Ứng dụng cây quyết định vào việc phân loại khách hàng vay tiêu dùng tại ngân hàng thương mại

ThS. NGUYỄN DƯƠNG HÙNG

Học viện Ngân hàng

Những năm gần đây, kỹ thuật khai phá dữ liệu (DM-Data Mining) đã được nhiều ngân hàng đề xuất, khuyến nghị sử dụng trong việc hỗ trợ ra quyết định của ngân hàng. Khai phá dữ liệu có thể giúp cho các ngân hàng có những chiến lược cạnh tranh tốt hơn trên thị trường trong việc phân khúc khách hàng, chấm điểm tín dụng, phê duyệt, quảng bá, giới thiệu sản phẩm, phát hiện các giao dịch gian lận. Một trong những công cụ khai phá dữ liệu hiệu quả hiện nay là sử dụng cây quyết định (Decision Tree) để tìm ra các luật phân lớp. Bài báo này sẽ nghiên cứu về việc ứng dụng cây quyết định để phân loại khách hàng vay tiêu dùng tại các ngân hàng thương mại, từ đó có cơ sở cho các quyết định cho vay.

1. Tổng quan về khai phá dữ liệu

Khai phá dữ liệu là một quá trình quan trọng trong quá trình tìm kiếm tri thức từ dữ liệu. Trong quá trình này, các chuyên gia bao gồm cả chuyên gia về công nghệ thông tin và chuyên gia của các doanh nghiệp, phải đặt ra được bài toán là cần các thông tin gì cho việc hỗ trợ kinh doanh, lấy các thông tin đó như thế nào và lấy ở đâu, bằng phương pháp nào cho hiệu quả nhất. Đó chính là bài toán về khai phá dữ liệu tìm kiếm tri thức hỗ trợ quyết định. Thông thường một bài toán như vậy gồm các bước:

Bước 1: Xác định vấn đề và lựa chọn nguồn dữ liệu *(Problem Understanding and Data Understanding).* Ở bước này, các chuyên gia trong lĩnh vực, ngành đặc thù cần thảo luận với các chuyên gia tin học, để xác định được chúng ta mong muốn khám phá những gì, thống nhất giải pháp cho quá trình khám phá dữ liệu (muốn có các qui luật hay muốn phân lớp, phân cụm dữ liệu…). Đây là một giai đoạn quan trọng vì nếu xác định sai vấn đề thì toàn bộ quá trình trở nên vô ích.

Bước 2: Chuẩn bị dữ liệu *(Data preparation)* gồm các bước sau: (i) Thu thập dữ liệu (Data gathering); (ii) Làm sạch dữ liệu (Data cleaning); (iii) Tích hợp dữ liệu (Data integeration); (iv) Chọn dữ liệu (Data selection); (v) Biến đổi dữ liệu

(Data transformation).

Đây cũng là một bước rất quan trọng vì nếu dữ liệu đầu vào không chính xác thì hiển nhiên sẽ không thể nào có một kết quả chính xác, không có giá trị hỗ trợ ra quyết định.

Bước 3: Khai phá dữ liệu *(Data Mining),* đây là bước xác định nhiệm vụ khai phá dữ liệu và lựa chọn kỹ thuật khai phá dữ liệu. Kết quả của quá trình này sẽ tìm ra các tri thức, mô hình hay các quy luật tiềm ẩn bên trong dữ liệu.

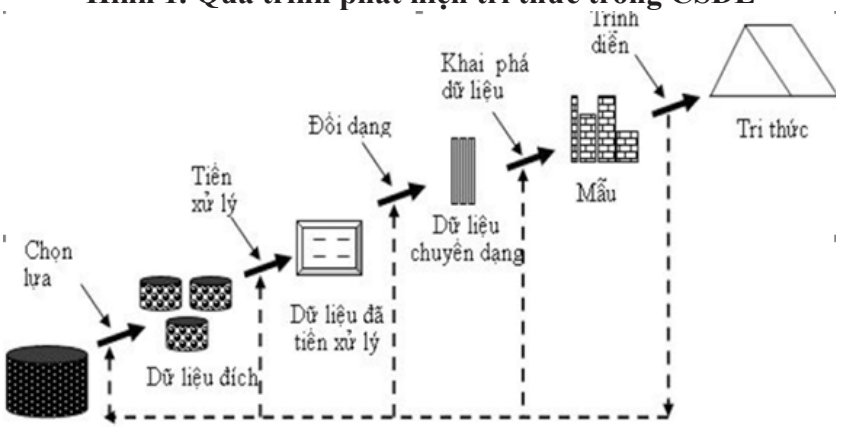
Bước 4: Đánh giá mẫu *(Partern Evalution)*: Đánh giá xem tri thức thu được có chính xác và có giá trị hay không, nếu không có thể quay lại các bước trên. Việc đánh giá này được thực hiện thông qua các chuyên gia trong từng lĩnh vực và người dùng cuối (end user) là chính, chứ không phải là các chuyên gia tin học.

Bước 5: Biểu diễn tri thức và triển khai *(Knowlegde Presentation and Deployment):* Biểu diễn tri thức phát hiện được dưới dạng tường minh, thân thiện và hữu ích với đa số người dùng và tiến hành đưa tri thức phát hiện được vào các ứng dụng cụ thể.

Một cách tổng quát, 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[2],[5].

Trong quá trình trên (Hình 1), khai phá dữ liệu là một khái niệm được dùng để mô tả quá trình phát hiện tri thức trong cơ sở dữ liệu (CSDL). Quá trình

Hình 1. Quá trình phát hiện tri thức trong CSDL



này kết xuất ra các tri thức ẩn chứa trong dữ liệu, giúp cho việc dự báo trong kinh doanh, các hoạt động sản xuất. Qui trình gồm có 6 giai đoạn[2]:

Giai đoạn 1: Thu thập dữ liệu *(Data Gathering).* Đây là bước tập hợp các dữ liệu được khai thác trong một CSDL, một kho dữ liệu và thậm chí các dữ liệu từ các nguồn ứng dụng Web.

Giai đoạn 2: Trích lọc dữ liệu *(Data Selection).* Ở giai đoạn này, dữ liệu được lựa chọn hoặc phân chia theo một số tiêu chuẩn nào đó, ví dụ chọn tất cả những khách hàng có tài khoản thế chấp là nhà ở của chính họ.

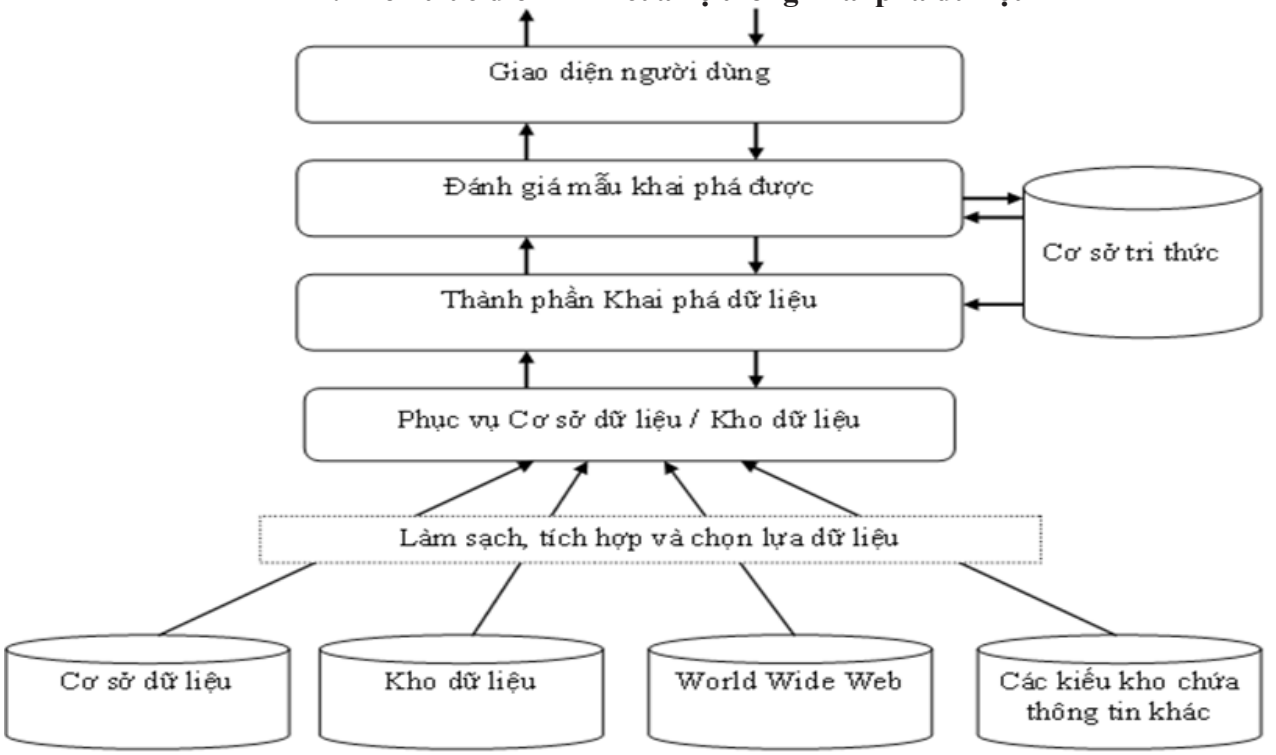
Giai đoạn 3: Làm sạch, tiền xử lý và chuẩn bị dữ liệu *(Cleansing, Pre-processing and Preparation).* Đây là một bước rất quan trọng trong quá trình khai phá dữ liệu. Một số lỗi thường mắc phải trong khi gom dữ liệu là dữ liệu không đủ tính chặt chẽ, logic; dữ liệu thường chứa các giá trị không có ý nghĩa và không có khả năng kết nối. Giai đoạn này sẽ tiến hành xử lý những dạng dữ liệu không chặt chẽ, không lôgic nói trên vì chúng là thông tin dư thừa, không có giá trị. Bởi vậy, đây là một quá trình rất quan trọng vì dữ liệu này nếu không được "làm sạch-tiền xử lý- chuẩn bị trước" thì sẽ dẫn đến những kết quả sai lệch nghiêm trọng, từ đó sẽ dẫn tới các quyết định không chính xác.

Giai đoạn 4: Chuyển đổi dữ liệu *(Data Transformation).* Dữ liệu thô sẽ được chuyển đổi sang các dạng dữ liệu phù hợp với mục đích khai thác.

Giai đoạn 5: Phát hiện và trích mẫu dữ liệu *(Pattern Extraction and Discovery).* Ở giai đoạn này, nhiều thuật toán khác nhau được sử dụng để trích ra các mẫu từ dữ liệu. Thuật toán thường dùng là nguyên tắc phân loại, nguyên tắc kết hợp hoặc các mô hình dữ liệu tuần tự.

Giai đoạn 6: ***Đánh giá kết quả mẫu*** *(Evaluation of Result).* Đây là giai đoạn cuối trong quá trình khai phá dữ liệu. Ở giai đoạn này, các mẫu dữ liệu được chiết xuất ra bởi phần mềm khai phá dữ liệu. Không phải bất cứ mẫu dữ liệu

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



nào cũng đều hữu ích. Vì vậy, cần phải ưu tiên những tiêu chuẩn đánh giá để chiết xuất ra các tri thức (Knowledge) cần thiết. Quá trình khai phá dữ liệu được mô hình hóa một cách tổng quát như Hình 2.

2. Ứng dụng khai phá dữ liệu trong lĩnh vực ngân hàng

Ngành công nghiệp ngân hàng trên thế giới đã trải qua những thay đổi to lớn trong cách thức kinh doanh của họ. Áp dụng công nghệ thông tin vào công việc kinh doanh ngân hàng đã tạo nên sự thay đổi lớn, theo đó, việc thực hiện giao dịch đã trở nên dễ dàng, đồng thời khối lượng dữ liệu từ các giao dịch này đã tăng lên đáng kể. Việc phân tích số lượng dữ liệu thô khổng lồ này và chuyển đổi nó thành thông tin hữu ích cho các ngân hàng nhằm hỗ trợ ra các quyết định kinh doanh trở thành một vấn đề thiết yếu. Bằng cách sử dụng khai phá dữ liệu để phân tích, các ngân hàng có thể dự đoán với độ chính xác tăng lên về những tình huống liên quan đến các quyết định kinh doanh của mình, ví dụ khách hàng sẽ phản ứng thế nào với việc điều chỉnh lãi suất, trong đó khách hàng nào sẽ có khả năng chấp nhận sự chào hàng sản phẩm mới, khách hàng nào sẽ có nguy cơ rủi ro cao hơn, và làm thế nào để mối quan hệ khách hàng ngày càng có lợi.

Lĩnh vực khai phá dữ liệu được được ứng dụng trong ngành công nghiệp ngân hàng tương đối rộng rãi, trong đó bao gồm việc phân khúc khách hàng và phân chia lợi nhuận, chấm điểm và phê duyệt tín dụng, dự đoán thanh toán mặc định, quảng bá sản phẩm, phát hiện các giao dịch gian lận, quản lý tiền mặt và các hoạt động dự báo, tối ưu hóa danh mục đầu tư chứng khoán và xếp hạng đầu tư. Các ngân hàng đã và đang sử dụng có hiệu quả kỹ thuật khai phá dữ liệu trong các lĩnh vực sau:

a. Marketing: Một trong những lĩnh vực được ứng dụng rộng rãi nhất cho ngành ngân hàng của kỹ thuật khai phá dữ liệu 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 CSDL về khách hàng. Bộ phận khai phá dữ liệu của các ngân hàng thực hiện các phân tích khác nhau trên bộ dữ liệu thu thập được để xác định hành vi của người tiêu dùng với sự tham khảo sản phẩm, giá và kênh phân phối. Với sự phản hồi của khách

hàng đối với các sản phẩm hiện có và các sản phẩm mới, các ngân hàng sẽ có các chiến lược quảng bá sản phẩm, nâng cao chất lượng sản phẩm và dịch vụ và đạt được lợi thế cạnh tranh. Kỹ thuật khai phá dữ liệu giúp ngân hàng phân tích các xu hướng trong quá khứ, xác định nhu cầu hiện tại và dự báo hành vi khách hàng với các sản phẩm và dịch vụ khác nhau để chuẩn bị cho các cơ hội kinh doanh mới. 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. Các kỹ thuật khai phá dữ liệu có thể được sử dụng để xác định phản hồi của khách hàng như thế nào khi ngân hàng thực hiện điều chỉnh lãi suất.

b. 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]. Khi cung cấp thẻ tín dụng mới cho khách hàng hay phê duyệt các khoản vay, các ngân hàng phải kiểm tra các thông tin khác nhau liên quan đến khoản tín dụng của mình. Kỹ thuật khai phá dữ liệu giúp phân biệt người trả nợ kịp thời với những người không có khả năng trả nợ kịp thời.

Trên thực tế, điểm tín dụng là một trong những công cụ quản lý rủi ro tài chính trước tiên được phát triển[4], là căn cứ giúp ngân hàng đưa ra những quyết định cho vay. Khai phá dữ liệu có thể tìm ra được hành vi tín dụng của từng khách hàng cá nhân với các khoản vay trả góp, thế chấp, tín dụng, bằng việc sử dụng các đặc điểm như lịch sử tín dụng, thời gian làm việc và thời gian cư trú, giúp ngân hàng đánh giá khách hàng và quyết định khách hàng đó có là một ứng cử viên tốt cho một khoản vay, hoặc có rủi ro nào tiềm ẩn nhằm giảm thiểu rủi ro trong cấp tín dụng.

c. 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. Với sự giúp đỡ của kỹ thuật khai phá dữ liệu, các hành động gian lận ngày càng được phát hiện nhiều hơn. Có hai phương pháp tiếp cận phổ biến đã được phát triển bởi tổ chức tài chính để phát hiện các mô hình gian lận[4]. Phương pháp tiếp cận thứ nhất, một ngân hàng cần phải sử dụng đến kho dữ liệu của bên thứ ba và sử dụng các kỹ thuật khai phá dữ liệu để xác định mô hình gian lận, sau đó, các ngân hàng có thể tham chiếu chéo những mẫu với CSDL riêng của mình. Phương pháp thứ hai, gian lận được nhận dạng mẫu dựa trên các mẫu thông tin nội bộ riêng của mình mà không phải nhờ vào bên thứ ba. Tuy nhiên, trên thực tế hầu hết các ngân hàng đang sử dụng kết hợp cả hai phương pháp tiếp cận trên.

d. Quản trị quan hệ khách hàng: Trong ngành ngân hàng, việc quản trị và phát triển các mối quan hệ khách hàng (CRM: Customer Relationship Management) một cách hiệu quả là một vấn đề quan trọng. Để làm được điều này, các ngân hàng cần phải đầu tư các nguồn lực để hiểu rõ hơn về khách hàng hiện tại và tiềm năng của họ. Sử dụng các công cụ khai phá dữ liệu phù hợp để tìm ra các sản phẩm và dịch vụ thích hợp có thể cung cấp cho khách hàng là một cách hiệu quả để đạt được mục tiêu này. Kỹ thuật 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. Bằng cách phân tích các dữ liệu trong quá khứ, khai phá dữ liệu có thể giúp các ngân hàng dự đoán số lượng khách hàng có khả năng thay đổi thẻ tín dụng của họ, từ đó họ có thể lập kế hoạch và triển khai ưu đãi đặc biệt khác nhau để giữ lại những khách hàng của mình.

Kỹ thuật khai phá dữ liệu giúp ngân hàng phân tích và nhận định được đâu là các khách hàng trung thành và đâu là các khách hàng có xu hướng chuyển sang các ngân hàng khác với mong muốn một dịch vụ tốt hơn, giúp các ngân hàng hoạt động tốt hơn và giữ chân khách hàng của mình.

3. Ứng dụng cây quyết định vào phân loại khách hàng trong quy trình tín dụng

a. Tổng quan về 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 để ra một quyết định 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. 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.

Về mặt quản lý, quy trình tín dụng là cơ sở cho việc phân định quyền, trách nhiệm cho các bộ phận trong hoạt động tín dụng; là cơ sở để thiết lập các hồ sơ, thủ tục vay vốn. Thông thường, một qui trình tín dụng gồm có: Lập hồ sơ vay vốn, phân tích tín dụng, ra quyết định, giải ngân, giám sát và thanh lý hợp đồng.

b. Lựa chọn thuật toán

Để ra quyết định tín dụng chính xác và để đảm bảo tính khách quan, các ngân hàng có thể sử dụng các tri thức/thông tin được trích xuất được từ hồ sơ khách hàng đã có. Các tri thức/thông tin này sẽ giúp ngân hàng tránh được rủi ro như từ chối một khách hàng tiềm năng hoặc cho một khách hàng không có khả năng thanh toán vay vốn. Thuật toán cây quyết định có thể dự đoán hoặc phân loại khách hàng bằng cách dựa trên cơ sở dữ liệu lịch sử đã có. Thuật toán cây quyết định bao gồm thuật toán ID3, thuật toán C4.5, thuật toán CART. Trong các thuật toán đó, thuật toán ID3 là một thuật toán được đánh giá có một cách thể hiện rõ ràng, dễ hiểu nhất. Do vậy, bài báo này sẽ sử dụng thuật toán ID3 để xây dựng cây quyết định phân loại khách hàng vay vốn tại ngân hàng.

Thuật toán cây quyết định là công cụ được dùng để phân lớp dữ liệu, mỗi cây quyết định tượng trưng cho một sự quyết định của một lớp các dữ kiện nào đó. Mỗi nút trong cây là tên của một lớp hay một phép thử thuộc tính cụ thể nào đó, phép thử này phân chia không gian trạng thái các dữ kiện tại nút đó thành các kết quả có thể đạt được của phép thử. Mỗi tập con được phân chia của phép thử là không gian con của các sự kiện, nó tương ứng với một vấn đề con của sự phân lớp. Các cây quyết định được dùng để hỗ trợ quá trình ra quyết định.

Cây quyết định (Decision Tree) có thể định nghĩa, diễn giả bằng một tập các luật ***IF–THEN***, với cách trình bày như vậy nó sẽ giúp cho người đọc dễ đọc và dễ hiểu. Cây quyết định có thể thực hiện được cả với các dữ liệu chứa lỗi (noisy data). Về bản chất, cây quyết định là một trong các phương pháp quy nạp được dùng phổ biến nhất trong quá trình xử lý dữ liệu. Một cách tổng thể, 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.

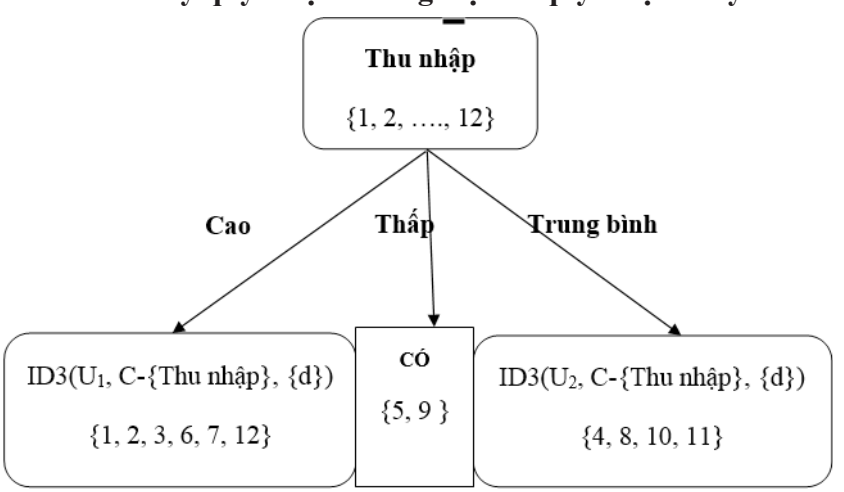
Nút lá (Leaf Node) hay còn gọi là nút trả lời biểu thị cho một lớp các trường hợp mà nhãn của nó là tên của lớp, nó biểu diễn một lớp (a classififcation).

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.

Để phân lớp mẫu dữ liệu chưa biết, giá trị các thuộc tính của mẫu được đưa vào kiểm tra trên cây quyết định. Mỗi mẫu tương ứng có một đường đi từ

Hình 3. Cây quyết định trong việc ra quyết định vay vốn



gốc đến lá và lá biểu diễn dự đoán giá trị phân lớp của mẫu đó.

Tiêu chí để đánh giá tìm điểm chia là rất quan trọng, chúng được xem là một tiêu chuẩn "Heuristic", là tiêu chuẩn mà việc tìm kiếm dựa vào tri thức hiện tại và trong quá khứ, thỏa mãn các tính chất : (i) Xác định phương án rõ ràng, không mập mờ và có thể thực thi được; (ii) có tính hữu hạn, sau một số bước phải có lời giải cho bài toán; (iii) tính đúng đắn, chắc chắn có những lời giải tốt mặt dù đó chưa phải là tốt nhất để phân chia dữ liệu. Ý tưởng chính trong việc đưa ra các tiêu chí là làm sao cho các tập con được phân chia càng trở nên "trong suốt" (tất cả các bộ thuộc về cùng một lớp) 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, 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:

$$\text{Entropy}(U) = -\sum p_i \log_2 p_i \quad (i = 1 \rightarrow k)$$

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 số 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, được xác định theo công thức sau[6]:

$$\text{IG}(U,c) = \text{Entropy}(U) - \sum |U_v| \text{Entropy}(U) / |U| \quad (v \in V_c)$$

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.

Thuật toán ID3[1] là giải thuật tìm kiếm tham lam (greedy search) dùng để xây dựng cây quyết định. Ý tưởng chính của thuật toán ID3 là xây dựng cây quyết định (Decission Tree) bằng cách ứng dụng từ trên xuống (Top-Down), bắt đầu từ một tập các đối tượng và các thuộc tính của nó. Tại mỗi nút của cây một thuộc tính được kiểm tra, kết quả của phép kiểm tra này được sử dụng để phân chia tập đối tượng theo kết quả kiểm tra trên. Quá trình này được thực hiện một cách lặp lại (đệ quy) cho tới khi tập đối tượng trong cây con được sinh ra thuần nhất theo một tiêu chí phân lớp nào đó, hay các đối tượng đó thuộc cùng một dạng giống nhau nào đó. Các lớp hay các dạng này được gọi là nhãn của nút lá của cây, còn tại mỗi nút không phải là nút lá thì nhãn của nó là tên thuộc tính được chọn trong số các thuộc tính được dùng để kiểm tra có giá trị **IG (Information Gain)** lớn nhất. Đại lượng **IG** được tính thông qua hàm **Entropy**. Như vậy, **IG** là đại lượng được dùng để đưa ra độ ưu tiên cho thuộc tính nào được chọn trong quá trình xây dựng cây quyết định.

c. Minh họa thuật toán ID3

Để minh họa cho thuật này, chúng tôi đưa một ví dụ về bài toán phân loại xem một người có được ngân hàng xét duyệt cho vay với các tham số (Tuổi, Tài khoản hiện tại, Thu nhập, Số con). Xét bảng quyết định DT = {U, C ∪ {d}}(Bảng 1)

Tập dữ liệu này gồm có 12 mẫu, một mẫu biểu diễn cho một khách hàng có được cho vay vốn hay không gồm các thuộc tính Tuổi, Số con, Thu nhập và Tài khoản hiện tại; và đều có một thuộc tính quyết định có cho vay vốn hay không. Thuộc tính quyết định chỉ có hai giá trị Có, Không. Mỗi thuộc tính đều có một tập các giá trị hữu hạn. Thuộc tính Tuổi có ba giá trị: Trẻ, Già, Trung niên; Số con có ba giá trị: Hai con, Không con, Một con; Thu nhập có hai giá trị: Cao, Trung bình và Tài khoản hiện tại có hai giá trị: Có, Không. Các giá trị này là ký hiệu dùng để biểu diễn bài toán.

Bảng 1. Dữ liệu mẫu

Khách hàng	Tuổi	Tài khoảnHT	Số con	Thu nhập	Quyết định
1	Trẻ	Có	Không con	Cao	Có
2	Trung niên	Không	Một con	Cao	Không
3	Trung niên	Không	Hai con	Cao	Không
4	Trẻ	Không	Hai con	Trung bình	Không
5	Trung niên	Có	Hai con	Thấp	Có
6	Già	Không	Một con	Cao	Không
7	Già	Không	Hai con	Cao	Không
8	Già	Không	Hai con	Trung bình	Không
9	Trẻ	Có	Hai con	Thấp	Có
10	Già	Không	Một con	Trung bình	Có
11	Già	Có	Hai con	Trung bình	Không
12	Già	Không	Hai con	Cao	Không

Thuật toán xây dựng cây quyết định với dữ liệu ở Bảng 1 như sau:

Trước tiên nút lá được khởi tạo gồm các mẫu từ 1 đến 12. Để tìm điểm chia tốt nhất, phải tính toán chỉ số IG của tất cả các thuộc tính trên. Trước tiên, 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) = − 4⁄12 $\log_2$4⁄12 − 8⁄12 $\log_2$8⁄12 = 0.918

Tính IG cho từng thuộc tính:

Thuộc tính "Tuổi có ba giá trị là "Trẻ", "Trung niên" và "Già". Căn cứ vào bảng dữ liệu ta có:

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

Giá trị của "Trẻ" có ba bộ {1, 9} có giá trị thuộc tính nhãn là "CÓ" và có một bộ {4} có nhãn lớp là "KHÔNG".

Giá trị của "Trung niên" có một bộ {5} có nhãn lớp là "CÓ" và có hai bộ {2, 3} có nhãn lớp là "KHÔNG";

Giá trị "Già" 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 "Tuổi" xét trên U là:

= 0.918 − [3⁄12(−2⁄3$\log_2$2⁄3 − 1⁄3$\log_2$1⁄3) + 3⁄12(−1⁄3$\log_2$1⁄3 − 2⁄3$\log_2$2⁄3) + 6⁄12(−1⁄6$\log_2$1⁄6 − 5⁄6$\log_2$5⁄6)] = 0.314

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

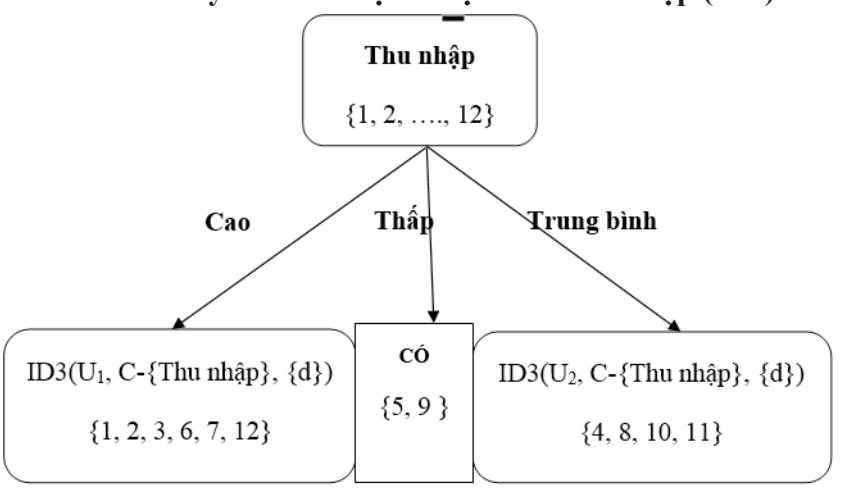
IG(U,Tài khoản hiện tại) = 0.918 − [4⁄12(−3⁄4$\log_2$3⁄4 − 1⁄4$\log_2$1⁄4) + 8⁄12(−1⁄8$\log_2$1⁄8 − 7⁄8$\log_2$7⁄8)] = 0.285

IG(U,Số con) = 0.918 − [3⁄12(−1⁄3$\log_2$1⁄3 − 2⁄3$\log_2$2⁄3) + 8⁄12(−2⁄8$\log_2$2⁄8 − 6⁄8$\log_2$6⁄8)] = 0.148

IG(U,Thu nhập) = 0.918 − [6⁄12(−1⁄6$\log_2$1⁄6 − 5⁄6$\log_2$5⁄6) + 4⁄12(−1⁄4$\log_2$1⁄4 − 3⁄4$\log_2$3⁄4)] = 0.323

Thuộc tính "Thu nhập" là thuộc tính có chỉ số IG lớn nhất nên sẽ được chọn là thuộc tính phân chia. Do đó, thuộc tính "Thu nhập" 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 Hình 4.

Hình 4. Cây sau khi chọn thuộc tính Thu nhập (ID3)



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

Tương tự, để tìm điểm chia tốt nhất tại thuật toán này, phải tính toán chỉ số IG của các thuộc tính "Tuổi", "Tài khoản hiện tại", "Số con".

- Trước 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 "Tuổi", thuộc tính này có ba giá trị là "Trẻ", "Trung niên" và "Già". Nhìn vào bảng dữ liệu:

+ Với giá trị "Trẻ" chỉ có một bộ {1} có giá trị thuộc tính nhãn là "CÓ ".

+ Tương tự giá trị "Trung niên" có hai bộ {2, 3} đều có nhãn lớp là "KHÔNG".

+ Với giá trị "Già" có ba bộ {6, 7, 12} đều có nhãn lớp "KHÔNG".

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

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

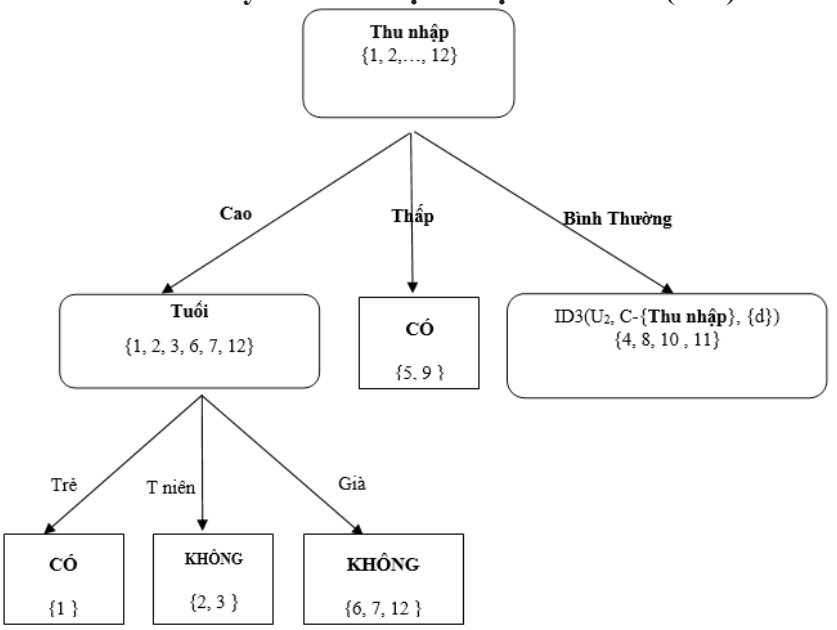
Tính tương tự ta cũng có:

IG(U_1, Tài khoản hiện tại) = $0.65 - [\frac{1}{6}(-\frac{1}{1}\log_2\frac{1}{1}) + \frac{5}{6}(-\frac{5}{5}\log_2\frac{5}{5})] = 0.65$

IG(U_1, Số con) = $0.65 - [\frac{1}{6}(-\frac{1}{1}\log_2\frac{1}{1}) + \frac{5}{6}(-\frac{5}{5}\log_2\frac{2}{2}) = 0.65$

Ta thấy chỉ số IG của ba thuộc tính "Tuổi", "Tài khoản hiện tại", "Số con" là như nhau, ta có thể chọn bất kỳ thuộc tính nào để phân chia.

Hình 5. Cây sau khi chọn thuộc tính Tuổi (ID3)



Giả sử ta chọn thuộc tính "Tuổi" để phân chia, do đó, thuộc tính "Tuổi" làm nhãn cho nút bên trái nối với nhánh "Cao". Thuộc tính này có ba giá trị "Trẻ", "Trung niên" và "Già" nên ta tiếp tục tạo thành ba nhánh mới là "Trẻ", "Trung niên" và "Già":

+ Với nhánh "Trẻ" gồm một mẫu {1} và có giá trị quyết định là "CÓ" nên ta tạo nút lá là "CÓ".

+ Với nhánh "Trung niên" gồm hai mẫu {2, 3} và có cùng giá trị quyết định là "KHÔNG" nên tạo nút lá là "KHÔNG".

+ Với nhánh "Già" 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-{Thu nhập}, {d})**, ta có cây như Hình 5.

Bước tiếp theo gọi thuật toán đệ quy: ID3(U2, C-{ Thu nhập}, {d})

Tính một cách tương tự như trên ta có:

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

IG(U_2, Tuổi) = $0.811 - [\frac{1}{4}(-\frac{1}{1}\log_2\frac{1}{1}) + \frac{3}{4}(-\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3})] = 0.811 - 0.689 = 0.123$

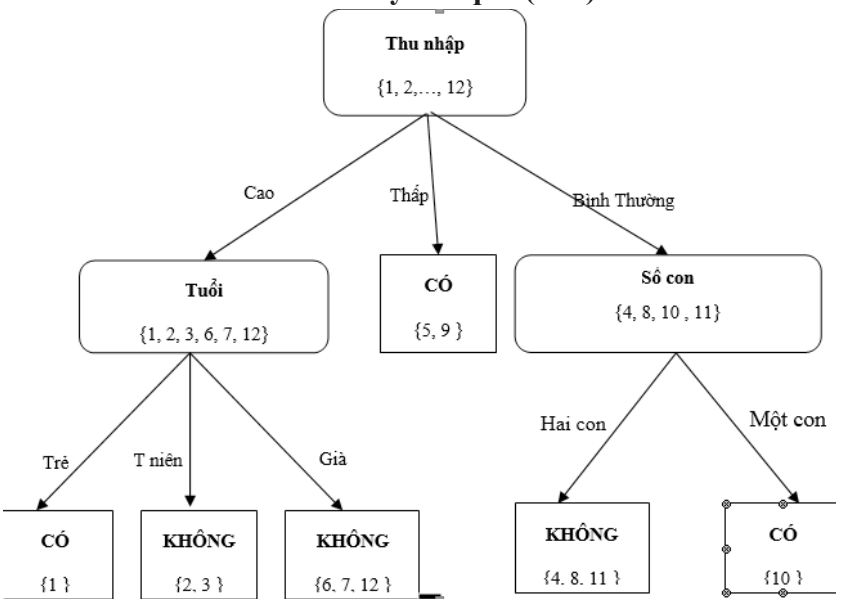
IG(U_2, Tài khoản hiện tại) = $0.811 - [\frac{1}{4}(-\frac{1}{1}\log_2\frac{1}{1}) + \frac{3}{4}(-\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3})] = 0.811 - 0.689 = 0.123$

IG(U_2, Số con) = $0.811 - [\frac{1}{3}(-\frac{1}{1}\log_2\frac{1}{1}) + \frac{3}{4}(-\frac{3}{3}\log_2\frac{3}{3})] = 0.811 - 0 = 0.811$

Ta thấy chỉ số IG của "Số con" là lớn nhất, nên nó được chọn để phân chia. Do đó, thuộc tính "Số con" làm nhãn cho nút bên phải nối với nhánh "Trung bình".

Trong U_2, thuộc tính này có hai giá trị "Hai con" và "Một con" nên ta tiếp tục tạo thành hai nhánh mới là "Hai con" và

Hình 6. Cây kết quả (ID3)



"Một con":

+ Với nhánh "Hai con" gồm ba mẫu {4, 8, 11} và đều có giá trị quyết định là "KHÔNG" nên ta tạo nút lá là "KHÔNG".

+ Với nhánh "Một con" gồm một mẫu {10} và có giá trị quyết định là "CÓ " nên tạo nút lá là "CÓ ".

Cuối cùng thu được cây như Hình 6.

Sau khi cây quyết định được hoàn thành, toàn bộ khách hàng có thể được phân lớp. Ví dụ, trong trường hợp thuộc tính Thu nhập= "Cao" và Tuổi= "Trẻ", khách hàng này được xếp vào lớp khách hàng thuộc diện "có" được vay tiêu dùng. Với kỹ thuật phân loại này, các ngân hàng có thể áp dụng nó vào quá trình ra quyết định cho khách hàng vay vốn tiêu dùng, vì đây là một phương pháp đảm bảo tính khách quan trong việc phân loại khách hàng.

4. Kết luận

Trong quá trình thử nghiệm, tác giả sử dụng tập dữ liệu Bank_data.csv gồm 600 đối tượng, 10 thuộc tính, sau khi tiền xử lí với phần mềm Weka và lưu dưới dạng file excel với tên: Dulieunganhang.xls. Tập dữ liệu này, ngoài các thuộc tính trên, còn có hai thuộc tính quyết định "result", quyết định một khách hàng là được vay hay không được vay.

Bài báo đã trình bày một ứng dụng cụ thể của kỹ thuật khai phá dữ liệu mà các ngân hàng có thể áp dụng để phân loại khách hàng của mình, căn cứ vào kết quả đó ngân hàng sẽ có thêm thông tin về khách hàng để quyết định có cho họ vay vốn hay không. Tuy nhiên, để có kết quả mang tính ứng dụng thực tế, kỹ thuật này cần phải có sự kết hợp với các thuật toán như: 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 dữ liệu mẫu để mô hình cây quyết định có độ tin cậy cao hơn và hoạt động hiệu quả hơn; tiếp tục phát triển hoàn thiện thuật toán theo hướng trở thành phần mềm khai phá dữ liệu trong tín dụng tiêu dùng, nhằm hỗ trợ cho ngân hàng đưa ra quyết định tín dụng cho khách hàng. Đồng thời, cần tìm hiểu nhu cầu thực tế để từ đó cải tiến chương trình, cài đặt bài toán theo các thuật toán đã nghiên cứu để làm việc tốt hơn với các CSDL lớn mang tính thực tế. ■

Tài liệu tham khảo

1. Nguyễn Hà Nam, Giáo trình Khai phá dữ liệu, ĐHQG Hà Nội, năm 2013

2. Hà Quang Thụy, Bài gảng Nhập môn khai phá dữ liệu, ĐHQG Hà Nội, năm 2010

3. http://www.sbv.gov.vn/

4. Hồ Tú Bảo (2001), Introduction to knowledge discovery and data mining, Intistute of Information Technology Nation Center for NaturalScience and Technology.

5. Ian H. Witten, Mark Hall and Eibe Frank (2005), "Data Mining, Practical Machine Learning Tools and Techniques", Second edition, Morgan Kaufmann Publisher.

6. Max Bramer (2007), Principles of Data Mining, University of Portsmouth, UK, Springer Publishers, 2002.

7. http://en.wikipedia.org/wiki/ID3_algorithm, truy nhập ngày 08/02/2014.