HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

Nguyễn Khắc Xuân Bách

NGHIÊN CỨU KHAI PHÁ DỮ LIỆU TRONG QUẢN LÝ RỦI RO TÍN DỤNG NGÂN HÀNG

Chuyên ngành: Hệ thống thông tin

Mã số: 8.48.01.04

TÓM TẮT LUẬN VĂN THẠC SĨ

HÀ NỘI - 2020

Luận văn được hoàn thành tại:

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

Người hướng dẫn khoa học: PGS.TS. Lê Hưu Lập

(Ghi rõ học hàm, học vị)

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

LỜI MỞ ĐẦU

1. Lý do chọn đề tài

Một trong những hoạt động chính của ngân hàng thương mại là hoạt động cho vay nên rủi ro tín dụng là một nhân tố hết sức quan trọng, đòi hỏi các ngân hàng phải có khả năng phân tích, đánh giá và quản lý rủi ro hiệu quả vì nếu ngân hàng chấp nhận nhiều khoản cho vay có rủi ro tín dụng cao thì ngân hàng có khả năng phải đối mặt với tình trạng thiếu vốn hay tính thanh khoản thấp. Điều này có thể làm giảm hoạt động kinh doanh thu lợi nhuận của ngân hàng, thậm chí phá sản. Đã có nhiều giải pháp về mặt nghiệp vụ nhằm hạn chế rủi ro tín dụng ngân hàng. Tuy nhiên, khi CNTT được ứng dụng rộng rãi thì người ta trông chờ vào một giải pháp quản lý rủi ro trong qua trình cho vay tín dụng một cách hiệu quả hơn. Một trong những phương pháp đó chính là ứng dụng khai phá dữ liệu vào lĩnh vực quản lý rủi ro nói chung và rủi ro tín dụng nói riêng nhằm giảm thiểu tình trạng nợ quá hạn, nâng cao chất lượng tín dụng, giảm thiểu khả năng mất vốn của các ngân hàng. Từ lý do đó đề tài luận văn: **"Nghiên cứu Khai phá dữ liệu trong quản lý rủi ro tín dụng ngân hàng"** có ý nghĩa về mặt khoa học và thực tiễn.

2. Tổng quan về đề tài nghiên cứu

Rủi ro tín dụng là một đề tài nghiên cứu quan trọng và rộng khắp trong ngành ngân hàng liên quan đến những quyết định cho vay và khả năng sinh lời. Đối với tất cả ngân hàng, tín dụng được coi là rủi ro lớn nhất và rất khó có thể được bù đắp. Việc áp dụng những kỹ thuật tiên tiến và có tính thống kê trong việc đánh giá rủi ro tín dụng và dự đoán phá sản đã trở thành một lĩnh vực nghiên cứu kể từ thập niên 70. Xếp hạng tín dụng đã trở thành một phương thức phân tích chủ yếu trong những trụ sở kinh tế có liên quan đến rủi ro tín dụng. Mục đích chính của xếp hạng tín dụng là phân chia những ứng viên thành hai nhóm: ứng viên tín dụng tốt và ứng viên với tín dụng xấu. Tính chính xác của xếp hạng tín dụng đóng vai trò rất quan trọng đối với lợi nhuận của tổ chức tài chính. Thậm chí 1% độ chính xác trong việc xếp hạng tín dụng của các ứng viên sẽ giảm tổn thất lớn cho các tổ chức tài chính.

Ngân hàng SHB là một trong những ngân hàng có nợ xấu tăng khá mạnh trong 6 tháng đầu năm 2018, với mức tăng 1 nghìn tỷ đồng, lên hơn 5,6 nghìn tỷ đồng (tương đương với mức tăng 21,7% so với 31/12/2017). Trong đó, nợ có khả năng mất vốn ở mức 3.273 tỷ đồng, tăng 14,2% và chiếm 58,2% tổng nợ xấu. Tỷ lệ nợ xấu của ngân hàng theo đó cũng tăng khá mạnh, từ mức 2,33% đầu năm lên 2,7%/tổng cho vay. Và cũng là ngân hàng có tỷ lệ nợ xấu cao thứ ba trong số 17 ngân hàng. Ở đây học viên chọn giải pháp khai phá dữ liệu để giải quyết bài toán xác định mức độ rủi ro tín dụng của ngân hàng.

3. Mục đích nghiên cứu

Mục đích của đề tài ứng dụng khai phá dữ liệu nhằm nâng cao chất lượng của hệ thống xếp hạng tín dụng của Ngân hàng SHB, để hệ thống xếp hạng tín dụng thực hiện phân loại khách hàng tốt hơn, phản ánh thực chất hơn tình trạng tín dụng của khách hàng.

4. Đối tượng và phạm vi nghiên cứu

- Dữ liệu khách hàng tại SHB.
- Ứng dụng khai phá dữ liệu vào việc đánh giá thông tin của khách hàng.
- Kho dữ liệu của ngân hàng SHB

5. Phương pháp nghiên cứu

- Nghiên cứu lý thuyết
- Thực nghiệm và phân tích kết quả

6. Cấu trúc của luận văn

Luận văn ngoài phần mở đầu và kết luận gồm 3 chương chính:

- Chương 1: Rủi ro tín dụng và quản lý rủi ro tín dụng tại ngân hàng

- Chương 2: Khai phá dữ liệu và bài toán phân lớp dự báo rủi ro tín dụng
- Chương 3: Thử nghiệm và đánh giá rủi ro tín dụng tại ngân hàng SHB

Trong đó, luận văn tập trung vào chương 2 và chương 3 với mục đích nghiên cứu khai phá dữ liệu trong bài toán phân lớp dự báo rủi ro tín dụng, sau đó thực nghiệm nhằm đánh giá mô hình này. Mặc dù có nhiều cố gắng nhưng do thời gian có hạn. Luận văn chắc chắn còn nhưng hạn chết khiếm khuyết. Kính mong các thầy cô và đồng nghiệp thông cảm và góp ý.

CHƯƠNG 1: RỦI RO TÍN DỤNG VÀ QUẢN LÝ RỦI RO TÍN DỤNG TẠI NGÂN HÀNG

Để có thể ứng dụng công nghệ khai phá dữ liệu và quản lý rủi ro tín dụng của ngân hàng, trước hết chúng ta cần phải rõ các khái niệm trong hoạt động tín dụng, phân loại tín dụng,xem xét đánh giá nguyên nhân dẫn đến rủi ro tín dụng, phương pháp quản lý của các ngân hàng nói chung và đặc biệt là ngân hàng SHB nơi tác giả đang công tác.

1.1. Hoạt động tín dụng

1.1.1. Tín dụng ngân hàng là gì?

Tín dụng ngân hàng là một giao dịch vay mượn tài sản giữa ngân hàng (bên cho vay) và khách hàng (bên đi vay), trong đó bên đi vay được sử dụng tài sản của bên cho vay trong một khoảng thời gian được thỏa thuận trước và phải hoàn trả vô điều kiện vốn gốc và lãi cho bên cho vay khi đến hạn thanh toán. Nói một cách khác, tín dụng ngân hàng là quan hệ chuyển nhượng quyền sử dụng vốn giữa ngân hàng và khách hàng trong một thời hạn nhất định với một khoản chi phí nhất định.

1.1.2. Bản chất của tín dụng

Bản chất của tín dụng là một giao dịch về tài sản trên cơ sở hoàn trả và có các đặc trưng sau:

- Tài sản giao dịch trong quan hệ tín dụng ngân hàng bao gồm hai hình thức là cho vay (bằng tiền) và cho thuê (bất động sản và động sản).
- Xuất phát từ nguyên tắc hoàn trả, vì vậy người cho vay khi chuyển giao tài sản cho người đi vay sử dụng phải có cơ sở để tin rằng người đi vay sẽ trả đúng hạn.
- Giá trị hoàn trả thông thường phải lớn hơn giá trị lúc cho vay, hay nói cách khác là người đi vay phải trả thêm phần lãi ngoài vốn gốc.
- Trong quan hệ tín dụng ngân hàng, tiền vay được cấp trên cơ sở bên đi vay cam kết hoàn trả vô điều kiện cho bên cho vay khi đến hạn thanh toán.

1.1.3. Vai trò của tín dụng

Thứ nhất: Đáp ứng nhu cầu vốn để duy trì quá trình sản xuất được liên tục đồng thời góp phần đầu tư phát triển kinh tế.

Thứ hai: Thúc đẩy quá trình tập trung vốn và tập trung sản xuất.

Thứ ba: Tín dụng là công cụ tài trợ cho các ngành kinh tế kém phát triển và ngành kinh tế mũi nhọn.

Thứ tư: Góp phần tác động đến việc tăng cường chế độ hạch toán kinh tế của các doanh nghiệp.

Thứ năm: Tạo điều kiện để phát triển các quan hệ kinh tế với nước ngoài.

1.1.4. Chức năng của tín dụng

Chức năng của tín dụng bao gồm 3 chức năng chính như sau:

- Phân phối lại nguồn vốn nhàn rỗi trên nguyên tắc hoàn trả lại cả gốc cả lãi
- Tạo điều kiện và lưu thông giá trị góp phần tiết kiệ được tiền mặt và chi phí lưu thông xã hội
- Kiểm soát đồng tiền với mọi hoạt động của kinh tế

1.2. Phân loại tín dụng trong ngân hàng

Công tác phân loại tín dụng dựa trên một số tiêu thức nhất định tùy theo yêu cầu của khách hàng và mục tiêu quản lý của ngân hàng. Có thể phân loại tín dụng trong ngân hàng theo nhiều cách như căn cứ vào thời hạn tín dụng, phân loại căn cứ theo đối tượng tín dụng, mục đích sử dụng vốn, căn cứ vào đối tượng trả nợ…Tuy nhiên do khuôn khổ luận văn tập trung vào phần dự báo rủi ro tín dụng nên luận văn chỉ dưa ra cách phân loại tín dụng dựa vào rủi ro, cách phân loại này giúp ngân hàng thường xuyên đánh giá lại

tính an toàn của các khoản tín dụng, trích lập dự phòng tổn thất kịp thời, được phân loại thành 5 nhóm[8]:

- **Nhóm 1:** Nợ đủ tiêu chuẩn, Các khoản nợ trong hạn mà tổ chức tín dụng đánh giá là có đủ khả năng thu hồi đầy đủ cả gốc và lãi đúng thời hạn.
- **Nhóm 2:** Nợ cần chú ý, bao gồm nợ quá hạn dưới 90 ngày và nợ cơ cấu lại thời hạn trả nợ.
- **Nhóm 3:** Nợ dưới tiêu chuẩn, bao gồm nợ quá hạn từ 90 ngày đến 180 ngày và nợ cơ cấu lại thời hạn trả nợ quá hạn dưới 90 ngày.
- **Nhóm 4:** Nợ nghi ngờ, bao gồm nợ quá hạn từ 181 ngày đến 360 ngày và nợ cơ cấu lại thời hạn trả nợ quá hạn từ 90 ngày đến 180 ngày.
- **Nhóm 5:** Nợ có khả năng mất vốn, gồm nợ quá hạn trên 360 ngày, nợ cơ cấu lại thời hạn trả nợ trên 180 ngày và nợ khoanh chờ Chính phủ xử lý.

1.3. Rủi ro tín dụng

Đây là rủi ro lớn nhất và thường xuyên xảy ra, có thể khiến ngân hàng rơi vào trạng thái tài chính khó khăn nghiêm trọng. "Rủi ro tín dụng trong hoạt động ngân hàng của tổ chức tín dụng là khả năng xảy ra tổn thất trong hoạt động ngân hàng của tổ chức tín dụng do khách hàng không thực hiện hoặc không có khả năng thực hiện nghĩa vụ của mình theo cam kết." [8].

1.3.1. Rủi ro tín dụng và nguyên nhân

a. Rủi ro tín dụng

b. Nguyên nhân chủ yếu dẫn đến rủi ro tín dụng

1.3.2. Các ảnh hưởng của rủi ro tín dụng đến hoạt động của ngân hàng

Các ảnh hưởng của rủi ro tín dụng đến hoạt động của ngân hàng là:

a. Đối với nền kinh tế

b. Đối với ngân hàng

c. Đối với khách hàng

1.4. Đánh giá phương pháp quản lý rủi ro tín dụng tại ngân hàng SHB hiện nay

- Bước đầu thì SHB đã thiết lập được hệ thống đánh giá xếp hạng tín dụng CSS nhằm giúp cán bộ quản lý tín dụng cũng như ban điều hành trong việc quản lý vận hành hoạt động tín dụng tại ngân hàng. Nhưng nó mới chỉ dừng ở mức thu thập thông tin liên quan về khách hàng vay vốn và tính điểm và xếp hạng theo một mô hình xếp hạng sẵn và xếp hạng khách hàng theo số điểm tính được một cách cứng nhắc. Việc đánh giá kết quả từ hệ thống vẫn dựa vào kinh nghiệm và trình độ đánh giá và phân tích của cán bộ tín dụng vì vậy trong thực tế chưa sát với thực tế của khách hàng. Ví dụ với các khách hàng đã được xếp hạng đôi khi được xếp hạng AAA, AA… (hạng cao nhất trong thang xếp hạng) thì việc trả nợ lại gặp khó khăn hoặc mặc dù có khách hàng điểm xếp hạng thấp nhưng lại trả nợ rất đúng hạn. Chính vì vậy việc khai thác triệt để những thông tin thu thập được từ khách hàng và dữ liệu thực tế thì hệ thống chưa đáp ứng được. Chính vì lý do đó mà việc áp dụng khai phá dữ liệu để thu được những thông tin hữu ích trong việc quản trị rủi ro và hỗ trợ việc ra quyết định là cần thiết.

1.5. Kết luận Chương 1

Căn cứ vào tình hình thực tế tại các ngân hàng Việt Nam nói chung và ngân hàng SHB nói riêng thì ngoài các phân tích về mặt nghiệp vụ cùng với các hệ thống đánh giá xếp hạng tín dụng thì cần tiếp tục nghiên cứu các giải pháp nhằm dự báo rủi ro tín dụng một cách hiệu quả hơn.Trong chương tiếp theo luận văn sẽ trình bày phương pháp khai phá dữ liệu nhằm quản lý rủi ro tín dụng ngân hàng.

CHƯƠNG 2: KHAI PHÁ DỮ LIỆU VÀ BÀI TOÁN PHÂN LỚP DỰ BÁO RỦI RO TÍN DỤNG

2.1. Tổng quan về khai phá dữ liệu

2.1.1. Khai phá dữ liệu là gì và tại sao phải khai phá dữ liệu

a. Khai phá dữ liệu là gì

Định nghĩa: Khai phá dữ liệu là một tập hợp các kỹ thuật được sử dụng để tự động khai thác và tìm ra các mối quan hệ lẫn nhau của dữ liệu trong một tập hợp dữ liệu khổng lồ và phức tạp, đồng thời cũng tìm ra các mẫu tiềm ẩn trong tập dữ liệu đó

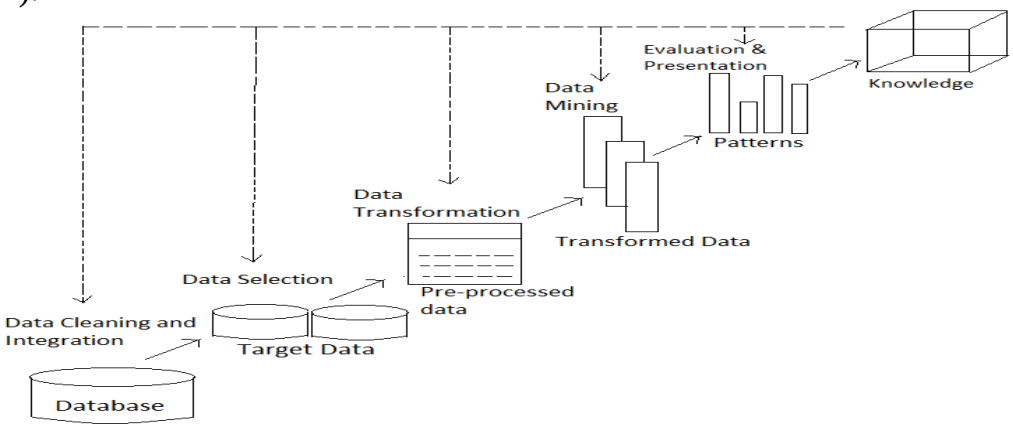
Khai phá dữ liệu được dùng để mô tả quá trình phát hiện ra tri thức trong CSDL. Quá trình này kết xuất ra các tri thức tiềm ẩn từ dữ liệu giúp cho việc dự báo trong kinh doanh, các hoạt động sản xuất,... Khai phá dữ liệu làm giảm chi phí về thời gian so với phương pháp truyền thống trước kia (ví dụ như phương pháp thống kê). Có nhiều thuật ngữ được dùng tương tự như Datamining như Knowledge Mining (khai phá tri thức), knowledge extraction (chắt lọc tri thức), data/parttern analysis (phân tích dữ liệu/mẫu), data archaeology (khảo cổ dữ liệu), data dredging (nạo vét dữ liệu) [9],…

b. Tại sao phải tiến hành khai phá dữ liệu trong các dịch vụ tài chính

Trong ngành công nghiệp dịch vụ tài chính trên toàn thế giới, phương thức liên lạc truyền thống của khách hàng mặt đối mặt (face-to-face) đang được thay thế bằng phương thức điện tử để giảm thời gian và chi phí xử lý các áp dụng cho sản phẩm khác nhau, và cuối cùng là cải thiện hiệu quả của việc sử dụng tài chính. Tin học hoá quá trình hoạt động tài chính, sử dụng internet và phần mềm tự động hoàn toàn có thể làm thay đổi các khái niệm cơ bản của kinh doanh và cách hoạt động kinh doanh đang được thực hiện. Hiển nhiên, lĩnh vực ngân hàng không phải là một ngoại lệ. Kể từ những năm 1990 toàn bộ khái niệm ngân hàng đã được chuyển sang cơ sở dữ liệu tập trung, giao dịch trực tuyến và máy ATM được thực hiện trên thế giới, đã làm cho hệ thống ngân hàng mặt mạnh mẽ hơn về mặt kỹ thuật và định hướng khách hàng tốt hơn. Dữ liệu có thể là một trong những nguồn tài nguyên có giá trị nhất của bất kỳ ngân hàng nào, tuy nhiên nó chỉ thực sự có giá trị khi nó biết cách tiếp cận với thông tin có giá trị ẩn chứa trong dữ liệu thô. Khai phá dữ liệu cho phép triết suất các thông tin từ các dữ liệu lịch sử, và dự đoán kết quả các tình huống trong tương lai. Nó giúp cho việc tối ưu hóa các quyết định kinh doanh, tăng giá trị của từng khách hàng và thông tin kết nối, đồng thời cải thiện sự hài lòng của khách hàng.

2.1.2. Quy trình và các bước khai phá dữ liệu

Khai phá dữ liệu là một bước trong bảy bước của quá trình KDD (Knowleadge Discovery in Database) và KDD được xem như 7 quá trình khác nhau theo thứ tự sau (Hình 2.1):



Hình 2.1: Các bước khai phá dữ liệu

- **Làm sạch dữ liệu (data cleaning**
- **Tích hợp dữ liệu (data intergation**
- **Lựa chọn dữ liệu (data selection)**
- **Chuyển đổi dữ liệu (data tranform)**
- **Khai phá dữ liệu (data mining)**
- **Đánh giá mẫu (pattern evaluation)**
- **Biểu diễn tri thức (Knowledge presentation)**

Từ những bước cơ bản trong khai phá dữ liệu, kiến trúc mẫu của một hệ thống khai phá dữ liệu có thể có những thành phần chính sau:

Hình 2.2: Các thành phần trong hệ thống Data Mining

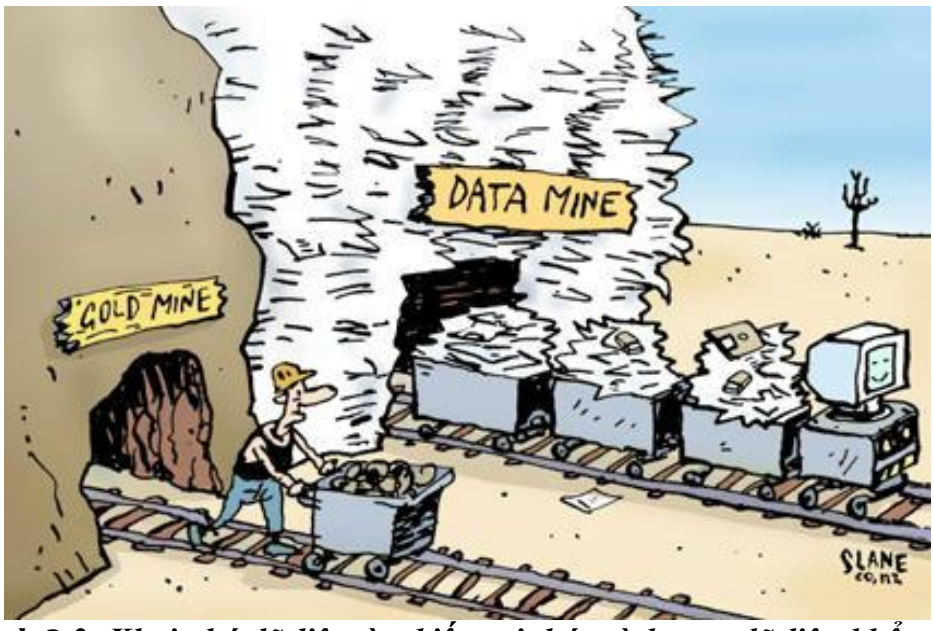
2.1.3. Các phương pháp khai phá dữ liệu

Các các phương pháp KPDL có thể được phân chia theo chức năng hay lớp các bài toán khác nhau. Sau đây là một số phương pháp phổ biến:

- **Phân lớp và dự đoán (classification & prediction)**
- **Luật kết hợp (association rules**
- **Khai phá chuỗi theo thời gian (sequential/temporal patterns)**
- **Phân cụm (clustering/segmentation)**
- **Mô tả khái niệm (concept description & summarization)**

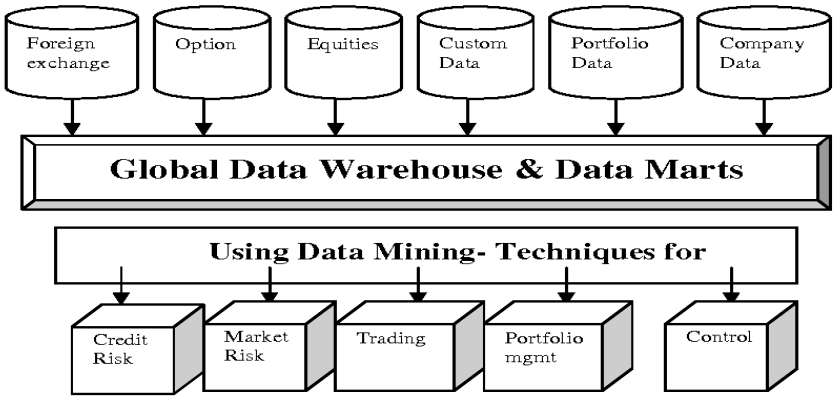
2.2. Ứng dụng của khai phá dữ liệu trong hệ thống thông tin ngân hàng

Hiện tại, các ngân hàng và tổ chức tài chính trên khắp thế giới đang phải duy trì những kho dữ liệu khổng lồ với nhiều thông tin có giá trị. Quy mô khổng lồ của các kho dữ liệu này gây khó khăn cho con người trong việc phân tích để đưa ra những thông tin hữu ích trong quá trình ra quyết định. Nhiều công ty thương mại đã nhanh nhạy nắm bắt được tình hình này, nhờ đó đã tạo nên một thị trường phần mềm về khai phá dữ liệu (data mining) rất phong phú nổi bật lên là các sản phẩm khai phá dữ liệu của Oracle (Oracle Dataminer), IBM, SAP…



Hình 2.3: Khai phá dữ liệu tìm kiếm tri thức từ lượng dữ liệu khổng lồ

Sự cạnh tranh toàn cầu, thị trường năng động và những chu kỳ đổi mới công nghệ càng ngày càng được rút ngắn đã tạo ra nhiều thách thức quan trọng cho ngành tài chính và ngân hàng. Việc có mặt nhanh chóng của thông tin ở phạm vi toàn cầu giúp làm tăng sự linh hoạt của các doanh nghiệp. Sự phát triển nhanh chóng về công nghệ thông tin trong các tổ chức tài chính đã tạo ra những nhu cầu lớn về việc liên tục phân tích dữ liệu.



Hình 2.4: Ứng dụng data mining trong ngân hàng [14]

Data mining góp phần giải quyết các vấn đề kinh doanh trong nganh ngân hàng và tài chính bằng cách tìm ra các dạng mẫu (patterns), nguyên nhân và mối tương quan trong các thông tin kinh tế, giá cả thị trường mà các nhà quản lý không thể dễ dàng nhận ra do khối lượng dữ liệu quá lớn hoặc xuất hiện quá nhanh. Cấp quản lý của các ngân hàng có thể tìm hiểu thêm về giai đoạn, chu kỳ của các diễn biến giao dịch của khách hàng nhằm phân khúc, xác đinh mục tiêu, thu hút và giữ nguồn khách hàng mang lại lợi nhuận. Business Intelligence và Data mining còn có thể giúp nhận diện các tầng lớp khách hàng khác nhau, để đưa ra các giải pháp về sản phẩm và giá cả phù hợp cho từng lớp khách hàng, góp phần tăng hiệu quả trong kinh doanh. Đó là:

- *Quản trị rủi ro*
- *Phát hiện gian lận*
- *Quản lý danh mục vốn*
- *Quảng cáo và chăm sóc khách hàng*

2.3. Bài toán phân lớp dự báo rủi ro tín dụng

Như đã trình bày ở chương 1, hoạt động tín dụng trong ngành ngân hàng là đặc biệt quan trọng vì vậy việc đánh giá và phân loại rủi ro là nhiệm vụ hàng đầu trong quản trị vận hành ngân hàng. Chính vì thế hiện nay hầu hết các ngân hàng trên thế giới nói chung và Việt Nam nói riêng đều có những hệ thống hỗ trợ việc đánh giá và phân loại rủi ro. Hầu hết các ngân hàng hiện nay đều sử dụng mô hình chấm điểm tín dụng để hỗ trợ đánh giá rủi ro và xếp hạng tín dụng từ đó có quyết định cho khách hàng vay hay không. Các đặc điểm về cấu trúc, thiết kế và vận hành của hệ thống xếp hạng tín dụng có thể khác nhau giữa các ngân hàng, ví dụ như: cơ cấu của các chỉ tiêu đánh giá, trọng số của các chỉ tiêu, số lượng các mức xếp hạng, ước tính mức rủi ro gắn liền với các mức xếp hạng, các chính sách khách hàng, chính sách tín dụng áp dụng cho từng mức xếp hạng. Nhưng nhìn chung thì cách tiếp cận chung là đều sử dụng các thông tin khách hàng cung cấp để đưa ra một giá trị điểm từ đó ứng với từng thang điểm mỗi khoản vay sẽ được xếp hạng theo từng thang điểm. Có thể thấy rằng đây là một mô hình khá phổ biến đang được thực hiện tại các NHTM Việt Nam, bởi lẽ mô hình này có nhiều lợi thế và khá phù hợp với các NHTM trong điều kiện Việt Nam hiện nay, cụ thể là:

- Tận dụng được kinh nghiệm và kiến thức chuyên sâu của các cán bộ tín dụng, các chuyên gia tài chính để phân tích các chỉ tiêu tài chính. Việc phân tích dựa trên công nghệ giản đơn, hệ thống lưu trữ thông tin ổn định, sử dụng hồ sơ sẵn có, dễ dàng thu thập thông tin
- Đây là mô hình tương đối đơn giản, song hạn chế của mô hình này là nó phụ thuộc vào trình độ phân tích, đánh giá của cán bộ tín dụng.
- Mô hình này có thể áp dụng cho các khoản vay riêng lẻ, mang tính đặc thù chịu ảnh hưởng các yếu tố vùng miền, phong tục, tập quán thì việc dựa trên các yếu tố định lượng, không đưa ra được quyết định chính xác mà phải dựa trên ý kiến và kinh nghiệm của cán bộ tín dụng.
- Các NHTM sử dụng mô hình này sẽ chịu chi phí cao do tốn nhiều thời gian để đánh giá và đòi hỏi cán bộ tín dụng phải có tính chuyên nghiệp, có thâm niên, kỹ năng.
- Mô hình này rất khó khăn đo lường vai trò của các yếu tố đến hạng tín nhiệm của khách hàng
- Đặc biệt là mô hình chấm điểm này chưa có khả năng dự báo được rủi ro mà mới chỉ đánh giá được phần nào rủi ro nhờ điểm xếp hạng

Chính vì những hạn chế của mô hình chấm điểm xếp hạng tín dụng hiện tại tôi xin đề xuất phương pháp áp dụng thuật toán phân lớp trong khai phá dữ liệu để dự báo khả năng hoàn vốn của các khách hàng dựa vào các thông tin sử dụng trong mô hình chấm điểm và dữ liệu lịch sử của các khách hàng đã vay vốn tại ngân hàng.

2.3.1. Phát biểu bài toán

Đầu vào:

- *Tập thông tin khách hàng và lịch sử trả nợ của các khách hàng nhằm mục đích xây dựng mô hình (tập training)*
- *Tập thông tin khách hàng và lịch sử trả nợ nhằm mục đích kiểm chứng mô hình (tập dữ liệu test)*
- *Tập thông tin khách hàng mới cần dự báo*

Đầu ra:

Đưa ra mô hình phân lớp dự báo, các chỉ số đánh giá mô hình, các luật rút ra từ mô hình giúp phân loại các khách hàng mới.

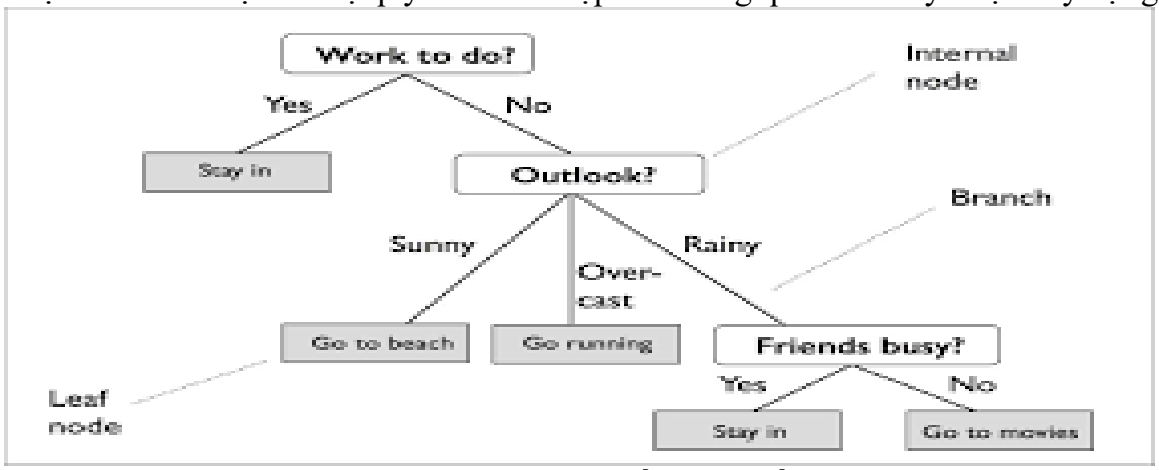
Ví dụ:

Đầu vào:
Thông tin khách hàng về khách hàng vay vốn: Mục đích vay *mua nhà*, có thu nhập *trên 10 triệu*, đang *ở cùng với bố mẹ*, làm tại *công ty cổ phần*, chức vụ *chuyên viên*, thời gian công tác trong lĩnh vực chuyên môn *dưới 3 năm*
Đầu ra: Dự báo khách hàng có khả năng rơi vào nhóm nợ cần chú ý (Nhóm nợ 2).
Mục tiêu của bài toán là phân loại khách hàng theo khả năng hoàn vốn dựa vào các thông tin đầu vào ban đầu khách hàng phải cung cấp từ đó dự báo được khách hàng này là khách hàng mục tiêu hay không. Việc dự báo chính xác sẽ giúp ngân hàng giảm thiểu các rủi ro có thể có từ các khách hàng có khả năng không trả được nợ.

2.3.2. Phân lớp sử dụng cây quyết định

a. Cây quyết định

Cuối những năm 70 đầu những năm 80, J. Ross Quinlan đã phát triển một thuật toán sinh cây quyết định. Đây là một tiếp cận tham lam, trong đó nó xác định một cây quyết dịnh được xây dựng từ trên xuống một cách đệ quy theo hướng chia để trị. Hầu hết các thuật toán sinh cây quyết định đều dựa trên tiếp cận top-down trình bày sau đây, trong đó nó bắt đầu từ một tập các bộ huấn luyện và các nhãn phân lớp của chúng. Tập huấn luyện được chia nhỏ một các đệ quy thành các tập con trong quá trình cây được xây dựng [15].



Hình 2.5: Ví dụ về cây quyết định

b. Ưu nhược diểm của cây quyết định

- **Ưu điểm**

Cây quyết định tương đối dễ hiểu
Đòi hỏi tiền xử lý dữ liệu đơn giản
Khả năng xử lý cả thuộc tính liên tục và rời rạc
Thể hiện rõ ràng những thuộc tính tốt nhất
Dễ dàng tính toán trong khi phân lớp

- **Nhược điểm**

Dễ xảy ra lỗi khi có quá nhiều lớp
Chi phí tính toán đắt để đào tạo

c. Quá trình xây dựng cây quyết định

Quá trình xây dựng cây quyết định gồm hai giai đoạn:

- **Giai đoạn thứ nhất phát triển cây quyết định**
- **Giai đoạn thứ hai cắt, tỉa bớt các cành nhánh trên cây quyết định**.

d. Thuật toán cây quyết định

Giải thuật cơ bản (giải thuật tham lam) được chia thành các bước như sau:

- Phát triển cây quyết định

- Chọn thuộc tính "tốt" nhất bằng một độ đo đã định trước
- Phát triển cây bằng việc thêm các nhánh tương ứng với từng giá trị của thuộc tính đã chọn
- Sắp xếp, phân chia tập dữ liệu đào tạo tới node con
- Nếu các ví dụ được phân lớp rõ ràng thì dừng.
- Ngược lại: lặp lại bước 1 tới bước 4 cho từng node con
- Cắt tỉa cây: nhằm đơn giản hóa, khái quát hóa cây, tăng độ chính xác

Điều kiện để dừng việc phân chia:

- Tất cả những mẫu huấn luyện đối với một nút cho trước thuộc về cùng một lớp.
- Không còn thuộc tính còn lại nào để phân chia tiếp.
- Không còn mẫu nào còn lại.

Trên cơ sở giải thuật cơ bản như đã nêu trên, đã có nhiều nghiên cứu để xây dựng cây quyết định mà nổi bật là các thuật toán CART, ID3, C4.5 [15]. Các thuật toán này chấp nhận sự tham lam (greedy) cách tiếp cận cây quyết định được xây dựng từ trên xuống một cách đệ quy, bắt đầu với một bộ dữ liệu huấn luyện tập và các nhãn lớp của họ. Hầu hết giải thuật cây quyết định đều theo cách tiếp cận từ trên xuống. Tập dữ liệu huấn luyện được phân vùng một cách đệ quy thành tập hợp con nhỏ hơn trong lúc cây được xây dựng.

Điểm khác biệt chính giữa các thuật toán này chính là tiêu chuẩn (hay còn gọi là thuộc tính phân chia) và độ đo để chọn lựa.

Có 3 loại tiêu chuẩn hay chỉ số để xác định thuộc tính tốt nhất phát triển tại mỗi node:

- **Gini-index** [15]
- ***Information–gain*** [15]
- χ^2 ***-bảng thống kê các sự kiện xảy ra ngẫu nhiên***

e. Thuật toán C4.5

C4.5 là sự kế thừa của của thuật toán học máy bằng cây quyết định dựa trên nền tảng là kết quả nghiên cứu của HUNT và các cộng sự của ông trong nửa cuối thập kỷ 50 và nửa đầu những năm 60 (Hunt 1962). Phiên bản đầu tiên ra đời là ID3 (Quinlan, 1979)- 1 hệ thống đơn giản ban đầu chứa khoảng 600 dòng lệnh Pascal, và tiếp theo là C4 (Quinlan 1987). Năm 1993, J. Ross Quinlan đã kế thừa các kết quả đó phát triển thành C4.5 với 9000 dòng lệnh C chứa trong một đĩa mềm. Mặc dù đã có phiên bản phát triển từ C4.5 là C5.0 - một hệ thống tạo ra lợi nhuận từ Rule Quest Research, nhưng nhiều tranh luận, nghiên cứu vẫn tập trung vào C4.5 vì mã nguồn của nó là sẵn dùng.

Tư tưởng phát triển cây quyết định của C4.5 là phương pháp. Chiến lược phát triển theo độ sâu (depth-first strategy) được áp dụng cho C4.5.

f. Chọn thuộc tính tốt nhất

Quinlan (1983) là người đầu tiên đề xuất việc sử dụng lý thuyết thông tin để tạo ra các cây quyết định và công trình của ông là cơ sở cho phần trình bày ở đây. Lý thuyết thông tin của Claude Shannon (1948) cung cấp khái niệm entropy để đo tính thuần nhất (hay ngược lại là độ pha trộn) của một tập hợp [9]. Một tập hợp là thuần nhất nếu như tất cả các phần tử của tập hợp đều thuộc cùng một loại, và khi đó ta nói tập hợp này có độ pha trộn là thấp nhất. Trong trường hợp của tập ví dụ, thì tập ví dụ được gọi là thuần nhất nếu như tất cả các ví dụ đều có cùng giá trị phân loại..

- ***Entropy đo tính thuần nhất của tập ví dụ***

Khái niệm entropy của một tập S được định nghĩa trong lý thuyết thông tin là số lượng mong đợi các bit cần thiết để mã hóa thông tin về lớp của một thành viên rút ra một cách

ngẫu nhiên từ tập S. Trong trường hợp tối ưu, mã có độ dài ngắn nhất. Theo lý thuyết thông tin, mã có độ dài tối ưu là mã gán –log2p bits cho thông điệp có xác suất là p [9]. Trong trường hợp S là tập ví dụ, thì thành viên của S là một ví dụ, mỗi ví dụ thuộc một lớp hay có một giá trị phân loại.

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

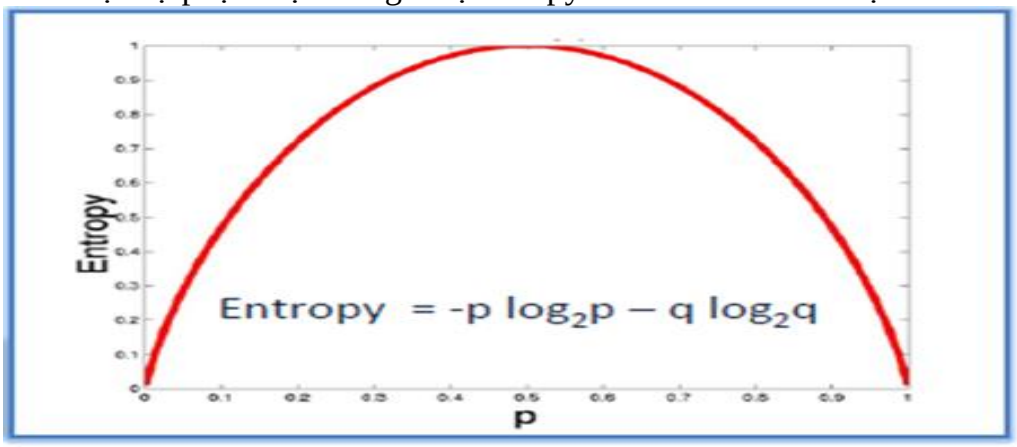
✓ Entropy(S) = 0: tập ví dụ S chỉ toàn ví dụ thuộc cùng một loại, hay S là thuần nhất.

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

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

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

Hình sau minh họa sự phụ thuộc của giá trị entropy vào xác suất xuất hiện của ví dụ dương:



Hình 2.6: Sự phụ thuộc của Entropy

Cho trước:

Tập S là tập dữ liệu huấn luyện, trong đó thuộc tính phân loại có hai giá trị, giả sử là âm (-) và dương (+). Trong đó:

p+ là xác suất các ví dụ dương trong tập S.

p_ là xác suất các ví dụ âm trong tập S.

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

$$Entropy(S) = -p_+ \log_2 p_+ - p_- \log_2 p_-$$

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

$$Entropy(S) \equiv \sum_{i=1}^{c} -pi \log 2\, pi$$

- ***Information gain***

Entropy là một số đo đo độ pha trộn của một tập ví dụ, bây giờ chúng ta sẽ định nghĩa một phép đo hiệu suất phân loại các ví dụ của một thuộc tính. Phép đo này gọi là lượng thông tin thu được (hay độ lợi thông tin), nó đơn giản là lượng giảm entropy mong đợi gây ra bởi việc phân chia các ví dụ theo thuộc tính này.

Một cách chính xác hơn, Gain (S, A) của thuộc tính A, trên tập S, được định nghĩa như sau:

$$Gain\ (S, A) = Entropy(S) - \sum \frac{|S_v|}{} Entropy(S_v)$$

$$\sum_{V \in Value(A)} |S|$$

Giá trị Value (A) là tập các giá trị có thể cho thuộc tính A, và Sv là tập con của S mà A nhận giá trị v.

- ***Tỷ suất lợi ích Gain Ratio***

Khái niệm độ lợi thông tin Gain có xu hướng ưu tiên các thuộc tính có số lượng lớn các giá trị. Nếu thuộc tính D có giá trị riêng biệt cho mỗi bản ghi, thì Entropy (S, D) = 0, như vậy Gain (S, D) sẽ đạt giá trị cực đại. Rõ ràng, một phân vùng như vậy thì việc phân loại là vô ích.

Thuật toán C4.5, một cải tiến của ID3, mở rộng cách tính Information Gain thành Gain Ratio để cố gắng khắc phục sự thiên lệch.

Gain Ratio được xác định bởi công thức sau:

$$\text{GainRatio (S, A)} = \frac{\text{Gain (S, A)}}{\text{SplitInformation(S, A)}}$$

Trong đó, SplitInformation (S, A) chính là thông tin do phân tách của A trên cơ sở giá trị của thuộc tính phân loại S. Công thức tính như sau:

$$\text{SplitInformation(S, A)} = -\sum_{i=1}^{c} \frac{|S_i|}{|S|} \log_2 \frac{|S_i|}{|S|}$$

g. Chuyển cây quyết định sang luật dễ hiểu

Thông thường, cây quyết định sẽ được chuyển về dạng các luật để thuận tiện cho việc cài đặt và sử dụng. Tuy nhiên việc tạo ra tập luật từ tập dữ liệu lớn và nhiều giá trị sai là vô cùng lớn. Vì vậy trong quá trình chuyển đổi từ cây quyết định sang luật cần phải cắt tỉa để thu được tập luật tối ưu

Việc chuyển đổi từ cây sang tập luật được thực hiện qua 4 bước

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

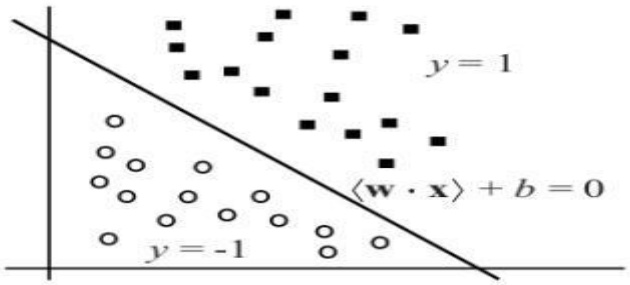
2.3.3. Phân lớp sử dụng SVM – Máy véctơ hỗ trợ

SVM –Support vector machine là một mô hình học có giám sát trong lĩnh vực học máy, SVM thường được dùng trong phân lớp dữ liệu (classification) và phân tích hồi quy (regression analysis). SVM là nền tảng cho nhiều thuật toán khai phá dữ liệu, SVM được giới thiệu bởi Vladimir Vapnik và các đồng sự vào năm 1995 [10]. Ý tưởng chính của SVM là phân chia dữ liệu bằng các siêu phẳng (hyperlane). Từ ý tưởng chính nhiều phương pháp cải tiến được tùy biên từ phương pháp nguyên thủy cho nhiều cách sử dụng khác nhau

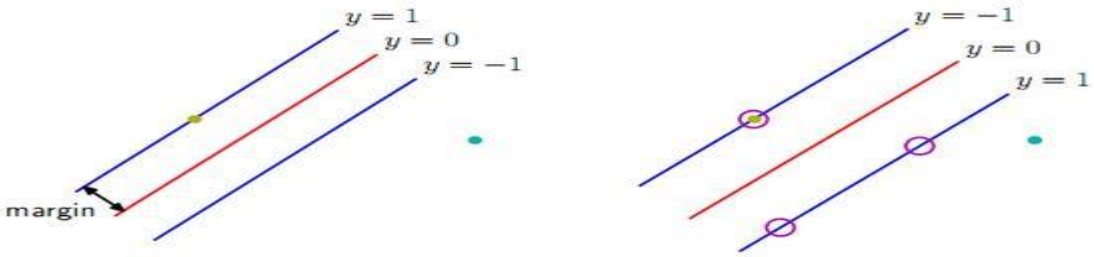
Xét bài toán phân lớp đơn giản nhất – phân lớp hai phân lớp với tập dữ liệu mẫu: $\{xi, yi|i = 1,2,3\ldots N|xi \in Rm\}$

Trong đó mẫu là các vector đối tượng được phân lớp thành các mẫu dương và mẫu âm:

- Các mẫu dương là các mẫu xi thuộc lĩnh vực quan tâm và được gán nhãn yi=1.
- Các mẫu âm là các mẫu xi không thuộc lĩnh vực quan tâm và được gán nhãn yi=-1.



Hình 2.7: Siêu phẳng phân tách



Hình 2.8: Khoảng cách từ siêu phẳng đến điểm gần siêu phẳng nhất

Các mặt siêu phẳng trong không gian đối tượng có phương trình là:

$$f(x) = wx + b = 0$$

Trong đó w là vector trọng số, b là độ dịch. Khi thay đổi w và b thì hướng và khoảng cách từ gốc tọa độ đến mặt siêu phẳng thay đổi.

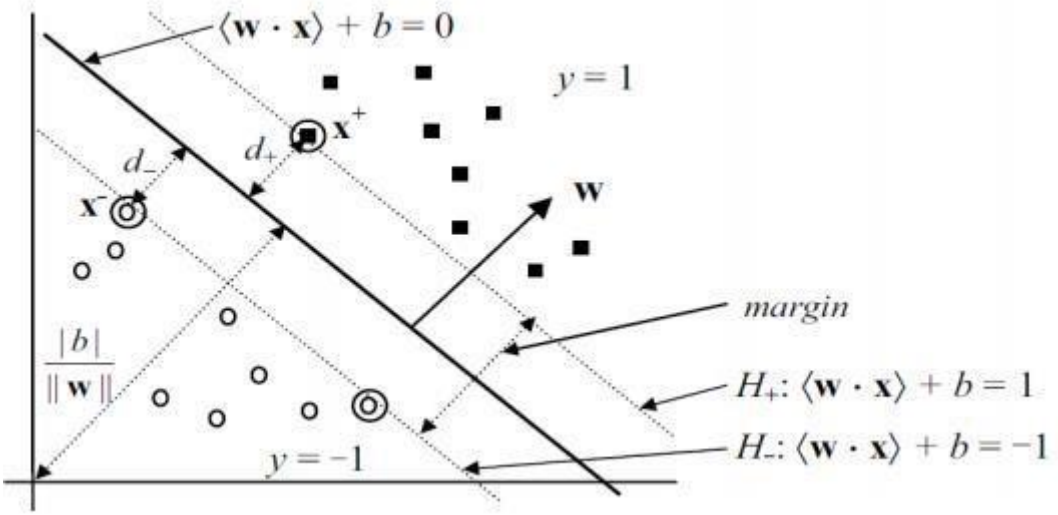
Bộ phân lớp nhị phân được xác định thông qua dấu của f(x):

$$yi = \begin{cases} -1, nếu\ f(xi) < 0 \\ 1, nếu\ f(xi) \geq 0 \end{cases}$$

- Nếu yi = 1 thì xi thuộc vào lớp dương
- Nếu yi = -1 thì xi thuộc vào lớp âm

Học máy SVM là một họ các mặt siêu phẳng phụ thuộc vào các tham số w, b. Mục tiêu của SVM là ước lượng w, b để cực đại lề hóa giữa lớp dương và lớp âm. Các giá trị của lề cho chúng ta các mặt siêu phẳng khác nhau

a. Phân lớp tuyến tính



Hình 2.9: Tập dữ liệu có thể tách tuyến tính

Bộ phân lớp tìm ra mặt siêu phẳng với lề cực đại được xác định bởi khoảng cách giữa các mẫu âm và mẫu dương gần mặt siêu phẳng nhất

Gọi d+ và d- là khoảng cách ngắn nhất từ siêu phẳng đến điểm dữ liệu dương và âm gần nhất. Khi đó lề siêu phẳng là margin = $d_+ + d_-$

Giả sử 2 điểm (x^+, 1) và (x^-, -1) là điểm gần siêu phẳng nhất. Khi đó chúng ta xác định được hai đường song song H_- và H_+. Thay đổi tỷ lệ w, b ta được:

$$H_+: wx^+ + b = 1$$

$$H_-: wx^- + b = -1$$

Các ràng buộc:

$$wx_i + b \geq 1 \; với \; nếu \; y_i = 1$$

$$wx_i + b \leq -1 \; với \; nếu \; y_i = -1$$

Không có dữ liệu huấn luyện nào nằm giữa H_+ và H

Gọi x_s là một điểm thuộc mặt siêu phẳng và d_+ là khoảng cách từ H_+ tới mặt siêu phẳng. Khi đó wx_s+ b = 0. Do vậy, ta có công thức sau:

$$d_+ = \frac{|wx_s + b - 1|}{||w||} = \frac{1}{||w||}$$

Trong đó ||w|| là độ dài vector w:

$$||w|| = \sqrt{w_1^2 + w_2^2 + \cdots + w_n^2}$$

Do vậy lề siêu phẳng được tính như sau:

$$margin = d_+ + d_- = \frac{2}{||w||}$$

Vấn đề cực đại lề (margin) được chuyển thành bài toán cực tiểu $||w||^2$ với các điều kiện: C với i = 1, 2, 3, …n

Vector w sẽ được tính theo công thức:

$$W = \sum_{i=0}^{n} x_i \, y_i \propto_i$$

Để xác định độ dịch chuyển b ta chọn mẫu xi sao cho mọi $\propto_i$ >0 sau đó sử dụng điều kiện Karush–Kuhn–Tucker (KKT) như sau:

$$\propto_i [y_i(wx_i + b) - 1] = 0$$

Các mẫu x_i tương ứng $\propto_i$ là những mẫu nằm gần siêu phẳng và được gọi là vector hỗ trợ. Support vector chính là cái mà ta quan tâm trong quá trình huấn luyện của SVM. Việc phân lớp cho một điểm dữ liệu mới sẽ chỉ phụ thuộc vào các support vector.

b. Phân lớp phi tuyến tính

Trường hợp không tách được tuyến tính chúng ta có thể giải quyết theo 2 phương pháp.

- **Phương pháp cực đại hóa biên mềm:** Năm 1995, Corinna Cortes và Vladimir N. Vapnik đề xuất một ý tưởng mới cho phép thuật toán gán nhãn sai cho một số ví dụ luyện tập Nếu không tồn tại siêu phẳng nào phân tách được hai lớp dữ liệu, thì thuật toán *biên mềm* sẽ chọn một siêu phẳng phân tách các ví dụ luyện tập tốt nhất có thể, và đồng thời cực đại hóa khoảng cách giữa siêu phẳng với các ví dụ được gán đúng nhãn. Phương pháp này sử dụng các biến bù ξ_i dùng để đo độ sai lệch của ví dụ x_i:

$$y_i(wx_i + b) \geq 1 - \xi_i \geq 0$$

Hàm mục tiêu có thêm một số hạng mới để phạt thuật toán khi ξi khác không, và bài toán tối ưu hóa trở thành việc trao đổi giữa lề lớn và mức phạt nhỏ. Nếu hàm phạt là tuyến tính thì bài toán trở thành:

$$\min_{w,\xi,b} \{\frac{1}{2}||w||^2 + C \sum_{i=1}^{n} \xi_i\}$$

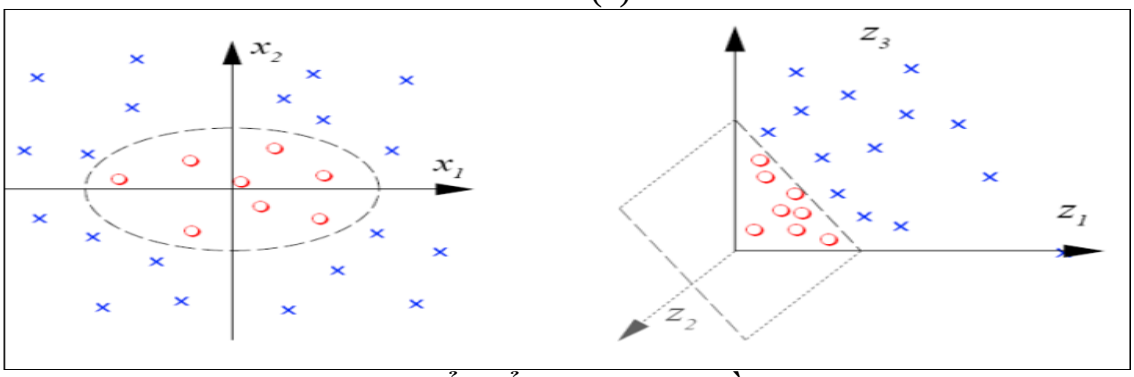
Với điều kiện:

$$y_i(wx_i + b) \geq 1 - \xi_i, \xi_i \geq 0$$

Phương pháp sử dụng thủ thuật hàm hạt nhân: Phương pháp này sử dụng một ánh xạ phi tuyến Φ để ánh xạ các điểm dữ liệu đầu vào từ không gian ban đầu sang một không gian

F mới có số chiều cao hơn. Trong không gian này các điểm dữ liệu có thể phân tách tuyến tính, hoặc có thể phân tách ít lỗi hơn so với không gian ban đầu. Siêu phẳng phân tách tuyến tính trong không gian mới sẽ tương ứng với mặt phân tách phi tuyến trong không gian ban đầu

$$\Phi: X \rightarrow F$$
$$X \rightarrow \Phi(x)$$



Hình 2.10: Chuyển đổi không gian bằng hàm nhân

Việc chuyển đổi sang không gian mới bằng cách sử dụng hàm nhân

Sau khi giải bài toán tuyến tính trong không gian đặc trưng ta có siêu phẳng phân lớp trong không gian đặc trưng. Dựa vào phương trình siêu phẳng ta xác định được các điểm support vector trong không gian đặc trưng. Sau đó ánh xạ các vector này về không gian ban đầu. Cuối cùng từ các support vector này ta xác định được đường phân lớp trong không gian ban đầu.

Các hàm nhân thường sử dụng:

Đa thức:

$$K(x, z) = (x.z + \theta)^d \text{ Trong đó } \theta \in N, d \in N$$

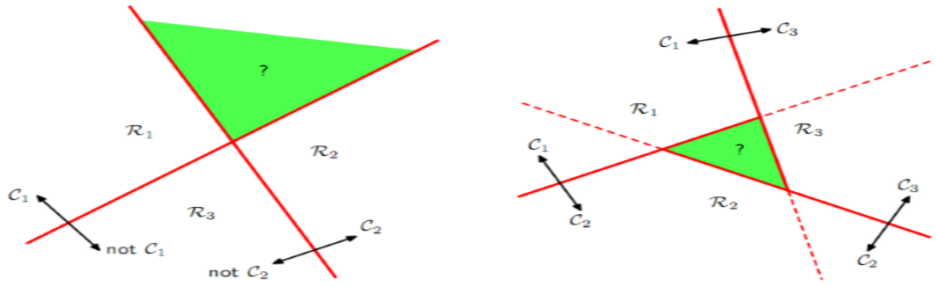
Gaussian RBF:

$K(x, z) = \exp(-\gamma|x - z|^2)$, với γ do người dùng định nghĩa.

Xích ma:

$K(x, z) = \tan(x.z + \theta)$

c. Phân đa lớp



Hình 2.11: Phân đa lớp

Bây giờ xét đến trường hợp phân nhiều lớp K > 2. Chúng ta có thể xây dựng việc phân K-class dựa trên việc kết hợp một số đường phân 2 lớp. Tuy nhiên, điều này sẽ dẫn đến một vài khó khăn (theo Duda and Hart, 1973).

Hướng one-versus-the-rest, ta sẽ dùng K-1 bộ phân lớp nhị phân để xây dựng Kclass.

Hướng one-versus-one, dùng K(K-1)/2 bộ phân lớp nhị phân để xây dựng Kclass.

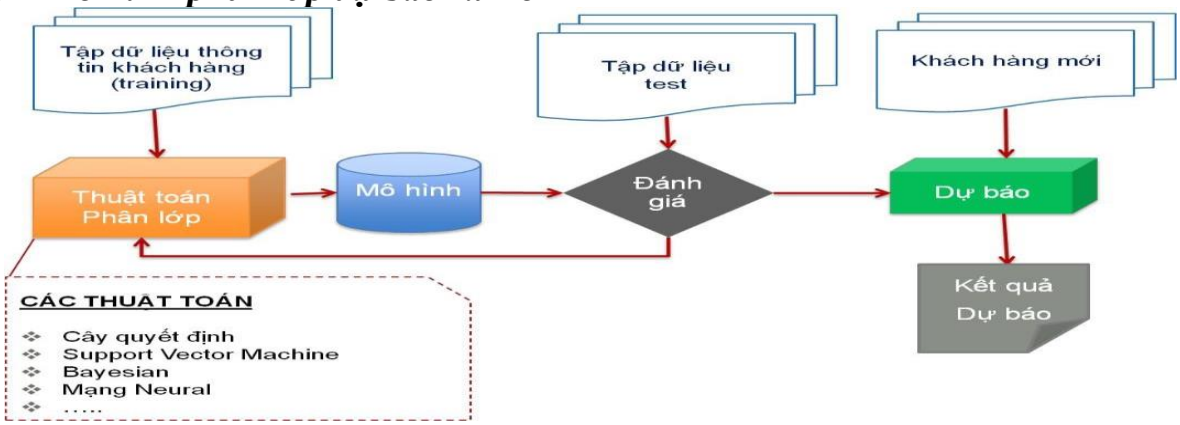
Cả 2 hướng đều dẫn đến vùng mập mờ trong phân lớp (như hình vẽ).

Ta có thể tránh được vấn đề này bằng cách xây dựng K-Class dựa trên K hàm tuyến tình có dạng:

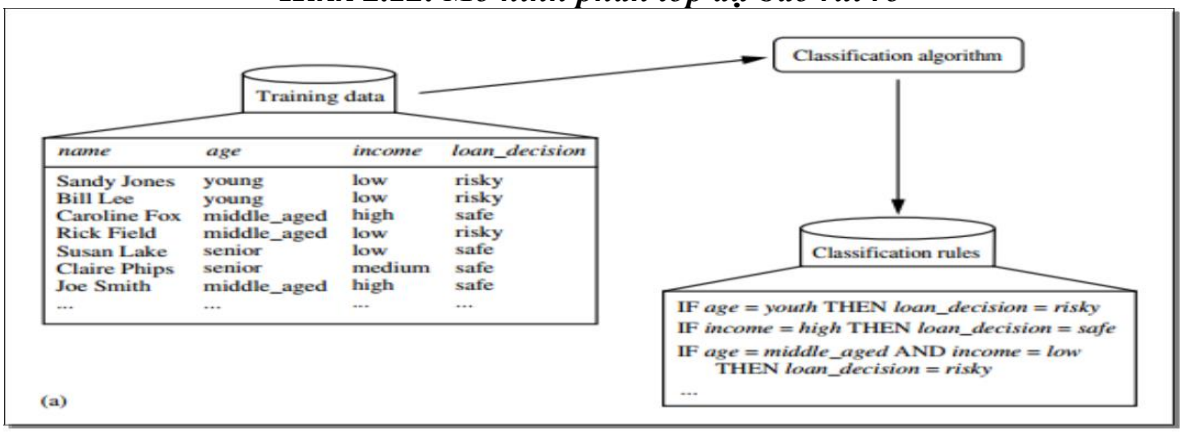
$$y_k(x) = w_k^t x + w_{k0}$$

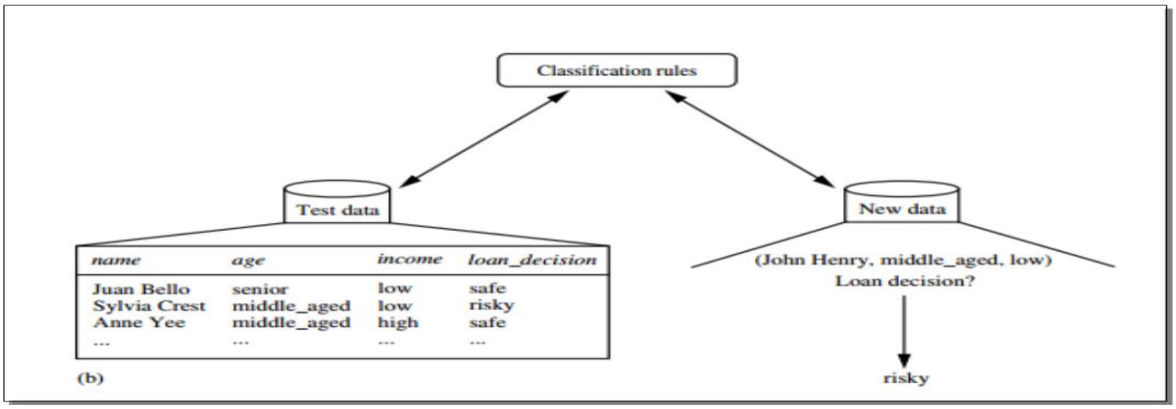
Và một điểm x được gán vào lớp C_k khi $y_k(x) > y_j(x)$ với mọi j ≠k.

2.4. Mô hình phân lớp dự báo rủi ro



Hình 2.12: Mô hình phân lớp dự báo rủi ro





Hình 2.13: Quy trình phân lớp

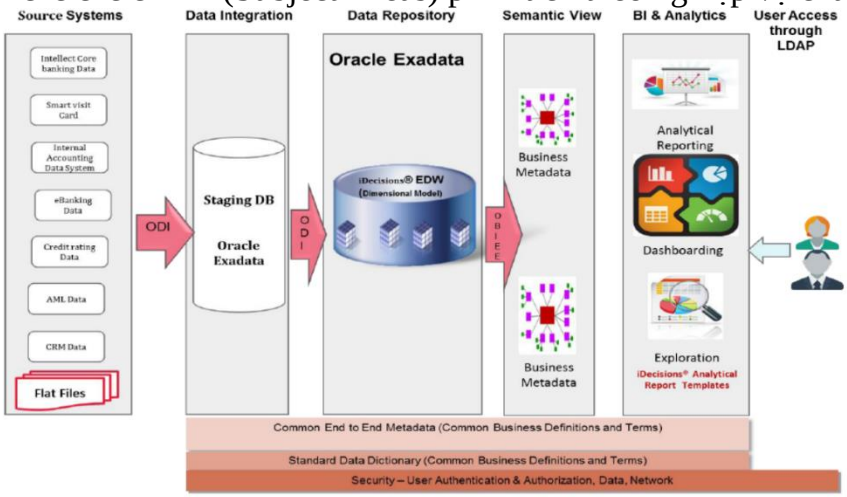
2.5. Kết luận chương 2

Qua tìm hiểu cơ sở lý thuyết về khai phá dữ liệu và ứng dụng thực tiễn của khai phá dữ liệu trong hệ thống các công ty tài chính và ngân hàng cụ thể là áp dụng bài toán phân lớp dự báo rủi ro tín dụng, chúng ta đã hiểu được tầm quan trọng của việc ứng dụng khai phá dữ liệu vào công tác quản lý rủi do tại ngân hàng.Trong chương tiếp theo luận văn sẽ thử nghiệm Phân lớp sử dụng cây quyết định áp dụng **thuật toán C4.5 và phân lớp sử dụng SVM** vào giải quyết bài toán phân lớp dự báo rủi ro tín dụng với tập dữ liệu mẫu là tập dữ liệu khách hàng tại SHB.

CHƯƠNG 3: THỬ NGHIỆM VÀ ĐÁNH GIÁ RỦI RO TÍN DỤNG TẠI NGÂN HÀNG SHB

3.1. Kho dữ liệu của SHB

Kho dữ liệu của SHB là giải pháp về kho dữ liệu lưu trữ các thông tin từ các hệ thống khác nhau trong ngân hàng như Core Banking(Intellect Polaris), Thẻ(Smart Vista), Kế toán nội bộ(IAS), CRM (Quản lý quan hệ khách hàng),Internet Banking, Mobile Banking,… và cung cấp dữ liệu tập trung từ nhiều nguồn dữ liệu của SHB phục vụ cho công tác khai thác và phân tích dựa trên các các chủ đề(Subject Areas) phần tích theo nghiệp vụ của ngân hàng.



Hình 3.1: Mô hình và kiến trúc kho dữ liệu của SHB

Dựa vào kiến trúc trên ta có thể thấy dữ liệu được chia ra thành 3 lớp rõ ràng:

- **Vùng làm tích hợp dữ liệu (Data Intergration**
- **Vùng lư trữ dữ liệu (Data Repository)**
- **Vùng dữ liệu theo chủ đề (Semantic Layer)**

3.2. Thử nghiệm các thuật toán phân lớp cho dự báo rủi ro tín dụng của SHB

Trước khi đi vào thử nghiệm bài toán phân lớp và dự báo rủi ro tín dụng của SHB. Chúng ta sẽ định nghĩa về bộ dữ liệu thông tin khách hàng cá nhân bao gồm những thông tin như sau:

- Thông tin về cá nhân khách hàng
- Thông tin về khả năng trả nợ của khách hàng
- Thông tin về quan hệ của khách hàng với SHB cũng như các tổ chức tín dụng khác
- Thông tin về phương án đầu tư của khách hàng

Bảng 3.1: Các trường thông tin chi tiết về khách hàng các nhân

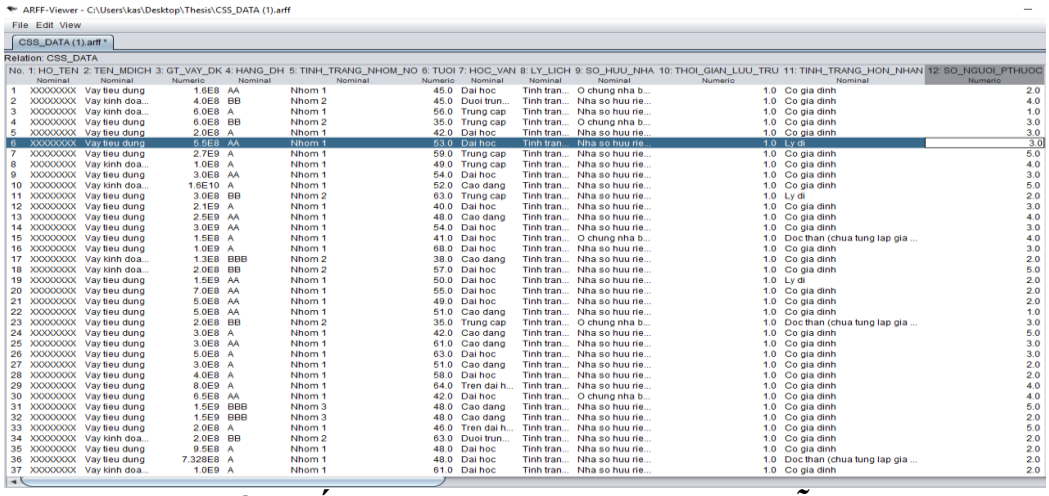
Số thứ tự	Chỉ Tiêu
I.	**Thông tin về cá nhân khách hàng**
1	Tuổi
2	Trình độ học vấn
3	Lý lịch tư pháp
4	Tình trạng sở hữu nhà ở/BĐS
5	Thời gian lưu trú tại địa chỉ hiện tại
6	Tình trạng hôn nhân
7	Số người trực tiếp phụ thuộc về kinh tế vào khách hàng
8	Giá trị hợp đồng bảo hiểm nhân thọ mà SHB là người thụ hưởng so với dư nợ hiện tại của khách hàng
9	Cơ cấu gia đình dựa trên tình trạng thực tế
10	Đánh giá mối quan hệ của KH vay với cộng đồng (uy tín trong công tác, kinh doanh, khu phố địa phương…)

11	Đánh giá mối quan hệ của khách hàng với các thành viên trong gia đình khách hàng
12	Năng lực hành vi dân sự của người thân trong gia đình
13	Đánh giá gia cảnh khách hàng so với mặt bằng chung của vùng
14	Tình trạng sức khỏe của khách hàng
II.	**Thông tin về khả năng trả nợ của khách hàng**
15	Loại hình cơ quan đang công tác
16	Triển vọng phát triển của cơ quan người tham gia trả nợ đang công tác
17	Thời gian làm trong lĩnh vực chuyên môn hiện tại
18	Thời giancông tác tại cơ quan hiện tại
19	Rủi ro nghề nghiệp (thất nghiệp, tai nạn nghề nghiệp, nhân mạng, …)
20	Vị trí công tác
21	Trả lương hoặc chuyển thu nhập qua SHB
22	Hình thức hợp đồng lao động
23	Tổng thu nhập hàng tháng của những người tham gia trả nợ
24	Mức thu nhập ròng ổn định hàng tháng của những người tham gia trả nợ
25	Tỷ lệ giữa tổng số tiền phải trả còn lại (gốc+lãi) và nguồn thu nhập trả nợ cho SHB
26	Đánh giá của cán bộ tín dụng về khả năng trả nợ của khách hàng
III.	**Thông tin về quan hệ của khách hàng với SHB cũng như các tổ chức tín dụng khác**
27	Số lần cơ cấu lại nợ hoặc nợ quá hạn trên 10 ngày trong 12 tháng gần nhất
28	Tỷ trọng nợ (nợ gốc, lãi) cơ cấu lại hoặc quá hạn từ 10 ngày trở lên trên tổng dư nợ của khách hàng vay tại SHB tại thời điểm đánh giá cấp tín dụng
29	Tình trạng dư nợ hiện tại
30	Tỷ trọng tiền gửi tiết kiệm tại SHB so với dư nợ hiện tại của khách hàng
31	Tình hình cung cấp thông tin của khách hàng theo yêu cầu của SHB trong 12 tháng gần nhất
32	Tình hình trả nợ gốc và lãi với các tổ chức tín dụng trong 12 tháng gần nhất (tính đến thời điểm đánh giá)
33	Thời gian khách hàng quan hệ với SHB
34	Số các Tổ chức tín dụng mà khách hàng đang có quan hệ tín dụng hiện tại
IV.	**Thông tin về phương án đầu tư của khách hàng**
35	Tỷ lệ vốn tự có của KH vay tham gia vào phương án đầu tư
36	Chiều hướng biến động của giá cả sản phẩm khách hàng đang tham gia đầu tư trong 6 tháng vừa gần nhất
37	Đánh giá phương án đầu tư của khách hàng
38	Đánh giá rủi ro gián đoạn hoạt động kinh doanh của khách hàng do tác động của môi trường kinh doanh
39	Tính ổn định của thị trường đầu ra
40	Quan hệ của khách hàng đối với các cá nhân tổ chức khác

a. Dữ liệu mẫu và xử lý dữ liệu nguồn

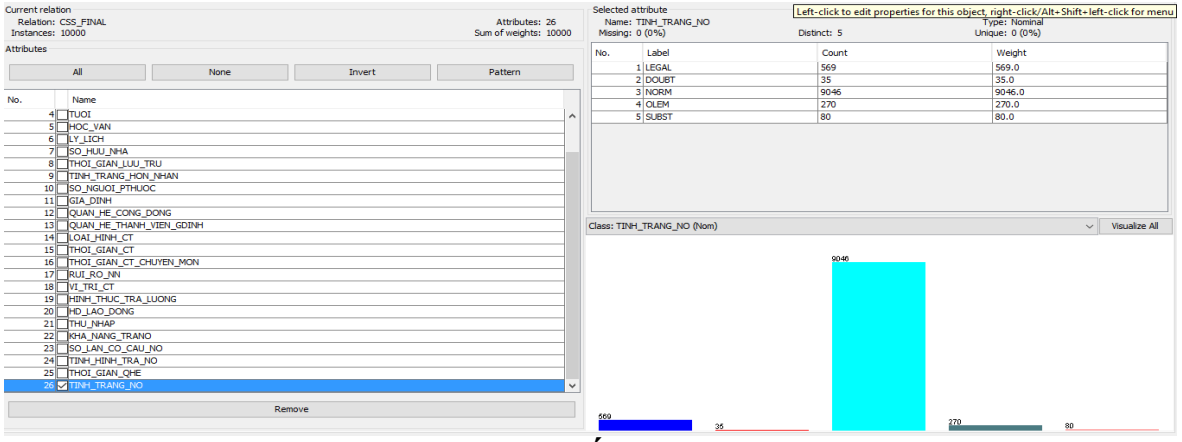
Phạm vi của luận văn cũng như bài toán đã được nêu ở chương số 2 chỉ thực hiện trên tập dữ liệu của khách hàng cá nhân không phải khách hàng cá nhân kinh doanh nên các chỉ tiêu về phương án đầu tư là không có giá trị. Vì vậy trước khi thực hiện thực nghiệm phải loại bỏ các trường không cần thiết này. Ngoài ra trong tập dữ liệu thực tế có một số trường có tỷ lệ các mẫu không có giá trị là cao nên cũng loại bỏ không tham gia vào quá trình xây dựng mô hình phân lớp.

Sau loại bỏ các trường không cần thiết, các trường có tỷ lệ rỗng cao thì còn 24 thuộc tính và có tổng cộng 10000 mẫu như trong hình dưới đây:

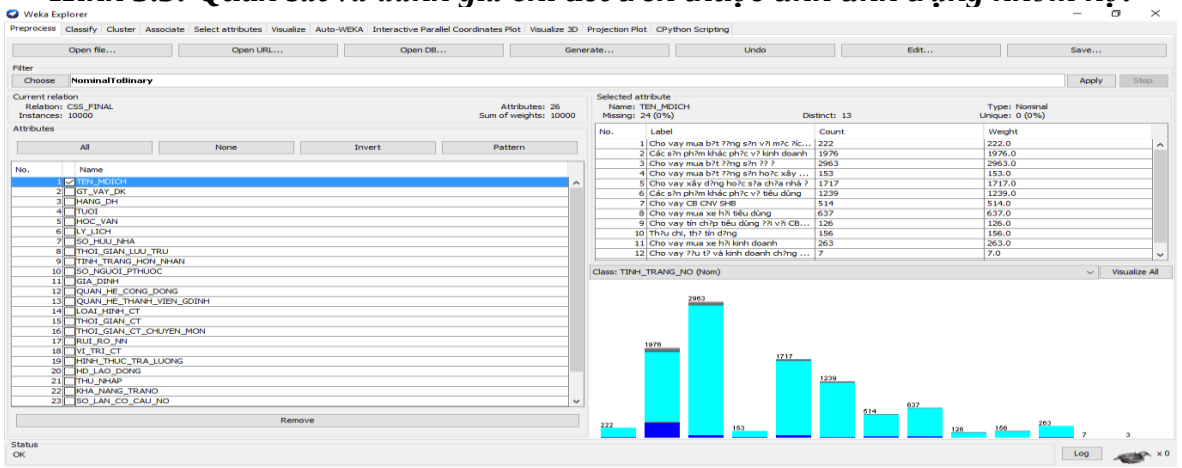


Hình số 3.2: Tập dữ liệu sử dụng làm mẫu.

Để trực quan hơn về các thông tin của một số thuộc tính trong tập dữ liệu mẫu, chúng ta công cụ Weka Explore cho phép xem các thông tin mô tả dữ liệu như tỷ lệ phân bổ chi tiết của của dữ liệu trên thuộc tính, được thể hiện bằng đồ thị rất dễ quan sát và đánh giá:



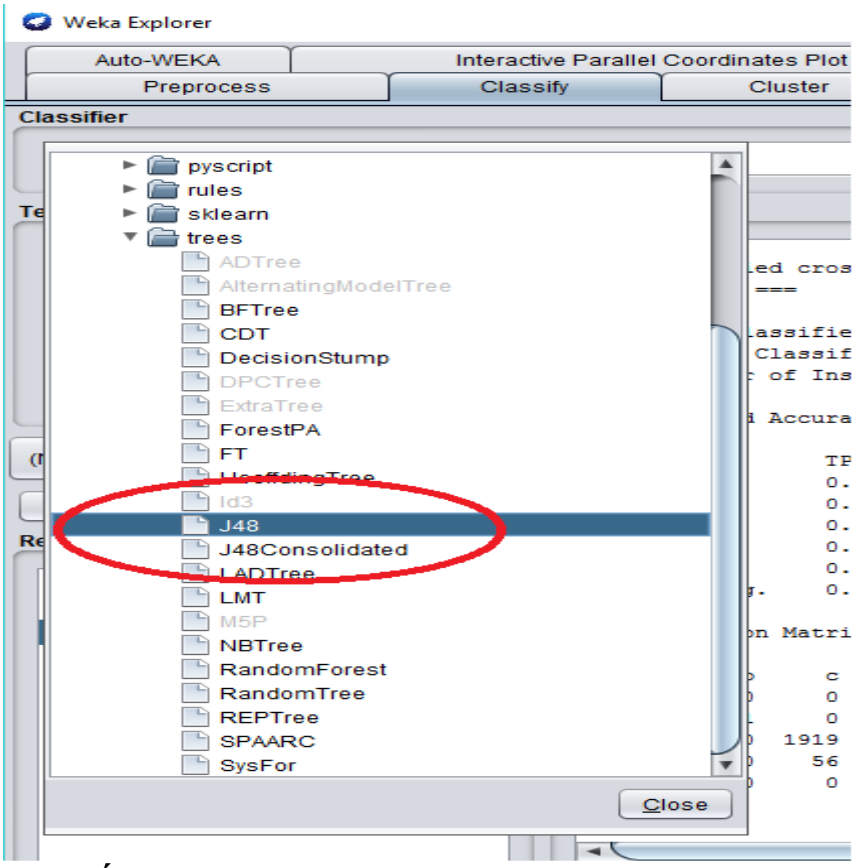
Hình 3.3: Quan sát và đánh giá chi tiết trên thuộc tính tình trạng nhóm nợ.



Hình số 3.4: Quan sát và đánh giá chi tiết trên thuộc tính tên mục đích vay.

b. Phân lớp sử dụng cây quyết định áp dụng thuật toán C4.5

Để kiểm nghiệm thuật toán phân lớp sử dụng cây quyết định C4.5 trên dữ liệu mẫu đã nêu ta thực hiện trên công cụ Weka Explore phiên bản 3.8.2 bằng cách chọn thuật toán J48 như sau:



Hình số 3.5: Cách cài đặt thuật toán C4.5 trên Weka Explore

Cài đặt thông số trên thuật toán: Trong giải thuật cây quyết định C4.5 hay J48 được cung cấp bởi Weka có 3 tham số quan trọng:

- ***confidenceFactor:*** *Nhân tố sử dụng cho việc cắt tỉa (Nếu giá trị này càng nhỏ thì cây sinh ra sẽ được cắt càng nhiều).*
- ***minNumObj:*** *Số thể hiện tối thiểu trên một nút lá trong cây.*
- ***unPruned:*** *nếu là True thì cây sinh ra sẽ được cắt tỉa và ngược lại.*

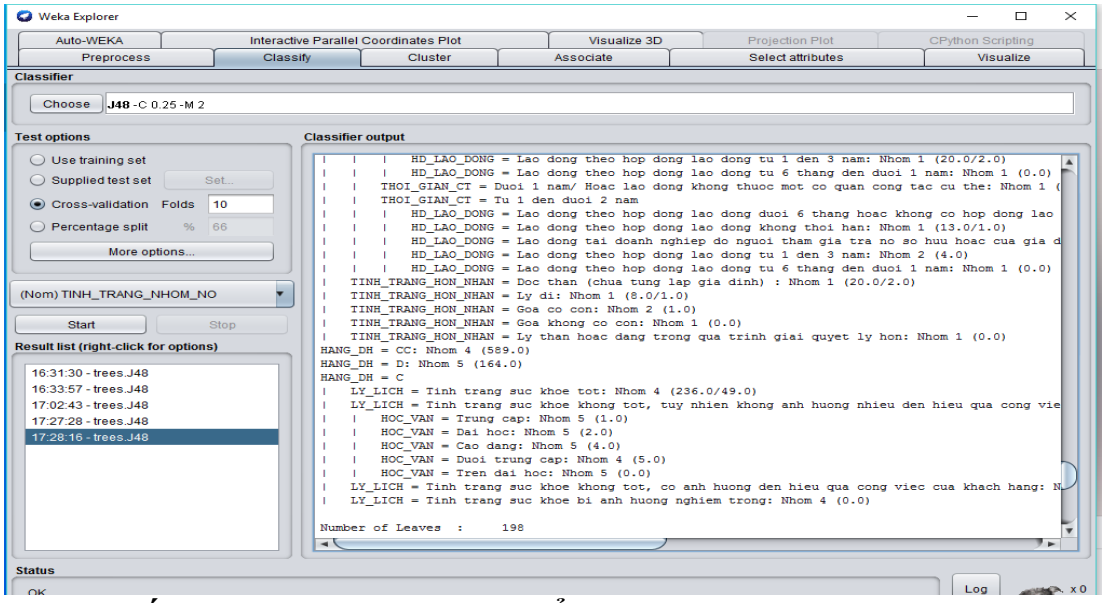
Sau khi điều chỉnh thông số và nghiên cứu ta chọn các giá trị tham số cho kết quả tốt nhất như sau:

- Chọn phương pháp test: Cross Validation
- Tham số thuật toán: **minNumObj**=10
- **confidenceFactor**=0.2
- **unpruned**=False

Kết quả của xây dựng mô hình phân lớp bằng cây quyết định áp dụng thuật toán C4.5 trên tập dữ liệu 10000 mẫu như sau:

Bảng 3.2: Bảng kết quả xây dựng cây quyết định áp dụng thuật toán C4.5

Thời gian xây dựng mô hình (Time taken to build model)	0.28 seconds
Số lá của cây (Number of Leaves)	81
Số nút của cây (Size of the tree)	104
Số mẫu phân lớp đúng (Correctly Classified Instances)	9667(Tỷ lệ: 96.67%)
Số mẫu phân lớp sai (Incorrectly Classified Instances)	333(Tỷ lệ: 3.33 %)

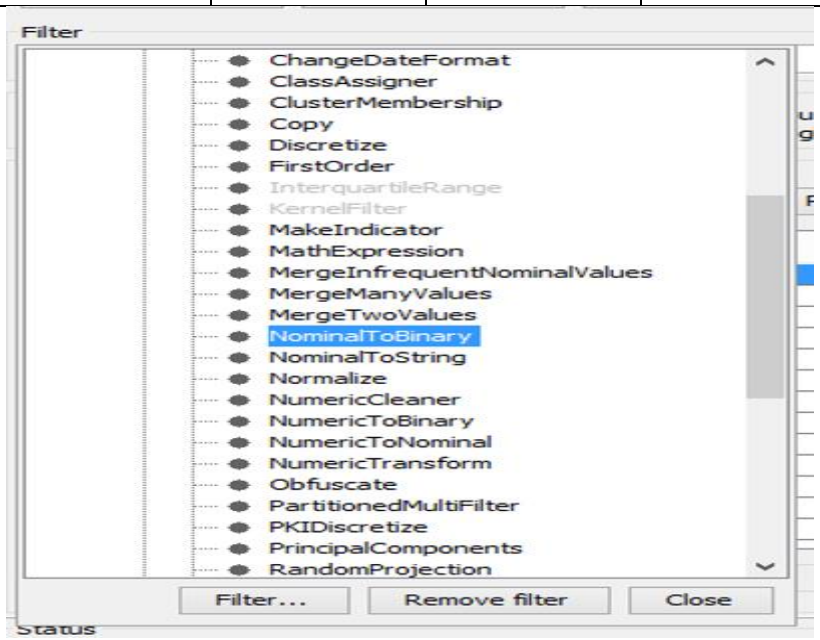


Hình số 3.5: Mô hình C4.5 được thể hiện trên màn hình Weka Explore

Ma trận thể hiện kết quả xây dựng trên tập 10000 mẫu là:

Bảng 3.3: Kết quả phân lớp C4.5 trên tập mẫu

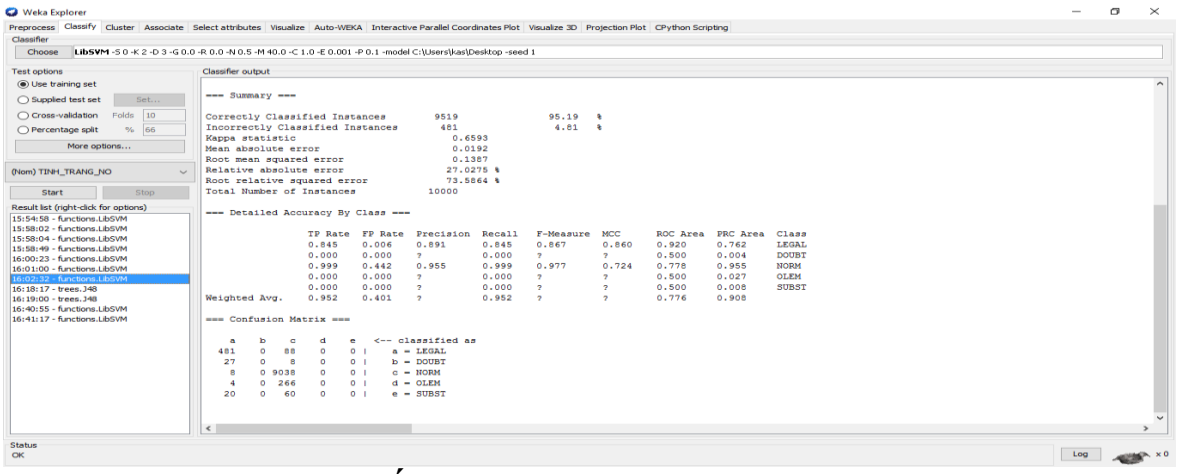
classified as	a	b	c	d	e
a = Nhom 2	0	0	205	62	2
b = Nhom 1	8	0	9036	0	2
c = Nhom 3	9	0	1919	0	55
d = Nhom 4	13	17	16	0	2
e = Nhom 5	496	0	72	1	0



Hình 3.6: Bộ chuyển đổi từ Nominal sang kiểu Binary

Bảng 3.4: Bảng kết quả xây dựng với mô hình phân lớp SVM

Thời gian xây dựng mô hình (Time taken to build model)	5.67 seconds
Số mẫu phân lớp đúng (Correctly Classified Instances)	9519 (Tỷ lệ: 95.19 %)
Số mẫu phân lớp sai (Incorrectly Classified Instances)	481 (Tỷ lệ: 4.81%)



Hình 3.7: Kết quả mô hình SVM trên Weka Explore

Ma trận thể hiện kết quả xây dựng trên tập 10000 mẫu là:

Bảng 3.5: Kết quả phân lớp SVM trên tập mẫu

classified as	a	b	c	d	e
a = Nhom 2	4	0	266	0	0
b = Nhom 1	8	9038	0	0	0
c = Nhom 3	20	0	266	0	0
d = Nhom 4	27	0	8	0	0
e = Nhom 5	481	0	88	0	0

3.3. So sánh kết quả đánh giá và đề xuất ứng dụng

Để đánh giá được hiệu quả của 2 phương pháp phân lớp và dự báo của 2 mô hình đã được thử nghiệm đánh giá ở trên phải dựa trên nhiều tiêu chí để đánh giá như độ chính xác (precision), độ hồi tưởng (recall), … các tiêu chí được đánh giá như dưới đây:

Bảng 3.6: Bảng tiêu chí đánh giá mô hình phân lớp

Lớp C_i		Dữ liệu thực	
		Thuộc lớp C_i	Không thuộc lớp C_i
Dự đoán	Thuộc lớp C_i	TP_i	TN_i
	Không thuộc lớp C_i	FP_i	FN_i

Trong đó:

- ***TP_i (true positives):*** *số lượng ví dụ dương được thuật toán phân đúng vào lớp C_i.*
- ***TN_i (true negatives):*** *số lượng ví dụ âm được thuật toán phân đúng vào lớp C_i.*
- ***FP_i (false positives):*** *số lượng ví dụ dương được thuật toán phân sai vào lớp C_i.*
- ***FN_i (false negatives):*** *số lượng ví dụ âm được thuật toán phân sai vào lớp C_i.*

Độ chính xác **Precision** của lớp $C_i$là tỷ lệ số ví dụ dương được thuật toán phân lớp cho giá trị đúng trên tổng số ví dụ được thuật toán phân lớp vào lớp C_i:

$$\boldsymbol{Pr = \frac{TP_i}{TP_i + TN_i}}$$

Độ chính xác **Recall** của lớp C_i là tỷ lệ số ví dụ dương được thuật toán phân lớp cho giá trị đúng trên tổng số ví dụ được thuật toán phân lớp vào lớp C_i:

$$\boldsymbol{Pr = \frac{TP_i}{TP_i + FN_i}}$$

Kết quả của các mô hình được đánh giá qua công thức trên được thể hiện qua các bảng kết như sau:

- Với mô hình xây dựng bằng cây quyết định C4.5:

Bảng 3.7: Bảng các chỉ số đánh giá phương pháp phân lớp C4.5

Class	TP Rate	FP Rate	Precision	Recall
Nhom 5	0.872	0.003	0.943	0.872
Nhom 4	0.486	0.000	1.000	0.486
Nhom 1	0.999	0.312	0.968	0.999
Nhom 2	0.233	0.000	0.984	0.233
Nhom 3	0.688	0.000	0.932	0.688
Weighted Avg.	**0.967**	**0.283**	**0.967**	**0.967**

Kết quả ở trên được đánh giá trên tập dữ liệu mẫu bảo gồm 10000

- Với mô hình dựa trên thuật toán SVM

Bảng 3.8: Bảng các chỉ số đánh giá phương pháp phân lớp SVM

Class	Recall	Precision	FP Rate	TP Rate
Nhom 5	0.845	0.891	0.006	0.845
Nhom 4	0.000	0.000	0.000	0.000
Nhom 1	0.999	0.955	0.442	0.999
Nhom 2	0.000	0.000	0.000	0.000
Nhom 3	0.000	0.000	0.000	0.000
Weighted Avg.	**0.952**	**0.977**	**0.401**	**0.952**

3.4. Kết luận chương 3

Đây là chương cuối và cũng là một trong những chương quan trọng nhất của luận văn đã thể hiện được ứng dụng của khai phá dữ liệu ứng vào bài toán phân dự báo rủi ro tín dụng tại ngân hàng SHB dựa vào kiến trúc kho dữ liệu liệu của SHB. Thử nghiệm bài toán với 2 thuật toán phân lớp là:

- Phân lớp sử dụng cây quyết định áp dụng thuật toán C4.5
- Phân lớp sử dụng thuật toán SVM

Đánh giá và so sánh ưu nhược điểm chi tiết trên thông số kỹ thuật của 2 thuật toán trên sử dụng cùng 1 bộ dữ liệu mẫu được mô phỏng dựa trên 10000 thông tin khách hàng.

KẾT LUẬN

1. Kết quả của luận văn

Luận văn đã giới thiệu những khái niệm và lý thuyết cơ bản hoạt động tín dụng của ngân hàng, tầm quan trọng của tín dụng trong ngân hàng và những rủi ro của hoạt động tín dụng. Cũng như ảnh hưởng của rủi ro tín dụng đến các ngân hàng thương mại và hệ thống ngân hàng ở Việt Nam. Đồng thời nêu ra hiện trạng tại ngân hàng SHB và nhu cầu cần thiết phải áp dụng công nghệ vào trong quản lý rủi ro tín dụng.

Trình bày các bước trong quá trình khai phá dữ liệu và các phương pháp khai phá dữ liệu hiện nay. Chính từ ý nghĩa thực tế của khai phá dữ liệu nên luận văn đã đưa ra các bài toán, các lĩnh vực mà ngành ngân hàng có thể áp dụng như áp dụng khai phá dữ liệu trong quản trị rủi ro ngân hàng, áp dụng trong phát hiện gian lận, trong kinh doanh, …

Với sự ứng dụng rộng rãi của khai phá dữ liệu trong ngành tài chính ngân hàng đó. Để chứng minh sự tính thực tế, luận văn đã đề xuất bài toán phân lớp dự báo để dự báo rủi ro tín dụng. Việc áp dụng các thuật toán phân lớp vào bài toán thực tế này thì có rất nhiều thuật toán song do thời lượng luận văn có hạn luận văn chỉ đề cập 2 phương pháp phân lớp thường được sử dụng là sử dụng cây quyết định áp dụng thuật toán C4.5 và phân lớp dựa trên thuật toán SVM. Từ đó đi sâu tìm hiểu về 2 thuật toán này.

Song song với nghiên cứu và tìm hiểu lý thuyết luận văn đã tìm hiểu về quy định quy trình về tín dụng và hệ thống đang có tại ngân hàng SHB để áp dụng các lý thuyết đã tìm hiểu trong việc khai phá dữ liệu áp dụng vào bài toán phân lớp dự báo rủi ro tín dụng tại ngân hàng SHB.

Kết quả thực nghiệm chỉ ra rằng thuật toán SVM cho kết quả phân lớp tốt hơn trong các lớp so với thuật toán cây quyết định.

2. Định hướng phát triển

Với rất nhiều ứng dụng thực tiễn của khai phá dữ liệu trong ngành tài chính ngân hàng, đặc biệt trong việc phân tích dự báo rủi ro tín dụng. Với thời gian có hạn luận văn mới chỉ nghiên cứu và thực nghiệm phân lớp dựa trên 2 thuật toán, vì vậy yêu cầu với bài toán trong tương lai là áp dụng các thuật toán khác như hồi quy dự báo, áp dụng mạng noron xây dựng các mô hình dự báo…Với sự ứng dụng rộng rãi của khai phá dữ liệu trong ngành tài chính ngân hàng như đã trình bày thì còn rất nhiều bài toán có thể tìm hiểu và nghiên cứu thêm trong tương lai.