ba thuộc tính "Quang cảnh", "Gió", "Nhiệt độ" là như nhau, ta có thể chọn bất kỳ thuộc tính nào để phân chia.

Giả sử ta chọn thuộc tính "Quang cảnh" để phân chia. Do đó, thuộc tính "Quang cảnh" làm nhãn cho nút bên trái nối với nhánh "Cao".

Thuộc tính này có ba giá trị "Âm u", "Nắng" và "Mưa" nên ta tiếp tục tạo thành ba nhánh mới là "Âm u", "Nắng" và "Mưa":

- Với nhánh "Âm u" gồm một mẫu {1} và có giá trị quyết định là "CÓ " nên ta tạo nút lá là "CÓ ".
- Với nhánh "Nắng" gồm hai mẫu {2, 3} và có cùng giá trị quyết định là "KHÔNG" nên tạo nút lá là "KHÔNG".

- Với nhánh "Mưa" có ba mẫu {6, 7, 12} và đều có giá trị quyết định là "KHÔNG" nên ta tạo nút lá là "KHÔNG". Sau khi thực hiện xong thuật toán đệ quy: **ID3(U_1, C-{Độ ẩm}, {d})**, ta có cây như sau:



Hình 5: Cây sau khi chọn thuộc tính Quang cảnh (ID3)

- ***Bước tiếp theo gọi thuật toán đệ quy: ID3(U2, C-{ Độ ẩm}, {d})***
- Tính một cách tương tự như trên ta có:

$$\text{Entropy } (U_2) = -\frac{1}{4}\log_2\frac{1}{4} - \frac{3}{4}\log_2\frac{3}{4} = 0.811$$

$$\text{IG}(U_2, \text{Quang cảnh}) = 0.811 - [\frac{1}{4}(-\frac{1}{1}\log_2\frac{1}{1}) + \frac{3}{4}(-\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3})]$$

$$= 0.811\text{-}0.689 = 0.123$$

$$\text{IG}(U_2, \text{Gió}) = 0.811 - [\frac{3}{4}(-\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3}) + 1/4(-\frac{1}{1}\log_2\frac{1}{1})]$$

$$= 0.811\text{-}0.689 = 0.123$$

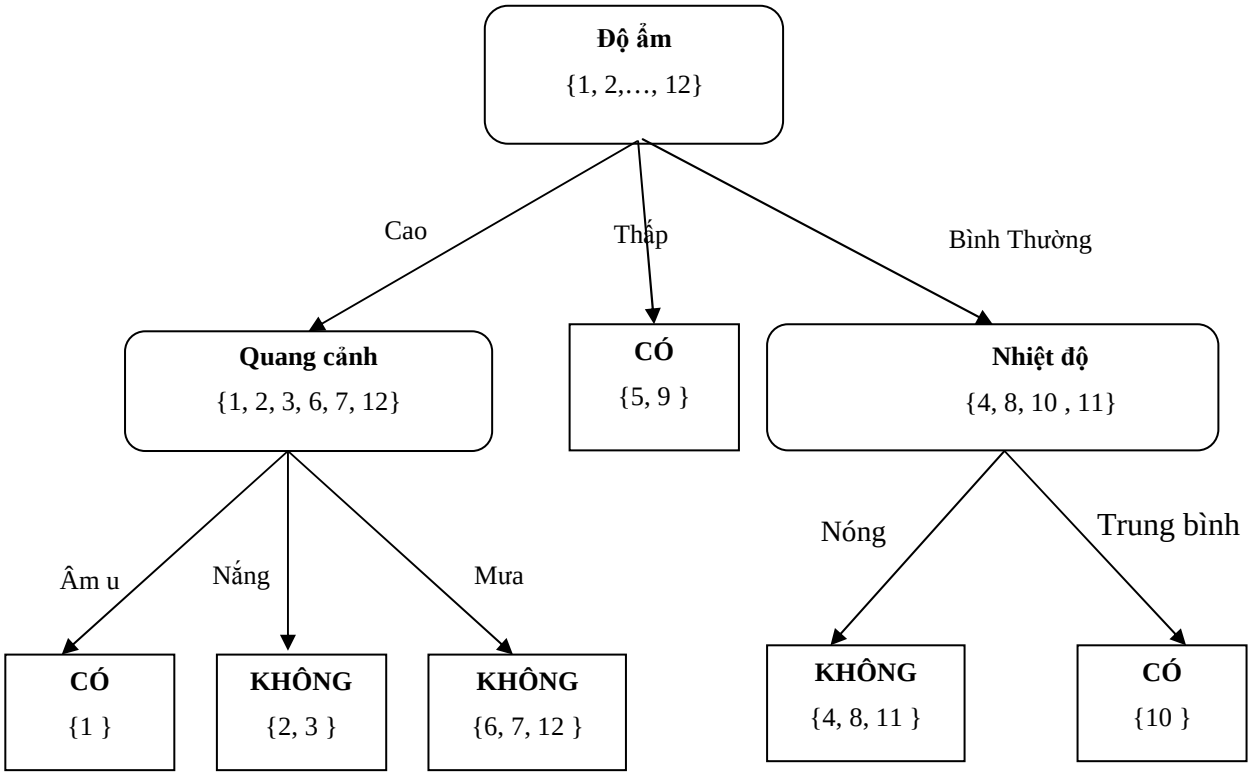
$$\text{IG}(U_2, \text{Nhiệt độ}) = 0.811 - [\frac{3}{4}(-\frac{3}{3}\log_2\frac{3}{3}) + \frac{1}{3}(-\frac{1}{1}\log_2\frac{1}{1})]$$

$$= 0.811\text{-}0 = 0.811$$

Ta thấy chỉ số IG của "Nhiệt độ" là lớn nhất, nên nó được chọn để phân chia. Do đó, thuộc tính "Nhiệt độ" làm nhãn cho nút bên phải nối với nhánh "Trung bình".Trong U_2, thuộc tính này có hai giá trị "Nóng" và "Ấm áp" nên ta tiếp tục tạo thành hai nhánh mới là "Nóng" và "Ấm áp":

- Với nhánh “Nóng” gồm ba mẫu {4, 8, 11} và đều có giá trị quyết định là “KHÔNG” nên ta tạo nút lá là “KHÔNG”.
- Với nhánh “Ấm áp” gồm một mẫu {10} và có giá trị quyết định là “CÓ ” nên tạo nút lá là “CÓ ”.

Cuối cùng thu được cây như sau:



Hình 6: Cây kết quả (ID3)

2.4 Kết luận

Chương này, luận văn đã trình bày tổng quan về một quy trình tín dụng cơ bản mà bất kỳ các ngân hàng thương mại nào cũng phải tuân theo. Bên cạnh đó luận văn cũng trình bày phương pháp tổng quát xây dựng cây quyết định; ba thuật toán xây dựng cây quyết định ID3; các ví dụ cụ thể để minh họa từng bước trên mỗi thuật toán. Trong chương sau, luận sẽ trình bày một ứng dụng cụ thể mà các ngân hàng có thể áp dụng để phân loại khách hàng của mình, căn cứ vào kết quả ngân hàng sẽ có thêm sự hỗ trợ để quyết định có cho họ vay vốn hay không.

Chương 3 - THỬ NGHIỆM VÀ ĐÁNH GIÁ

3.1 Giới thiệu bài toán

Trong chương này, luận văn tập trung nghiên cứu đối với công tác tín dụng tiêu dùng đặc biệt là trong việc mua nhà giá thấp như hiện nay của khách hàng với tập dữ liệu Dulieu_nganhang.xls. Dựa vào tập Dulieunganhang.xls sẽ xây dựng mô hình cây quyết định, từ cây quyết định rút ra các luật quyết định. Dựa vào các luật quyết định đó ta sẽ phân lớp được tập dữ liệu mới (dữ liệu về khách hàng xin vay tiêu dùng, nhưng chưa được phân lớp) và tập dữ liệu sau khi được phân lớp sẽ hỗ trợ cho các cán bộ tín dụng ra quyết định cho khách hàng vay hay không.

3.2 Cơ sở dữ liệu

Luận văn sử dụng tập dữ liệu: Dulieunganhang.xls gồm 600 đối tượng với 10 thuộc tính điều kiện và thuộc tính quyết định "result" quyết định một khách hàng là được vay và không được vay.

Bảng 2: Bảng các thuộc tính của tập dữ liệu Dulieunganhang

Thứ tự	Thuộc tính	Giá trị	Ý nghĩa
1	Tuoi	Tre, Trungnien, Gia	Trẻ, trung niên, già
2	Gioitinh	Nam, Nu	Nam, Nữ
3	Hokhau	NongThon, ThiTran, NgoaiO, ThanhPho	Nông thôn, Thị trấn, Ngoại ô, Thành phố
4	Thunhap	Thap,Trungbinh, Cao	Thấp, trung bình, cao
5	Kethon	Co, Khong	Có, không
6	SoCon	Khongcon, Motcon, Haicon, Bacon	Không con, Một con, Hai con, Ba con
7	XeOto	Co, Khong	Có, không
8	TaikhoaTietkiem	Co, Khong	Có, không
9	TaikhoanHientai	Co, Khong	Có, không
10	TaisanThechap	Co, Khong	Có, không
11	RESULT(Chovay)	True, false	Có (True), Không (False)

3.3 Cài đặt ứng dụng

Chương trình gồm các mô đun chính:

1. Đọc dữ liệu đầu vào từ file Excel.
2. Kiểm tra dữ liệu.
3. Tạo cây quyết định.
4. Tạo luật được sinh ra từ cây quyết đinh.
5. Đánh giá độ chính xác của thuật toán.

- Thuật toán tạo cây quyết được cụ thể hóa bằng việc lập trình với nội dung câu lệnh như sau:

```
'Lớp xây dựng cây quyết định của thuật toán DecisionTree*/
    Public Class DecisionTree
    Private mSamples As DataTable
    Private mTotalPositives As Integer = 0
    Private mTotal As Integer = 0
    Private mTargetAttribute As String = "RESULT
    Public mTrueValue As String = "True"
    Private mFalseValue As String = "False"
    Private mEntropySet As Double = 0.0
    'Trả về số phần tử True trong bảng quyết định
    Private Function countTotalPositives(ByVal samples As DataTable) As Integer
    Dim result As Integer = 0
    For Each aRow As DataRow In samples.Rows
    Dim s As String = "True"
    If Not (aRow(mTargetAttribute).ToString().Trim().ToUpper() = mTrueValue.ToUpper()) Then s = "False"
    If Boolean.Parse(s) = True Then result = result + 1
    Next
    Return result
    End Function
```

```
    ' Duyệt qua bảng và kiểm tra thuộc tính có giá trị là value và trả về số phần tử True và số phần tử âm.  */
    Private Sub getValuesToAttribute(ByVal samples As DataTable, ByVal attribute As Attribute, ByVal value As String, ByRef positives As Integer, ByRef negatives As Integer)
    positives = 0
    negatives = 0
    For Each aRow As DataRow In samples.Rows
    If CType(aRow(attribute.AttributeName), String) = value Then
    Dim s As String = "True"
    IF Not (aRow(mTargetAttribute).ToString().Trim().ToUpper() = mTrueValue.ToUpper()) Then s = "False"
    If Boolean.Parse(s) = True Then
    positives = positives + 1
    Else
    negatives = negatives + 1
    End If
    End If
    Next
    End Sub
```

- **Thủ tục tính toán Entropy:** $Entropy(U) = -\sum_{i=1}^{k} p_i \log_2 p_i$

```
    'Tính entropy -p+log(p+,2) + p-log(p-,2)
    Private Function calcEntropy(ByVal positives As Integer, ByVal negatives As Integer) As Double
    Dim total As Integer = positives + negatives
    Dim ratioPositive As Double = CType(positives /
    Dim ratioNegative As Double = CType(negatives / total, Double)
```

```
    ' Cây sẽ ngưng làm việc khi phát hiện
root.Attribute.value chứa giá trị  null
    If total = 0 Then Return 0
    If Not (ratioPositive = 0) Then ratioPositive = -
(ratioPositive) * System.Math.Log(ratioPositive, 2)
    If Not (ratioNegative = 0) Then ratioNegative = -
(ratioNegative) * System.Math.Log(ratioNegative, 2)
    Dim result As Double = ratioPositive + ratioNegative
    Return result
    End Function
```

- **Thủ tục tính lượng thông tin IG**

```
    'Tính lượng thông tin thu thêm (IG):
```

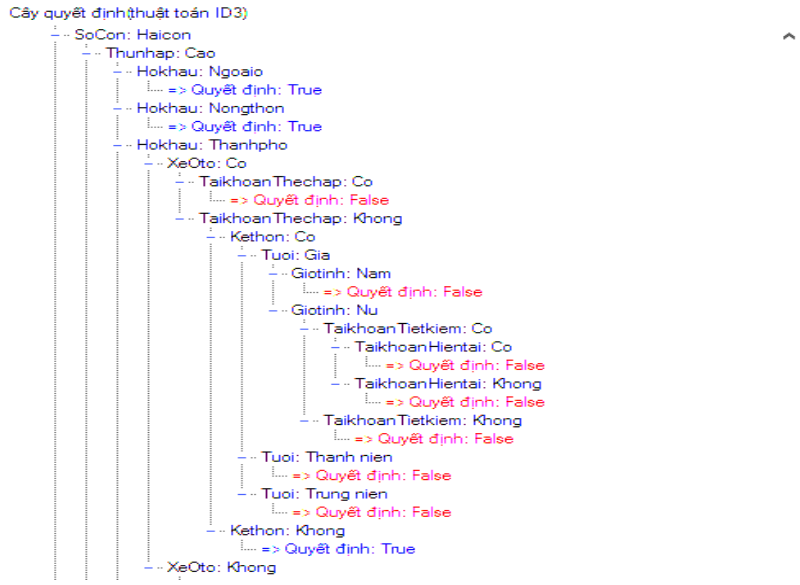
$$IG(U,c) = Entropy(U) - \sum_{v \in V_c} \frac{|U_v|}{|U|} Entropy(U_v)$$

```
    Private Function gain(ByVal samples As DataTable, ByVal
attribute As Attribute) As Double
    Dim values() As String = attribute.values
    Dim sum As Double = 0.0
    Dim _len As Integer = values.Length - 1
    For i As Integer = 0 To _len
    Dim positives, negatives As Integer
    positives = negatives = 0
    getValuesToAttribute(samples, attribute, values(i),
positives, negatives)
    Dim entropy As Double = calcEntropy(positives, negatives)
    sum += -CType((positives + negatives) / mTotal * entropy,
Double)
    Next i
    Return mEntropySet + sum
    End Function
```

3.4 Kết quả và đánh giá

3.4.1 Mô hình cây quyết định và các luật

Với tập dữ liệu như đã trình bày (Dulieunganhang.xls: 600 mẫu,10 thuộc tính, 1 thuộc tính kết luận), sau khi được chạy với chương trình, nó sinh ra 238 luật với mô hình cây quyết định như sau:



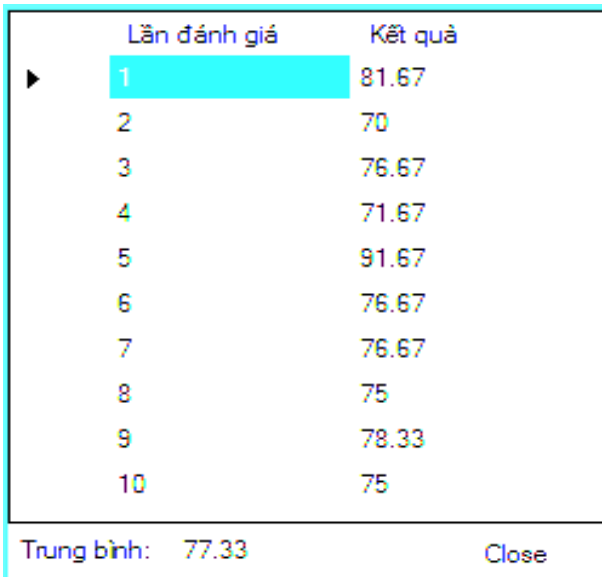
Hình 7: Cây quyết định ứng với thuật toán ID3

- Các luật quyết định ứng với cây quyết định ID3

Hình 8: Một số luật của cây quyết định ID3

3.4.2 Đánh giá thuật toán và ứng dụng của cây quyết định trong việc hỗ trợ cán bộ tín dụng

Để đánh giá hiệu suất của một cây quyết định người ta thường sử dụng một tập ví dụ tách rời, tập này khác với tập dữ liệu huấn luyện, để đánh giá khả năng phân loại của cây trên các ví dụ của tập này. Tập dữ liệu này gọi là tập kiểm tra. Thông thường, tập dữ liệu sẵn có sẽ được chia thành hai tập: tập huấn luyện thường chiếm 2/3 tổng số mẫu và tập kiểm tra chiếm 1/3 tổng số mẫu. Luận văn cũng sử dụng phương thức này để đánh giá thuật toán ID3 theo tập dữ liệu: Dulieunganhang.xls Đánh giá độ chính xác của thuật toán với số lần là 10 trên bộ dữ liệu Dulieunganhang.xsl , ta được kết quả như sau:



Hình 9: Độ chính xác của thuật toán ID3

3.5 Kết luận

Trong chương này, luận văn đã sử dụng bộ dữ liệu Dulieunganhang.xls để kiểm chứng các thuật toán xây dựng cây quyết định ở chương 2. Bộ dữ liệu này với 600 bản ghi và 10 thuộc tính, nó rất phù hợp trong việc sử dụng cây quyết định để phân loại khách hàng vay vốn tại các ngân hàng thương mại. Đồng thời, dựa vào mô hình cây quyết định (các luật quyết định) đã được xây dựng, luận văn cũng đánh giá, phân tích các luật trong quá trình phân loại khách hàng để từ đó tiếp tục hỗ trợ việc ra quyết định cho khách hàng vay vốn tại các ngân hàng thương mại .

KẾT LUẬN

Qua hai năm học tập, tìm tòi, nghiên cứu, đặc biệt là trong khoảng thời gian làm luận văn, tác giả đã hoàn thiện luận văn với các mục tiêu đặt ra ban đầu. Cụ thể luận văn đã đạt được những kết quả sau:

- Trình bày các kiến thức cơ bản về khám phá tri thức và khai phá dữ liệu.
- Giới thiệu phương pháp tổng quát xây dựng cây quyết định, trình bày thuật toán xây dựng cây quyết định ID3 cùng một số ví dụ minh họa cho các phương pháp xây dựng cây quyết định.
- Cài đặt bằng Visual Basic thuật toán xây dựng cây quyết định ID3 trên cơ sở dữ liệu mẫu Dulieunganhang.xsl. Đánh giá độ chính xác của các thuật toán trên và đánh giá độ chính xác của từng luật trong mô hình cây quyết định.

Một số vấn đề luận văn phải tiếp tục nghiên cứu, tìm hiểu:

- Cần tiếp tục nghiên cứu các thuật toán khai phá dữ liệu bằng cây quyết định: thuật toán ADTCCC (dựa vào CORE và đại lượng đóng góp phân lớp của thuộc tính), thuật toán ADTNDA (dựa vào độ phụ thuộc mới của thuộc tính) …
- Cần bổ sung thêm dữ liệu cho tập huấn luyện để mô hình cây quyết định có độ tin cậy cao hơn và hoạt động hiệu quả hơn. Tiếp tục phát triển hoàn thiện theo hướng trở thành phần mềm khai phá dữ liệu trong tín dụng tiêu dùng nhằm hỗ trợ cho cán bộ tín dụng đưa ra quyết định cho khách hàng vay hay không.
- Tìm hiểu nhu cầu thực tế để từ đó cải tiến chương trình, cài đặt lại bài toán theo các thuật toán đã nghiên cứu để làm việc tốt hơn với các cơ sở dữ liệu lớn và có thể có được sản phẩm trên thị trường.