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

LUẬN VĂN THẠC SĨ KỸ THUẬT

(THEO ĐỊNH HƯỚNG ỨNG DỤNG)

HÀ NỘI - 2020

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

LUẬN VĂN THẠC SĨ KỸ THUẬT

(Theo định hướng ứng dụng)

NGƯỜI HƯỚNG DẪN KHOA HỌC:

PGS.TS. LÊ HỮU LẬP

HÀ NỘI - 2020

LỜI CAM ĐOAN

Tôi xin cam đoan kết quả đạt được trong luận văn là sản phẩm của riêng cá nhân tôi, không sao chép lại của người khác. Trong toàn bộ nội dung của luận văn, những điều đã trình bày là của cá nhân tôi hoặc được tôi tổng hợp từ nhiều nguồn tài liệu. Tất cả các nguồn tài liệu tham khảo có xuất xứ rõ ràng và được trích dẫn hợp pháp.

Tôi xin chịu toàn bộ trách nhiệm và chịu mọi hình thức kỷ luật theo quy định cho lời cam đoan của tôi.

Tác giả luận văn

Nguyễn Khắc Xuân Bách

LỜI CẢM ƠN

Lời đầu tiên tôi xin gửi lời cảm ơn và lòng biết ơn sâu sắc đến **PGS.TS. Lê Hữu Lập**, người đã giúp tôi chọn đề tài, định hình hướng nghiên cứu, tận tình hướng dẫn và chỉ bảo tôi trong suốt quá trình thực hiện luận văn tốt nghiệp.

Tôi xin bày tỏ lòng biết ơn trân thành tới các thầy, cô giáo trong trường Học viện Công nghệ và Bưu chính Viễn thông. Các thầy, cô giáo đã dạy bảo và truyền đạt cho tôi rất nhiều kiến thức, giúp tôi có được một nền tảng kiến thức vững chắc sau những ngày tháng học tập tại trường. Và xin gửi lời cảm ơn đến Ban Lãnh đạo và các đồng nghiệp tại Khối công nghệ thông tin – Ngân hàng thương mại cổ phần Sài Gòn – Hà Nội (SHB) đã hết sức tạo điều kiện thuận lợi cho tôi trong suốt quá trình học tập và thực hiện luận văn. Tôi xin gửi sâu sắc các bạn khóa 2018 đợt 1 đã ủng hộ khuyến khích tôi trong suốt quá trình học tập tại trường.

Cuối cùng, tôi muốn gửi lời cảm ơn sâu sắc nhất đến gia đình và bạn bè – những người thân yêu luôn kịp thời động viên và giúp đỡ tôi vượt qua những khó khăn trong học tập cũng như trong cuộc sống.

Hà Nội, tháng 12 năm 2019

Nguyễn Khắc Xuân Bách

MỤC LỤC

DANH MỤC CÁC THUẬT NGỮ, CÁC CHỮ VIẾT TẮT

Viết tắt	Tiếng Anh	Tiếng Việt
BI	Business Intelligence	Kinh doanh thông minh
CSS	Credit Scoring System	Hệ thống xếp hạng tín dụng
DWH	Data Warehouse	Kho dữ liệu
IAS	Internal Accounting System	Hệ thống kế toán nội bộ
KDD	Knowledge Discovery and Data Mining	Khai phá dữ liệu tri thức
KPDL	Data Mining	Khai phá dữ liệu
NHTM	Commercial Bank	Ngân hàng thương mại cổ phần
SHB	Sai Gon – Ha Noi commercial Bank	Ngân hàng thương mại cổ phần Sài Gòn – Hà Nội

DANH SÁCH BẢNG

DANH SÁCH HÌNH VẼ

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. Xin trân trọng cảm ơn !.

Tác giả

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.

- Việc phân phối vốn tín dụng đã góp phần điều hoà vốn trong toàn bộ nền kinh tế, tạo điều kiện cho quá trình sản xuất được liên tục. Tín dụng còn là cầu nối giữa tiết kiệm và đầu tư. Nó là động lực kích thích tiết kiệm đồng thời là phương tiện đáp ứng nhu cầu về vốn cho đầu tư phát triển.
- Trong nền kinh tế sản xuất hàng hoá, tín dụng là một trong những nguồn vốn hình thành vốn lưu động và vốn cố định của doanh nghiệp, vì vậy tín dụng đã góp phần động viên vật tư hàng hoá đi vào sản xuất, thúc đẩy tiến bộ khoa học kỹ thuật đẩy nhanh quá trình tái sản xuất xã hội.

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

- Hoạt động của ngân hàng là tập trung vốn tiền tệ tạm thời chưa sử dụng, trên cơ sở đó cho vay các đơn vị kinh tế. Mặt khác quá trình đầu tư tín dụng được thực hiện một cách tập trung, chủ yếu là cho các xí nghiệp lớn, những xí nghiệp kinh doanh hiệu quả.

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

- Trong thời gian tập trung phát triển nông nghiệp và ưu tiên cho xuất khẩu … Nhà nước đã tập trung tín dụng để tài trợ phát triển các ngành đó, từ đó tạo điều kiện phát triển các ngành khác.

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

- Đặc trưng cơ bản của vốn tín dụng là sự vận động trên cơ sở hoàn trả và có lợi tức, nhờ vậy mà hoạt động của tín dụng đã kích thích sử dụng vốn có hiệu quả. Bằng cách tác động như vậy, đòi hỏi các doanh nghiệp khi sử dụng vốn tín dụng phải quan tâm đến việc nâng cao hiệu quả sử dụng vốn, giảm chi phí sản xuất, tăng vòng quay của vốn, tạo điều kiện nâng cao doanh lợi của doanh nghiệp.

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

- Trong điều kiện kinh tế "mở", tín dụng đã trở thành một trong những phương tiện nối liền các nền kinh tế các nước với nhau.

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

Rủi ro 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.

Rủi ro tín dụng là khả năng tiềm ẩn có thể gây tổn thất về vốn và thu nhập cho Ngân hàng phát sinh khi đối tác không đáp ứng được một phần hoặc toàn bộ các điều khoản của Hợp đồng tín dụng hay không thực hiện đầy đủ như đã thỏa thuận theo các điều khoản của Hợp đồng tín dụng.

Rủi ro tín dụng xuất hiện trong quá trình cho vay, chiết khấu giấy tờ có giá, cho thuê tài chính, bảo lãnh ngân hàng, bao thanh toán và các hình thức cấp tín dụng khác của ngân hàng.

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

Trong quan hệ tín dụng có hai đối tượng tham gia là ngân hàng cho vay và người đi vay. Ngân hàng và người đi vay hoạt động tuân theo sự chi phối với những điều kiện cụ thể của môi trường kinh doanh. Môi trường kinh doanh là đối tượng thứ ba có mặt trong quan hệ tín dụng. Rủi ro tín dụng xuất phát từ môi trường kinh doanh gọi là rủi ro do nguyên nhân khách quan. Rủi ro xuất phát từ người vay và ngân hàng cho vay gọi là rủi ro do nguyên nhân chủ quan. Sự tiếp cận các yếu tố, nguyên nhân gây rủi ro sau đây giúp chúng ta nhìn nhận một cách đầy đủ, toàn diện, khách quan hơn, từ đó sẽ đưa ra được những đề xuất phòng ngừa, giảm thiểu rủi ro trong kinh doanh của NHTM một cách hữu ích, thiết thực hơn.

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

a. Đối với nền kinh tế

- Hệ thống ngân hàng có mối quan hệ chặt chẽ với nền kinh tế, là kênh thu hút và cung cấp tiền cho các tổ chức, doanh nghiệp và cá nhân trong nền kinh tế. Do đó, rủi ro tín dụng có ảnh hưởng trực tiếp đến nền kinh tế.
- Ở mức độ thấp, rủi ro tín dụng khiến cơ hội tiếp cận vốn mở rộng hoạt động sản xuất kinh doanh hoặc tiêu dùng của các khách hàng bị hạn chế, ảnh hưởng xấu đến khả năng tăng trưởng của nền kinh tế.
- Ở mức độ cao hơn, khi có một ngân hàng lâm vào tình trạng khó khăn dẫn đến phá sản, thì hiệu ứng dây chuyền rất dễ xảy ra trong toàn bộ hệ thống ngân hàng, gây nên khủng hoảng đối với toàn bộ nền kinh tế, ảnh hưởng tiêu cực đến đời sống xã hội và sự phát triển của đất nước.

b. Đối với ngân hàng

- Việc không thu hồi được nợ (gốc, lãi và các khoản phí) làm cho nguồn vốn của các NHTM bị thất thoát, trong khi đó, các ngân hàng này vẫn phải chi trả tiền lãi cho nguồn vốn hoạt động, làm cho lợi nhuận bị giảm sút. Nếu lợi nhuận không đủ thì ngân hàng còn phải dùng chính vốn tự có của mình để bù đắp thiệt hại. Điều này có thể làm ảnh hưởng đến quy mô hoạt động của các NHTM.
- Mặt khác, tỷ lệ nợ quá hạn cao làm cho uy tín, niềm tin vào tiềm lực tài chính của ngân hàng bị suy giảm, dẫn đến làm giảm khả năng huy động vốn của ngân hàng, nghiêm trọng hơn nó có thể dẫn đến rủi ro thanh khoản, đẩy ngân hàng đến bờ vực phá sản và đe dọa sự ổn định của toàn bộ hệ thống ngân hàng.

c. Đối với khách hàng

- Đối với bản thân chủ thể không có khả năng hoàn trả vốn (lãi) cho ngân hàng thì họ gần như không có cơ hội tiếp cận với nguồn vốn ngân hàng và thậm chí là cả những nguồn khác trong nền kinh tế do đã mất đi uy tín.

- Cơ hội tiếp cận vốn ngân hàng của các chủ thể đi vay khác cũng bị hạn chế hơn khi rủi ro tín dụng buộc các NHTM hoặc thắt cho vay hay thậm chí phải thu hẹp quy mô hoạt động.
- Các chủ thể gửi tiền vào ngân hàng có nguy cơ không thu hồi được khoản tiền gửi và lãi nếu như các ngân hàng lâm vào tình trạng phá sản.

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

- Đặc điểm chung của các ngân hàng thương mại Việt Nam hiện nay là danh mục tín dụng vẫn chiếm phần lớn trong tổng tài sản (từ 60% đến 70% tổng tài sản của ngân hàng). Do vậy, việc thu thập thông tin về khách hàng vay để đánh giá khả năng thu hồi vốn, dự báo rủi ro là nhu cầu cần thiết của các ngân hàng nói chung và ngân hàng SHB nói riêng.
- Hệ thống xếp hạng tín dụng nội bộ có thể được sử dụng trong các quy trình quản lý rủi ro tín dụng sau: Ban hành chính sách tín dụng, Quy trình cho vay, Giám sát rủi ro danh mục tín dụng, Lập báo cáo quản trị rủi ro, Chính sách dự phòng rủi ro tín dụng, Xác định mức vốn an toàn tối thiểu, Phân tích hiệu quả sinh lời của danh mục tín dụng và Xác định khung lãi suất tiêu chuẩn… Tóm lại, hệ thống xếp hạng tín dụng nội bộ là một cấu phần quan trọng và là một công cụ đắc lực trong quản trị kinh doanh ngân hàng.
- 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

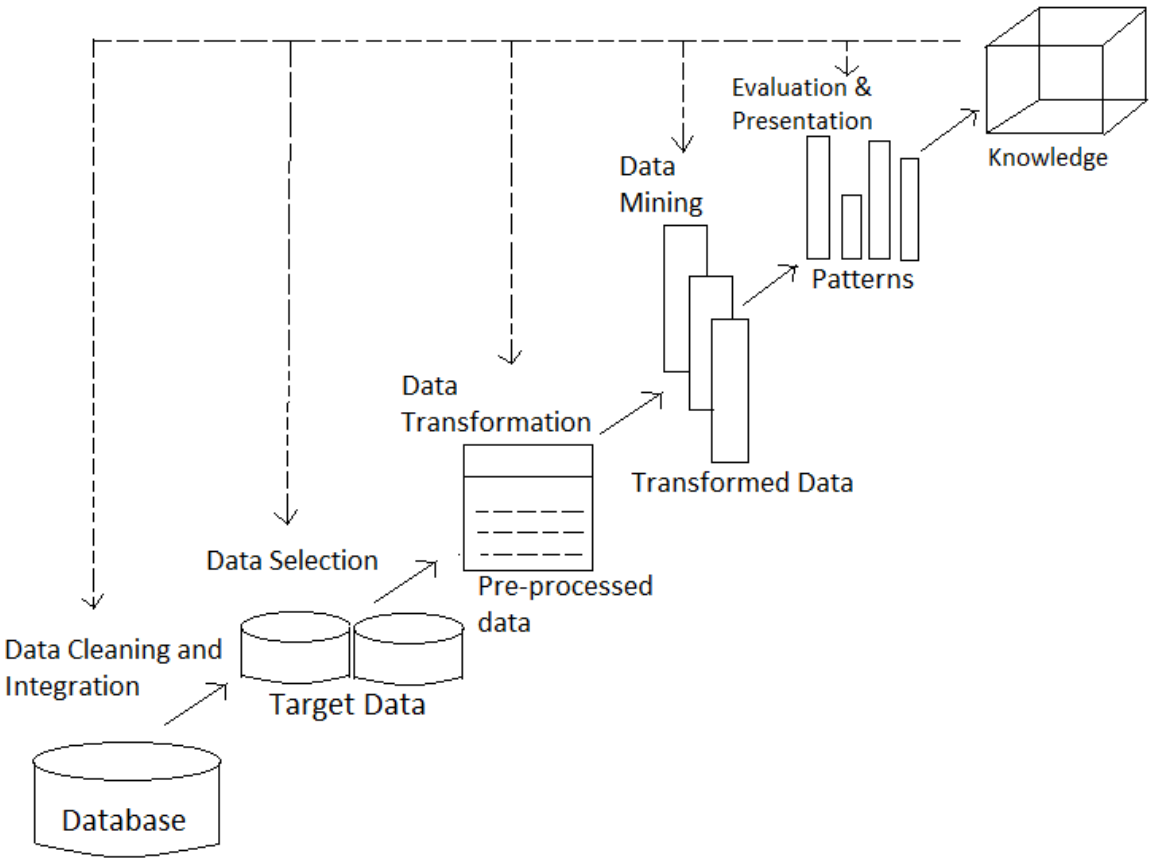
Trong ngành công nghiệp dịch vụ tài chính trên toàn thế giới, phương thức liên lạc truyền thống của khách hàng mặt đối mặt (face-to-face) đang được thay thế bằng phương thức điện tử để giảm thời gian và chi phí xử lý các áp dụng cho sản phẩm khác nhau, và cuối cùng là cải thiện hiệu quả của việc sử dụng 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.

Số lượng dữ liệu được thu thập bởi các ngân hàng đã tăng nhanh chóng trong những năm gần đây. Với những kỹ thuật phân tích số liệu thống kê hiện khó có thể quản lý tốt với khối lượng lớn dữ liệu hiện có như hiện tại. Sự tăng trưởng bùng nổ này đã dẫn đến sự cần thiết của kỹ thuật phân tích dữ liệu mới và các công cụ mới để tìm ra các thông tin thực sự có ích ẩn chứa trong dữ liệu này. Ngân hàng là lĩnh vực mà tại đây một lượng lớn dữ liệu được thu thập. Dữ liệu này có thể được tạo ra từ các giao dịch của các tài khoản ngân hàng, hồ sơ vay vốn, trả nợ, thẻ tín dụng, v.v... Người ta cho rằng thông tin có giá trị về các hồ sơ tài chính của khách hàng được ẩn chữa trong các cơ sở dữ liệu hoạt động lớn và các thông tin này có thể được sử dụng để cải thiện hiệu suất kinh doanh của các ngân hàng. Tại thời điểm ban đầu tại các trung tâm tin học đầu mối của các ngân hàng, nhiều gói phần mềm đang được sử dụng cho các giao dịch hàng ngày. Từ đó, nếu như thiết kế mới một Hệ thống thông tin (MIS: Management Information System) mới hoặc cơ cấu lại những cơ sở hạ tầng hiện sẽ khó thể thực hiện được bởi không chỉ đơn giản là cần thay thế các gói phầm mềm tại các trung tin học đó. Giải pháp cho vấn đề này là để thực hiện các khái niệm về kho dữ liệu và khai phá dữ liệu (Data Warehouse and Data Mining).

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):

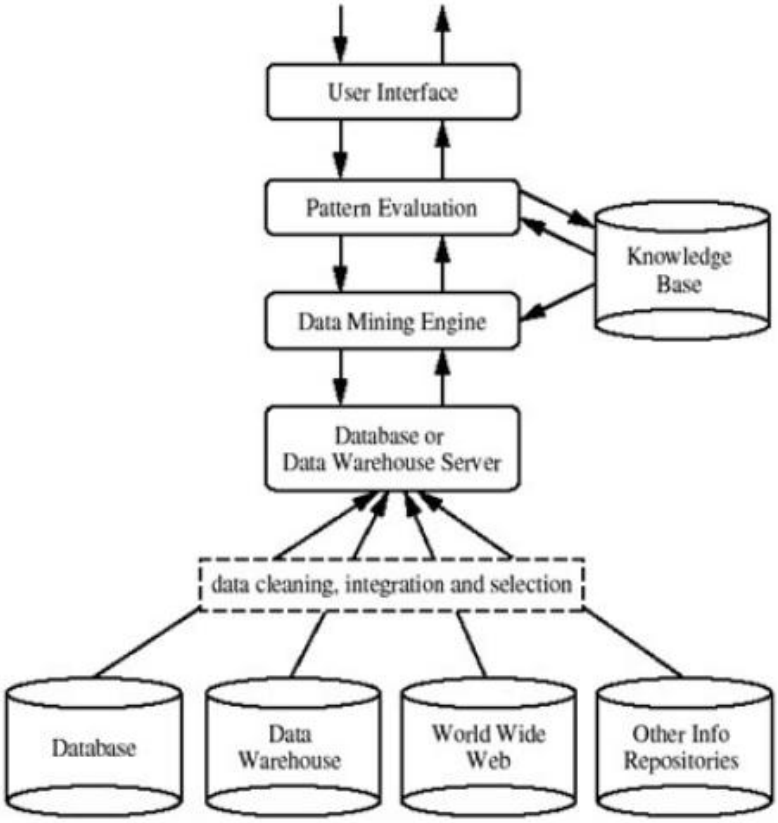


***Hình 2.1: Các bước khai phá dữ liệu*[9]**

- **Làm sạch dữ liệu (data cleaning)**: ở bước này các nhiễu và dữ liệu không nhất quán sẽ được loại bỏ.
- **Tích hợp dữ liệu (data intergation)**: dữ liệu từ nhiều nguồn khác nhau có thể được tổ hợp lại.
- **Lựa chọn dữ liệu (data selection)**: những dữ liệu thích hợp với nhiệm vụ phân tích sẽ được trích rút ra từ cơ sở dữ liệu
- **Chuyển đổi dữ liệu (data tranform)**: dữ liệu sau khi được chọn lọc sẽ được chuyển đổi hay hợp nhất về dạng thích hợp cho việc khai phá.
- **Khai phá dữ liệu (data mining)**: đây là quá trình cốt lõi, tất yếu trong đó các phương pháp thông minh sẽ được áp dụng nhằm trích rút ra các mẫu dữ liệu.
- **Đánh giá mẫu (pattern evaluation)**: các nhà phân tích dữ liệu sẽ dựa trên một số độ đo nào đó để xác định lợi ích thực sự, độ quan trọng của các mẫu biểu diễn tri thức.

- **Biểu diễn tri thức (Knowledge presentation)**: ở giai đoạn này các kĩ thuật biểu diễn và hiển thị tri thức sẽ được sử dụng để đưa tri thức đã lấy ra đến người dùng.

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[9]

Trong đó:

- **Database, Data warehouse, World Wide Web, và Information repositories:** Đây là các nguồn dữ liệu đầu vào cho quá trình khai phá dữ liệu, nguồn dữ liệu sử dụng trong khai phá dữ liệu rất đa dạng như có thể từ cơ sở dữ liệu quan hệ, từ một kho dữ liệu hoặc từ dữ liệu trên web….Trong những tình huống cụ thể, thành phần này là nguồn nhập (input) của các kỹ thuật tích hợp và làm sạch dữ liệu.

- **Data warehouse server:** Thành phần chịu trách nhiệm chuẩn bị dữ liệu thích hợp cho các yêu cầu khai phá dữ liệu.
- **Knowledge base:** Thành phần chứa tri thức miền, được dùng để hướng dẫn quá trình tìm kiếm, đánh giá các mẫu kết quả được tìm thấy. Tri thức miền có thể là các phân cấp khái niệm, niềm tin của người sử dụng, các ràng buộc hay các ngưỡng giá trị, siêu dữ liệu, …
- **Data mining engine:** Thành phần chứa các khối chức năng thực hiện các tác vụ khai phá dữ liệu.
- **Pattern evaluation module:** Thành phần này làm việc với các độ đo (và các ngưỡng giá trị) hỗ trợ tìm kiếm và đánh giá các mẫu sao cho các mẫu được tìm thấy là những mẫu được quan tâm bởi người sử dụng. Thành phần này có thể được tích hợp vào thành phần Data mining engine.
- **User interface:** Thành phần hỗ trợ sự tương tác giữa người sử dụng và hệ thống khai phá dữ liệu.
    - Người sử dụng có thể chỉ định câu truy vấn hay tác vụ khai phá dữ liệu.
    - Người sử dụng có thể được cung cấp thông tin hỗ trợ việc tìm kiếm, thực hiện khai phá dữ liệu sâu hơn thông qua các kết quả khai phá trung gian.
    - Người sử dụng cũng có thể xem các lược đồ cơ sở dữ liệu/kho dữ liệu, các cấu trúc dữ liệu; đánh giá các mẫu khai phá được; trực quan hóa các mẫu này ở các dạng khác nhau.

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):** xếp một đối tượng vào một trong những lớp đã biết trước. Ví dụ: phân lớp vùng địa lý theo dữ liệu thời tiết. Hướng tiếp cận này thường sử dụng một số kỹ thuật của

machine learning như cây quyết định (decision tree), mạng nơron nhân tạo (neural network), v.v. Phân lớp còn được gọi là học có giám sát (học có thầy – supervised learning).

- **Luật kết hợp (association rules):** là dạng luật biểu diễn tri thứ ở dạng khá đơn giản. Ví dụ: "60 % nam giới vào siêu thị nếu mua bia thì có tới 80% trong số họ sẽ mua thêm thịt bò khô". Luật kết hợp được ứng dụng nhiều trong lĩnh vực kinh doanh, y học, tin-sinh, tài chính & thị trường chứng khoán, v.v.
- **Khai phá chuỗi theo thời gian (sequential/temporal patterns):** tương tự như khai phá luật kết hợp nhưng có thêm tính thứ tự và tính thời gian. Hướng tiếp cận này được ứng dụng nhiều trong lĩnh vực tài chính và thị trường chứng khoán vì nó có tính dựbáo cao.
- **Phân cụm (clustering/segmentation):** xếp các đối tượng theo từng cụm (số lượng cũng như tên của cụm chưa được biết trước. Phân cụm còn được gọi là học không giám sát (học không có thầy – unsupervised learning).
- **Mô tả khái niệm (concept description & summarization):** thiên về mô tả, tổng hợp và tóm tắt khái niệm. Ví dụ: tóm tắt văn bản.

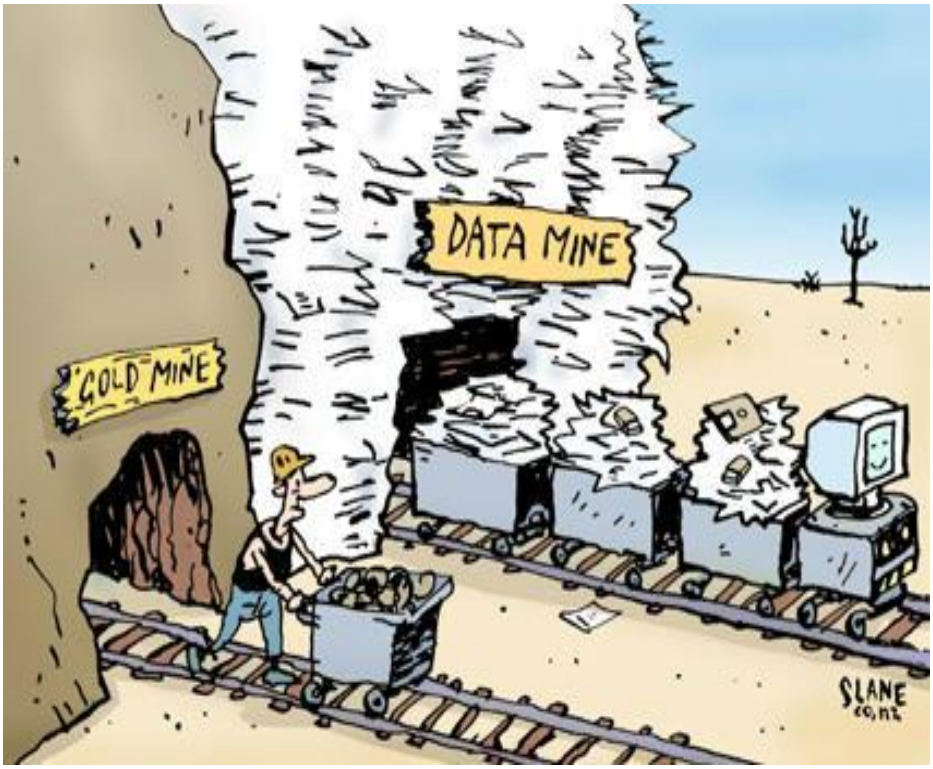
2.2. Ứng dụng của khai phá dữ liệu trong hệ thống thông tin ngân hàng

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…

Khai phá dữ liệu ra đời như một xu hướng tất yếu để hỗ trợ việc khai thác chắt lọc thông tin, và giờ đây khai phá dữ liệu đã và đang trở thành những hướng nghiên cứu chính của lĩnh vực máy tính và khoa học tri thức. Hiện nay khai phá

dữ liệu là một phần không thể thiếu trong hệ thống doanh nghiệp thông minh (Business Intelligence)

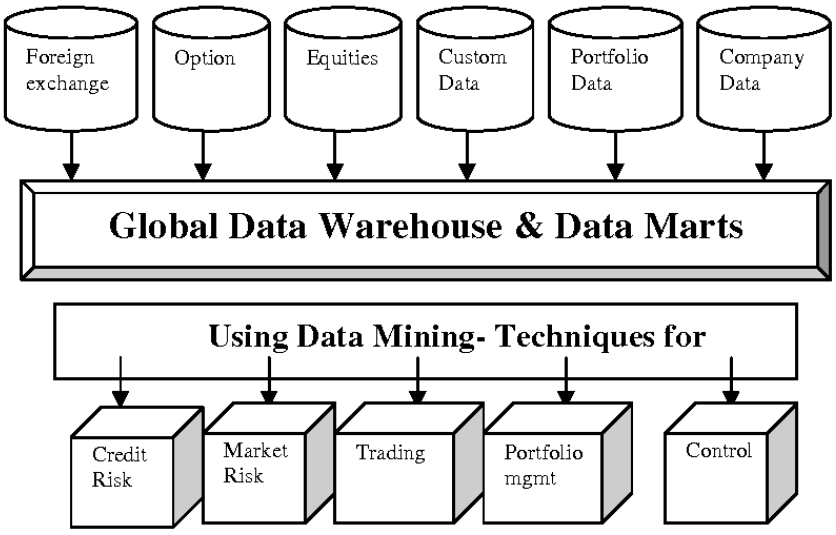
Business Intelligence tập trung vào việc tim kiếm kiến thức từ nhiều kho dữ liệu điện tử cả trong và ngoài để hỗ trợ quá trình ra quyết định. Các kỹ thuật data mining đã trở nên quan trọng cho việc tìm kiếm kiến thức từ các cơ sở dữ liệu. Trong những năm gần đây, Business Intelligence đóng vai trò nòng cốt trong việc hỗ trợ các doanh nghiệp trong việc xây dựng các mục tiêu kinh doanh như giữ lại khách hàng, thâm nhập thị trường, tăng lợi nhuận và hiệu suất. Trong phần lớn các trường hợp, những tri thức này có được từ việc phân tích các dữ liệu lịch sử.



Hình 2.3: Khai phá dữ liệu tìm kiếm tri thức từ lượng dữ liệu khổng lồ[13]

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.

2.2.1. Quản trị rủi ro

Quản trị và đo lường rủi ro là một vấn đề trọng tâm của tất cả các tổ chức tài chính. Thử thách chính của ngành tài chính ngân hàng là sự vận hành hệ thống quản trị rủi ro nhằm nhận diện, đo lường, và kiểm soát khả năng tổn thất. Rủi ro tín dụng và rủi ro thị trường là các thử thách chính. Có nhiều giải thuật

thống kê được áp dụng để hỗ trợ công tác dự báo, đo lường rủi ro tín dụng nhưng chúng là chưa đủ, vì vậy xu hướng hiện nay người ta thường áp dụng các kỹ thuật học máy, kỹ thuật khai phá dữ liệu để đưa ra các mô hình phân tích, dự báo, mô tả lại các tri thức, mối quan hệ giữa các thông tin được thu thập từ các hệ thống khác nhau trong mỗi doanh nghiệp, ngân hàng, ứng dụng khai phá dữ liệu trong quản trị rủi ro ngân hàng thường sử dụng trong đánh giá rủi ro thị trường tài chính và rủi ro tín dụng.

a. Rủi ro trong thị trường tài chính

Với mỗi công cụ tài chính như chỉ số chứng khoán, lãi suất, ngoại tệ, rủi ro thị trường được đo lường, dự báo bằng các mô hình khai phá dữ liệu tạo nên từ một bộ các yếu tố rủi ro phụ thuộc như lãi suất, chỉ số chứng khoán và chỉ số phát triển kinh tế. Người ta quan tâm đến mối liên hệ giữa giá cả hoặc mức độ rủi ro của các công cụ và các yếu tố rủi ro phụ thuộc cũng như sự phụ thuộc của chính các yếu tố rủi ro đó.

Ví dụ: Chúng ta có thể xây dựng mô hình dự báo chứng khoán, tỷ giá ngoại tệ … bằng cách áp dụng các kỹ thuật khai phá dữ liệu để đưa ra mô hình dự báo chỉ số chứng khoán, tỷ giá ngoại tệ để có thể quyết định các chính sách trong quản trị kinh doanh.

b. Rủi ro tín dụng

Đánh giá rủi ro tín dụng là môt bước quan trọng trong hoạt động cho vay trong ngành tài chính ngân hàng. Thiếu bước này người cho vay sẽ không thể đưa ra quyết định khách quan về việc có nên cho vay khách hàng hay không, hay đưa ra lãi suất bao nhiêu là hợp lý.

Trong lĩnh vực cho vay thương mại, đánh giá rủi ro thường là sự cố gắng định lượng độ rủi ro mất mát của người cho vay khi thực hiện một quyết định cho vay nhất định. Ở đây, rủi ro tín dụng có thể được định lượng bằng sự thay đổi giá trị của tài sản thế chấp hoặc, các yếu tố thông tin về người vay, của yếu tố khả năng mất vốn, và tỷ lệ thu hồi của công cụ trong trường hợp không có khả năng trả nợ. Vì vậy, việc áp dụng các kỹ thuật khai phá ở đây chủ yếu là phân lớp, hồi

quy hoặc các kỹ thuật mô tả như phân cụm, mô tả quy luật của tập dữ liệu khách hàng bằng luật kết hợp…

Quản trị rủi ro sử dụng nhiều phương pháp, mô hình dự đoán. Các phương pháp hữu dụng có thể được phân loại rộng theo hai cách tùy thuộc vào vào loại thông tin dự đoán hoặc biến số dự đoán, còn có thể gọi là những biến số mục tiêu. Nếu một loại giá trị dự đoán là giá trị phân lớp, kỹ thuật phân lớp được ưu tiên dùng.

Ví dụ:

- ***Phương pháp phân lớp:***

Theo cách này, các mức độ rủi ro được phân vào hai nhóm dựa trên lịch sử mất vốn. Ví dụ, những khách hàng đã từng không trả nợ có thể được phân vào nhóm "rủi ro", trong khi số còn lại là nhóm "an toàn". Thông tin phân loại là mục tiêu của việc dự đoán, kỹ thuật cây quyết định và nguyên tắc quy nạp được dùng để xây dựng những mô hình dự đoán mức độ rủi ro mất vốn của đăng ký vay vốn mới.

- ***Phương pháp dự báo giá trị:***

Ví dụ, cách này thử dự đoán lượng vốn mất ước tính của những khoản cho vay mới thay cho việc phân loại các khoản vay. Giá trị dự đoán là giá trị bằng số và như vậy nó yêu cầu những kỹ thuật tạo mô hình có thể sử dụng dữ liệu bằng số làm biến số mục tiêu (hoặc dự đoán). Các thuật toán thường dùng trong phương pháp này là mạng noron (Neutral Network) và phương pháp hồi quy. Các kỹ thuậtkhai phá dữ liệu phổ biến nhất được sử dụng cho quản trị rủi ro là:

- *Phân cụm (mô tả)*
- *Phân lớp (dự báo) và hồi quy (dự báo)*
- *Khai phá luật kết hợp...*

2.2.2. Phát hiện gian lận

Các ngân hàng mất hàng triệu đô la mỗi năm bởi các hành vi gian lận. Phát hiện các giao dịch gian lận có thể giúp ngân hàng để hành động sớm và hạn

chế thiệt hại. Phát hiện gian lận là quá trình xác định các hành vi sử dụng các phương tiện bất hợp pháp để có được tiền, thông tin của ngân hàng hoặc của khách hàng sử dụng dịch vụ của ngân hàng. Thông thường gian lận trong ngân hàng thường được thực hiện trên thẻ tín dụng bởi tính chất đặc thù và tiện ích của nó. Việc áp dụng khai phá dữ liệu trong phát hiện gian lận thẻ tín dụng thường sử dụng các thuật toán phân cụm để phân loại các giao dịch hợp pháp và các giao dịch bất thường

Ngoài ra, gian lận cũng thường gặp trong trong các báo cáo tài chính của khách hàng cung cấp cho ngân hàng. Một trong các chỉ tiêu để ngân hàng quyết định cho vay hay không là chỉ tiêu về báo cáo tài chính của khách hàng (khách hàng doanh nghiệp). Các báo cáo tài chính mà khách hàng cung câp có thể phóng đại, doanh số bán hàng và lợi nhuận có thể không chính xác mặc dù báo cáo có thể đã được kiểm toán, các loại gian lậnrất khó phát hiện bằng cách sử dụng thủ tục kiểm toán bình thường. Phương phápphân lớp dựa trên mạng nơron, hồi quy và cây quyết định được sử dụng để phân phân lớp xác định tỷ lệ gian lận trong các báo cáo từ các dữ liệu không gian lận (Sharma và Panigrahi, 2012).

2.2.3. Quản lý danh mục vốn

Phương pháp đánh giá rủi ro ở mức độ tổng thể, quản lý danh mục vốn sẽ xác định độ rủi ro của một nhóm công cụ hoặc khách hàng. Mô hình dự báo sẽ đưa ra mô hình dự báo về thu nhập hoặc giá cả, chi phí, lợi nhuận ước tính từ các danh mục đầu tư để hỗ trợ quản lý trong việc đưa ra chiến lược trong kinh doanh

Với việc Data mining và kỹ thuật tối ưu hóa, nhà đầu tư có thể phân bổ vốn vào các hoạt động giao dịch để tối đa hóa lợi nhuận hoặc tối thiểu hóa rủi ro. Tính năng này hỗ trợ khả năng đưa ra các khuyến cáo trong giao dịch và trong cơ cấu danh mục đầu tư.

Kỹ thuật Data mining tạo điều kiện cho việc phân tích tình huống liên quan đến đánh giá ước tính của tài sản hoặc thu nhập và rủi ro một cách thấu đáo. Với chức năng này, có thể dùng các mô hình thị trường giả định (ví dụ như giả định về lãi suất và tỉ giá hối đoái) để đánh giá tác động của giá trị và rủi ro của

danh mục, đối tác kinh doanh, hoặc phòng kinh doanh. Nhiều tình huống có thể được chú trọng thông qua việc xem xét tính hình thị trường. Phân tích lợi nhuận tổn thất sẽ giúp người dùng đánh giá các lớp tài sản, các vùng, các đối tác, và các tiểu danh mục có thể được so sánh với các mức chuẩn chung quốc tế.

2.2.4. Ứng dụng kinh doanh

Một trong những chủ đề nghiên cứu quan trọng trong những năm gần đây là việc xây dựng các công cụ định lượng trong kinh doanh sử dụng các phương pháp Data mining lấy dữ liệu quá khứ làm đầu vào để dự đoán những biến động ngắn hạn của tỷ giá hối đoái, lãi suất hay chứng khoán thường gọi là diễn biến thị trường.

Mục đích của kỹ thuật này là nhằm phát hiện ra những thời điểm thị trường mất giá hoặc tăng giá bằng cách nhận diện các nhân tố quan trọng quyết định đến lợi nhuận thị trường. Kỹ thuật khai phá dữ liệu nghiên cứu mối quan hệ giữa các thông tin thị trường từ đó giúp nhà quản lý đưa ra các chiến lược phù hợp với thị trường. Ví dụ tăng lãi suất, tăng vốn huy động, thúc đẩy cho vay…

Các giao dịch được thực hiện dựa trên việc dự đoán những biến động trong ngắn hạn của giá cả trên thị trường (ngoại tệ/chứng khoán/lãi suất v.v.). Hoạt động giao dịch được thực hiện dựa trên bản năng của người giao dịch. Người đó có thể mua hoặc bán nếu nghĩ sản phẩm đang không được đánh giá đúng giá trị, bản năng nay thường được dựa trên kinh nghiệm trong quá khứ hoặc qua một số phân tích về điều kiện thị trường. Tuy nhiên, số lượng các nhân tố mà người giao dịch, kể cả các chuyên gia, có thể dựa vào thường là hạn chế. Vì thế, các dự đoán này thường là sai lầm.

Giá cả của các tài sản tài chính bị tác động bởi nhiều yếu tố, có thể được phân loại thành các nhóm lớn bao gồm các yếu tố kinh tế, chính trị và các yếu tố thị trường. Những người tham gia thị trường quan sát mối quan hệ giữa các yếu tố này và giá của tài sản, xem xét cả giá trị hiện tại và tương lai của các yếu tố này để định giá tài sản trong tương lai và dựa vào đó mà thực hiện các giao dịch. Thông thường, tại thời điểm một nhà giao dịch nhiều kinh nghiệm phát hiện ra

những yếu tố có lợi này thì nhiều người khác cũng đã khám phá ra cơ hội đó, vì vậy lợi nhuận có thể mang về từ giao dịch cũng giảm đi. Ngoài ra, những nhân tố này cũng có thể liên quan đến nhiều yếu tố khác, khiến cho việc dự đoán trở nên khó khăn hơn.

Kỹ thuật khai phá dữ liệu được dùng để khám phá ra các kiến thức ẩn, các dạng mẫu (patterns) chưa biết và những quy tắc mới từ một bộ dữ liệu lớn. Những thông tin này có thể có ích trong nhiều quyết định. Trong điều kiên kinh tế toàn cầu hóa cùng với những tiến bộ của công nghệ thông tin, một khối lượng lớn dữ liệu tài chính được tạo ra và lưu trữ. Khối lượng dữ liệu này có thể được khai thác nhằm phát hiện ra những dạng mẫu (patterns) ẩn và dự đoán về xu hướng trong tương lai và các động thái của thị trường tài chính. Với sự nhanh nhạy của kỹ thuật khai phá dữ liệu, các dữ liệu mới nhất có thể được sử dụng để tạo ra các thông tin quan trọng trong thời gian nhanh nhất. Điều này sẽ giúp cải thiện phản ứng thị trường và nhận thức thị trường, góp phần làm giảm chi phí va tăng doanh thu.

Những tiến bộ trong lĩnh vực kỹ thuật đã tạo điều kiện cho sự ra đời của những hệ thống dự báo nhanh nhạy và chính xác hơn. Những hệ thống này kết hợp giữa kỹ thuật Data mining và Business Intelligence như Case Based Reasoning (CBA) và mạng noron - Neutral Networks (NN). Sự kết hợp các hệ thống dự báo này với nhau cùng với một chiến lược giao dịch tốt tạo rất nhiều cơ hội kiếm lợi nhuận khổng lồ.

Giá trị của một sản phẩm tài chính phụ thuộc vào cả các yếu tố kinh tế vĩ mô và vi mô. Những thông tin này tồn tại dưới nhiều định dạng khác nhau. Data mining sẽ giúp khám phá ra những thông tin và dạng mẫu (patterns) ẩn từ những khối lượng dữ liệu lớn và dưới nhiều định dạng. Kỹ thuật NN và CBR có thể được áp dụng rộng rãi cho việc dự báo các nhân tố tài chính.

Mạng nơron được biết đến bởi khả năng học hỏi và khả năng cải thiện hiệu suất, tính năng qua thời gian. Mạng nơron cũng có khả năng khái quát hóa, tức là nhận biết được các vật thể mới tương tự nhưng không hoàn toàn giống như

các vật thể trước. Với khả năng rút ra được ý nghĩa từ các thông tin chưa chính xác, NN cũng được dùng để phát hiện ra các dạng mẫu (patterns) quá phức tạp đối với con người. NN đóng vai trò chuyên gia trong lĩnh vực mà con người được đào tạo để làm, có thể được dùng để dự báo về tình hình mới và hoạt động tại thời gian thực. Vì vậy, dữ liệu lịch sử về thị trường tài chính và những biến liên quan có thể được dùng đẻ huấn luyện NN trong việc mô phỏng thị trường. Dựa vào giá trị của các biến trên thị trường, NN có thể dự đoán trạng thái của ngày tiếp theo hoặc có thể dùng để đưa ra các khuyến cáo mua hay bán.

Phương pháp CBR dựa vào suy luận từ diễn biến lịch sử. Phương pháp này sử dụng một kho dữ liệu lớn dưới dạng các trường hợp (case) bao gồm nhiều biến số. Khi một case mới được đưa vào, thuật toán CBR sẽ dự đoán kết quả của case này dựa vào các case tồn tại trong kho lưu trữ. Kỹ thuật Data mining có thể được dùng để tìm ra các dạng mẫu (patterns) ẩn trong các case này hỗ trợ cho việc ra quyết định. Phương pháp CBR có thể được dùng trong thời gian thực để phân tích nhanh và giúp đưa ra quyết định tạo ra lợi nhuận kịp thời.

Vì vậy, kỹ thuật Data mining có thể được dùng kết hợp với thị trường tài chính để dự đoán diễn biến thị trường và hỗ trợ việc ra quyết định kinh doanh.

2.2.5. Quảng cáo và chăm sóc khách hàng

Trong môi trường tài chính cạnh tranh cao, những quyết định thông minh về marketing đang trở nên quan trọng hơn bao giờ hết nhằm thu hút, giữ khách hàng và cải thiện quan hệ khách hàng. Sự quan tâm khách hàng và các chiến lược marketing là cần thiết cho sự sống còn và thành công của doanh nghiệp. Data mining và phương pháp phân tích dự báo có thể hỗ trợ doanh nghiệp đưa ra các chiến lược này.

Các tổ chức tài chính gặp khó khăn trong việc tìm kiếm khách hàng không chủ động tiếp cận, vì thế các tổ chức này đang tích cực tiến hành các biện pháp marketing nhằm lôi kéo khách hàng từ các đối thủ cạnh tranh. Sự thiếu chắc chắn về khách hàng khiến việc lên kế hoạch cho các dịch vụ mới và việc sử dụng các phương tiện truyền thông gần như là điều không thể. Một phương pháp thường

dùng là áp dụng kiến thức chuyên môn chủ quan của con người làm quy tắc. Cho đến gần đây, việc thay thế con người bằng công nghệ kỹ thuật vẫn còn gặp nhiều khó khăn.

Một công cụ thú vị được sử dụng trong các tổ chức tài chính và marketing là công cụ phân tích thông tin khách hàng. Những phân tích và tính toán về chỉ báo chính giúp cho ngân hàng nhận diện những yếu tố ảnh hưởng đến nhu cầu của khách hàng trong quá khứ và tương lai.

Thông tin dữ liệu cá nhân của khách hàng cũng có thể đưa ra những dấu hiệu nhận tác động đến nhu cầu trong tương lai. Trong những trường hợp phân tích về các bên cho vay cá nhân và các doanh nghiệp nhỏ, những nhiệm vụ marketing thường bao gồm các yếu tố về chính khách hàng, hồ sơ tín dụng và xếp hạng của các tổ chức xếp hạng tín dụng bên ngoài.

Với những tiến bộ trong các công cụ Data mining và Business Intelligence, các ngân hàng đã có thể tăng cường thu hút khách hàng qua hình thức marketing trực tiếp hoặc thiết lập nhiều kênh tiếp xúc khách hàng, cải thiện phát triển khách hàng thông qua việc bán chéo hoặc bán thêm (up sell) sản phẩm và tăng độ thu hút khách hàng thông qua việc quản lý hành vi. Các ngân hàng có thể dùng dữ liệu sẵn có để giữ các khách hàng tốt nhất và để nhận diện các cơ hội có thể bán thêm dịch vụ. Có thể xây dựng hồ sơ của tất cả các tài khoản có giá trị và 5-10% tài khoản tốt nhất có thể được giao cho các quản lý khách hàng. Những người này sẽ nhận diện những cơ hội bán các sản phẩm cho các khách hàng này.

Cũng có thể nhóm nhiều nhiều sản phẩm vào thành gói đáp ứng nhu cầu của những khách hàng lớn. Data mining cũng có thể giúp ngân hàng điều chỉnh những phương pháp chào hàng khác nhau tùy trường hợp. Ví dụ, người ta có thể điều chỉnh các bức thư trực tiếp theo từng phân khúc của những người có tài khoản tại ngân hàng, Các ngân hàng cũng có thể nhận diện những khách hàng gặp vấn đề nhiều khả năng không trả được nợ trong tương lai thông qua tìm hiểu các hồ sơ trả nợ trong quá khứ và các mẫu thực tế trong dữ liệu sẵn có. Việc này

cũng giúp các ngân hàng điều chỉnh mối quan hệ với các khách hàng này để hạn chế rủi ro tương lai đến mức tối thiểu.

Data mining làm tăng tỷ lệ phản hồi trong chiến dịch gửi thư trực tiếp trong khi thời gian yêu cầu cho việc phân loại khách hàng giảm xuống. Điều này sẽ làm tăng thu nhập, cải thiện hiệu quả đội ngũ bán hàng trong nhóm mục tiêu. Data mining cũng giúpcác ngân hàng tối đa hóa danh mục dịch vụ và kênh phân phối của họ. Một sao kê các giao dịch trong quá khứ có thể cung cấp những thông tin hữu ích cho ngân hàng, và các chi nhánh/địa điểm khác nhau trong cùng một chi nhánh cũng có thể xuất hiện những mẫu mà nếu phát hiện ra có thể dùng dữ liệu quá khứ để học hỏi và làm cơ sở cho những hành động trong tương lai.

Kỹ thuật Data mining có thể trở nên vô cùng hữu ích cho các ngân hàng và tổ chức tài chính trong việc nhắm mục tiêu và giành được khách hàng một cách tốt hơn, phát hiện gian lận nhanh chóng (real time), cung cấp sản phẩm dựa trên các phân khúc để nhắm khách hàng mục tiêu tốt hơn, phân tích về diễn biến mua của khách hàng qua thời gian để giữ khách hàng và tạo mối quan hệ với khách hàng tốt hơn, phát hiện những xu hướng mới xuất hiện để chủ động hành động trong một thị trường có tính cạnh tranh cao, để bổ sung thêm giá trị vào những sản phẩm và dịch vụ hiện có và tung ra những gói sản phẩm và dịch vụ mới.

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).

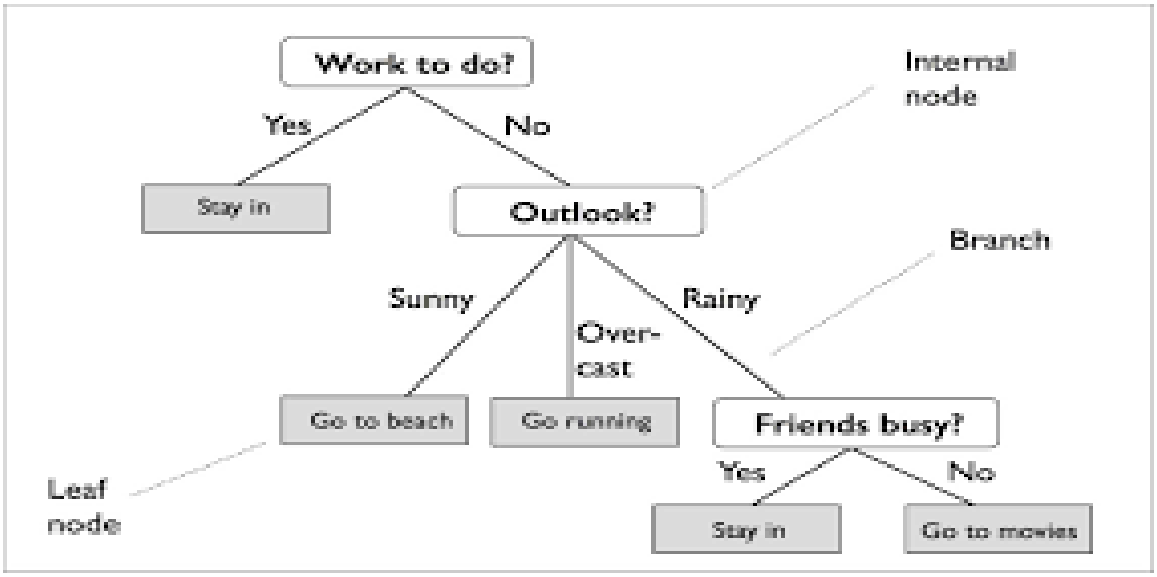
Mục tiêu của bài toán là phân loại khách hàng theo khả năng hoàn vốn dựa vào các thông tin đầu vào ban đầu khách hàng phải cung cấp từ đó dự báo được khách hàng này là khách hàng mục tiêu hay không. Việc dự báo chính xác sẽ giúp ngân hàng giảm thiểu các rủi ro có thể có từ các khách hàng có khả năng 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: Cây quyết định có thể dễ dàng sinh ra các quy tắc dạng If…Then… Else…. Hoặc các câu lệnh SQL. Đây chính là ưu điểm của phương pháp này. Với tập dữ liệu lớn khiến hình dạng của cây quyết định có thể phức tạp nhưng việc xây dựng các quy tắc là không khó

Đòi hỏi tiền xử lý dữ liệu đơn giản: Yêu cầu của các thuật toán phân lớp sử dụng cây quyết định không đòi hỏi xử lý dữ liệu đầu vào phức tạp

Khả năng xử lý cả thuộc tính liên tục và rời rạc: Cây quyết định xử lý "tốt" như nhau với thuộc tính liên tục và thuộc tính rời rạc. Tuy rằng với thuộc tính liên tục cần nhiều tài nguyên tính toán hơn. Những thuộc tính rời rạc đã từng gây ra những vấn đề với mạng neural và các kỹ thuật thống kê lại thực sự dễ dàng thao tác với các tiêu chuẩn phân chia (splitting criteria) trên cây quyết định:

mỗi nhánh tương ứng với từng phân tách tập dữ liệu theo giá trị của thuộc tính được chọn để phát triển tại node đó. Các thuộc tính liên tục cũng dễ dàng phân chia bằng việc chọn ra một số gọi là ngưỡng trong tập các giá trị đã sắp xếp của thuộc tính đó. Sau khi chọn được ngưỡng tốt nhất, tập dữ liệu phân chia theo test nhị phân của ngưỡng đó.

Thể hiện rõ ràng những thuộc tính tốt nhất: Các thuật toán xây dựng cây quyết định đưa ra thuộc tính mà phân chia tốt nhất tập dữ liệu đào tạo bắt đầu từ node gốc của cây. Từ đó có thể thấy những thuộc tính nào là quan trọng nhất cho việc dự đoán hay phân lớp.

Dễ dàng tính toán trong khi phân lớp: Mặc dù như chúng ta đã biết, cây quyết định có thể chứa nhiều định dạng, nhưng trong thực tế, các thuật toán sử dụng để tạo ra cây quyết định thường tạo ra những cây với số phân nhánh thấp và các test đơn giản tại từng node. Những test điển hình là: so sánh số, xem xét phần tửcủa một tập hợp, và các phép nối đơn giản. Khi thực thi trên máy tính, những test này chuyển thành các toán hàm logic và số nguyên là những toán hạng thực thi nhanh và không đắt. Đây là một ưu điểm quan trọng bởi trong môi trường thương mại, các mô hình dự đoán thường được sử dụng để phân lớp hàng triệu thậm trí hàng tỉ bản ghi.

- **Nhược điểm**

Dù có những sức mạnh nổi bật trên, cây quyết định vẫn không tránh khỏi có những điểm yếu. Đó là cây quyết định không thích hợp lắm với những bài toán với mục tiêu là dự đoán giá trị của thuộc tính liên tục như thu nhập, huyết áp hay lãi suất ngân hàng, … Cây quyết định cũng khó giải quyết với những dữ liệu thời gian liên tục nếu không bỏ ra nhiều công sức cho việc đặt ra sự biểu diễn dữ liệu theo các mẫu liên tục.

Dễ xảy ra lỗi khi có quá nhiều lớp: Một số cây quyết định chỉ thao tác với những lớp giá trị nhị phân dạng yes/no hay accept/reject. Số khác lại có thể chỉ định các bản ghi vào một số lớp bất kỳ, nhưng dễ xảy ra lỗi khi số ví dụ đào tạo

ứng với một lớp là nhỏ. Điều này xẩy ra càng nhanh hơn với cây mà có nhiều tầng hay có nhiều nhánh trên một node.

Chi phí tính toán đắt để đào tạo: Điều này nghe có vẻ mâu thuẫn với khẳng định ưu điểm của cây quyết định ở trên. Nhưng quá trình phát triển cây quyết định đắt về mặt tính toán. Vì cây quyết địnhcó rất nhiều node trong trước khi đi đến lá cuối cùng. Tại từng node, cần tính một độ đo (hay tiêu chuẩn phân chia) trên từng thuộc tính, với thuộc tính liên tục phải thêm thao tác xắp xếp lại tập dữ liệu theo thứ tự giá trị của thuộc tính đó. Sau đó mới có thể chọn được một thuộc tính phát triển và tương ứng là một phân chia tốt nhất. Một vài thuật toán sử dụng tổ hợp các thuộc tính kết hợp với nhau có trọng số để phát triển cây quyết định. Quá trình cắt cụt cây cũng "đắt" vì nhiều cây con ứng cử phải được tạo ra và so sánh.

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 này phát triển bắt đầu từ gốc, đến từng nhánh và phát triển quy nạp theo cách thức chia để trị cho tới khi đạt được cây quyết định với tất cả các lá được gán nhãn lớp.
- **Giai đoạn thứ hai cắt, tỉa bớt các cành nhánh trên cây quyết định:** Giai đoạn này nhằm mục đích đơn giản hóa và khái quát hóa từ đó làm tăng độ chính xác của cây quyết định bằng cách loại bỏ sự phụ thuộc vào mức độ lỗi (noise). Nghiên cứu các thuật toán phân lớp dữ liệu dựa trên cây quyết định của dữ liệu đào tạo mang tính chất thống kê, hay những sự biến đổi mà có thể là đặc tính riêng biệt của dữ liệu đào tạo. Giai đoạn này chỉ truy cập dữ liệu trên cây quyết định đã được phát triển trong giai đoạn trước và quá trình thực nghiệm cho thấy giai đoạn này không tốn nhiều tài nguyên tính toán, như với phần lớn các thuật toán, giai đoạn này chiếm khoảng dưới 1% tổng thời gian xây dựng mô hình phân lớp.

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: đi từ gốc, đến các nhánh, phát triển quy nạp theo hình thức chia để trị.
- 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]: Loại tiêu chuẩn này lựa chọn thuộc tính mà làm cực tiểu hóa độ không tinh khiết của mỗi phân chia. Các thuật toán sử dụng tiêu chuẩn này là CART, SLIQ, SPRINT.
- ***Information–gain*** [15]: Khác với Gini-index, tiểu chuẩn này sử dụng entropy để đo độ không tinh khiết của một phân chia và lựa chọn thuộc tính theo mức độ cực đại hóa chỉ số entropy. Các thuật toán sử dụng tiêu chuẩn này là ID3, C4.5.
- ***χ^2 -bảng thống kê các sự kiện xảy ra ngẫu nhiên***: χ^2 đo độ tương quan giữa từng thuộc tính và nhãn lớp. Sau đó lựa chọn thuộc tính có độ tương quan lớn nhất. CHAID là thuật toán sử dụng tiêu chuẩn này.

Do số lượng thuật toán về cây quyết định khá nhiều nên trong khuôn khổ luận văn chỉ trình bày về thuật toán C4.5 được sử dụng rộng rãi trong các ứng dụng tài chính ngân hàng.

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.

```
FormTree(T)
(1) ComputerClassFrequency(T);
(2) if OneClass or FewCases
        return a leaf;
        Create a decision node N;
(3) ForEach Attribute A
        ComputeGain(A);
(4) N.test=AttributeWithBestGain;
(5) if N.test is continuous
        find Threshold;
(6) ForEach T' in the splitting of T
(7) if T' is Empty
        Child of N is a leaf
    else
(8) Child of N=FormTree(T');
(9) ComputeErrors of N;
    return N
```

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.

Khi tập ví dụ là thuần nhất thì có thể nói: ta biết chắc chắn về giá trị phân loại của một ví dụ thuộc tập này, hay ta có lượng thông tin về tập đó là cao nhất. Khi tập ví dụ có độ pha trộn cao nhất, nghĩa là số lượng các ví dụ có cùng giá trị phân loại cho mỗi loại là tương đương nhau, thì khi đó ta không thể đoán chính xác được một ví dụ có thể có giá trị phân loại gì, hay nói khác hơn, lượng thông

tin ta có được về tập này là ít nhất. Vậy, điều ta mong muốn ở đây là làm sao chọn thuộc tính để hỏi sao cho có thể chia tập ví dụ ban đầu thành các tập ví dụ thuần nhất càng nhanh càng tốt. Vậy trước hết, ta cần có một phép đo để đo độ thuần nhất của một tập hợp, từ đó mới có thể so sánh tập ví dụ nào thì tốt hơn.

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

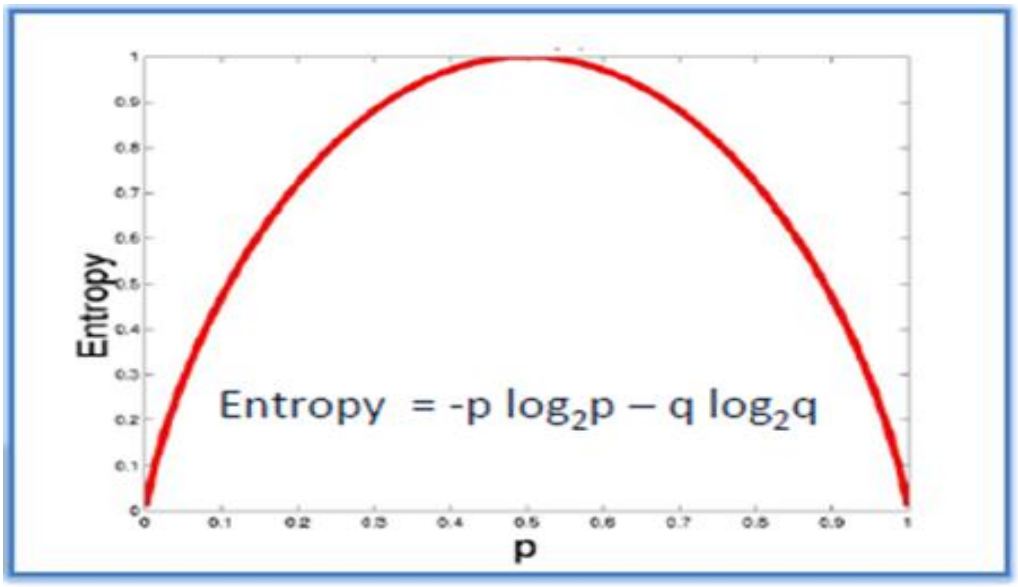
Khái niệm entropy của một tập S được định nghĩa trong lý thuyết thông tin là số lượng mong đợi các bit cần thiết để mã hóa thông tin về lớp của một thành viên rút ra một cách ngẫu nhiên từ tập S. Trong trường hợp tối ưu, mã có độ dài ngắn nhất. Theo lý thuyết thông tin, mã có độ dài tối ưu là mã gán –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:



Entropy = -0.5 $\log_2$0.5 – 0.5 $\log_2$0.5 = 1

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:

$$\text{Gain (S, A)} = \text{Entropy(S)} - \sum_{V \in \text{Value(A)}} \frac{|S_v|}{|S|} \text{Entropy}(S_V)$$

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:*** Luật khởi tạo ban đầu là đường đi từ gốc đến lá của cây quyết định. Một cây quyết định có một lá thì tương ứng tập luật sản xuất sẽ có một luật khởi tạo. Từng điều kiện trong luật được xem xét và loại bỏ nếu không ảnh hưởng tới độ chính xác của luật đó.
- ***Lựa chọn:*** Các luật đã cắt tỉa được nhóm lại theo giá trị phân lớp, tạo nên các tập con chứa các luật theo lớp. Sẽ có k tập luật con nếu tập training có k giá trị phân lớp. Từng tập con trên được xem xét để chọn ra một tập con các luật mà tối ưu hóa độ chính xác dự đoán của lớp gắn với tập luật đó.
- ***Sắp xếp:*** Sắp xếp K tập luật đã tạo ra từ trên bước theo tần số lỗi. Lớp mặc định được tạo ra bằng cách xác định các case trong tập training không chứa trong các luật hiện tại và chọn lớp phổ biến nhất trong các case đó làm lớp mặc định.
- ***Ước lượng, đánh giá:*** Tập luật được đem ước lượng lại trên toàn bộ tập training, nhằm mục đích xác định xem liệu có luật nào làm giảm độ chính xác của sự phân lớp. Nếu có, luật đó bị loại bỏ và quá trình ước lượng được lặp cho đến khi không thể cải tiến thêm.

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

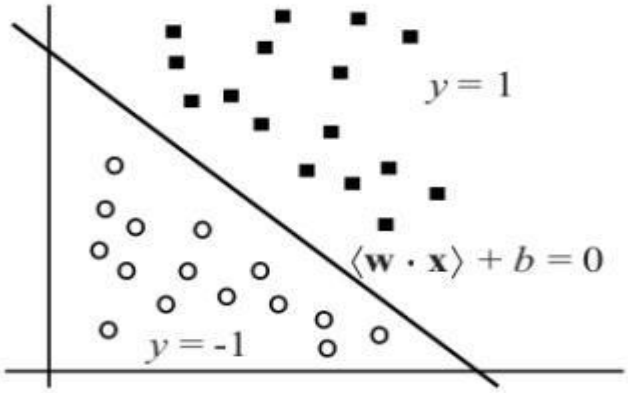
SVM –Support vector machine là một mô hình học có giám sát trong lĩnh vực học máy, SVM thường được dùng trong phân lớp dữ liệu (classification) và phân tích hồi quy (regression analysis). SVM là nền tảng cho nhiều thuật toán khai phá dữ liệu, SVM được giới thiệu bởi Vladimir Vapnik và các đồng sự vào 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:

$$\{xi, yi | i = 1,2,3 \ldots N | xi \in Rm\}$$

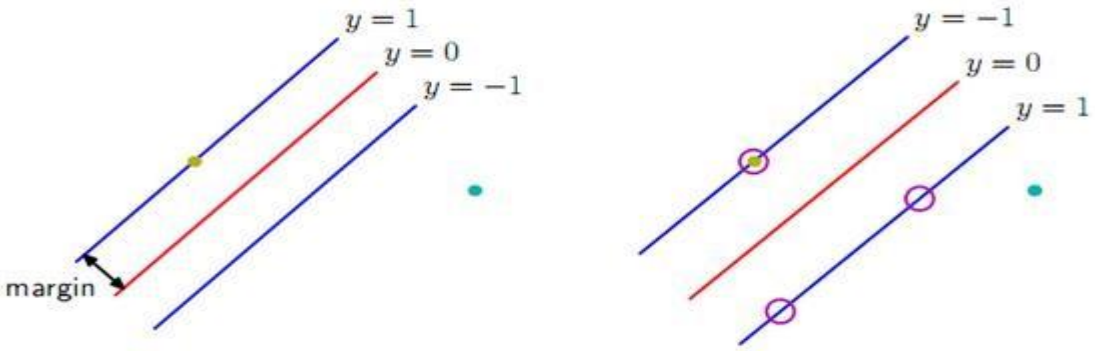
Trong đó mẫu là các vector đối tượng được phân lớp thành các mẫu dương và mẫu âm:

- 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

Trong trường hợp này, bộ phân lớp SVM là mặt siêu phẳng phân tách các mẫu dương khỏi các mẫu âm với độ chênh lệch cực đại, trong đó độ chênh lệch này gọi là lề (margin) xác định bằng khoảng cách mẫu dương và mẫu âm gần mặt siêu phẳng nhất. Mặt phẳng này được gọi là mặt siêu phẳng lề tối ưu.



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.

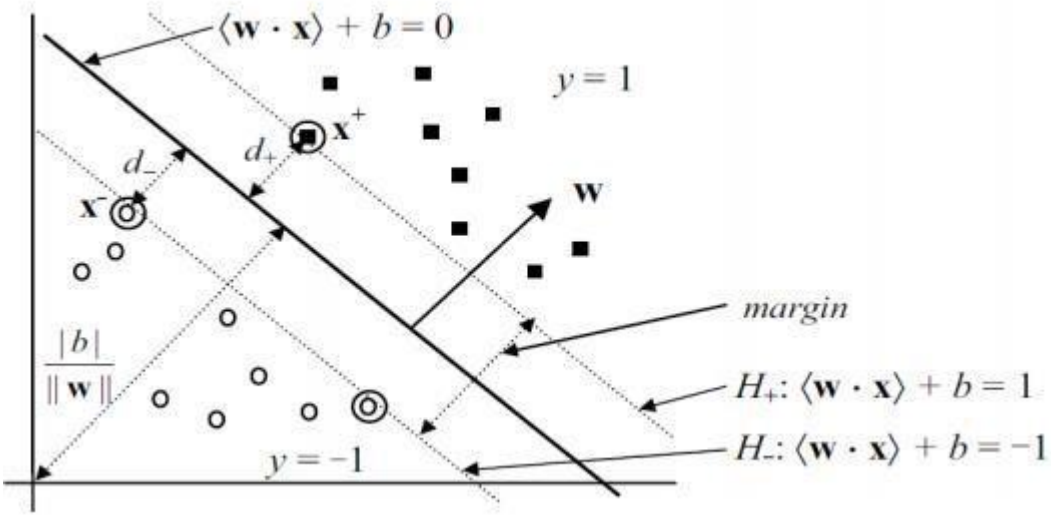
Bộ phân lớp nhị phân được xác định thông qua dấu của f(x):

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

- 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:

$$y_i(\mathrm{w}x_i + \mathrm{b}) \geq 1 - \xi_i \geq 0$$

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:

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

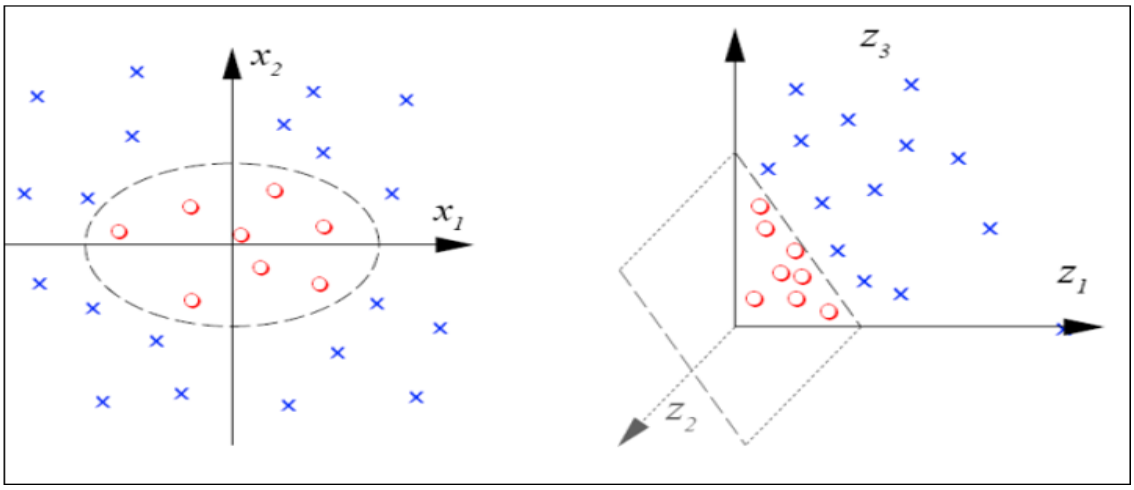
Với điều kiện:

$$y_i(\mathrm{w}x_i + \mathrm{b}) \geq 1 - \xi_i, \xi_i \geq 0$$

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

$$\Phi: \mathrm{X} \rightarrow \mathrm{F}$$

$$\mathrm{X} \rightarrow \Phi(\mathrm{x})$$



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:

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

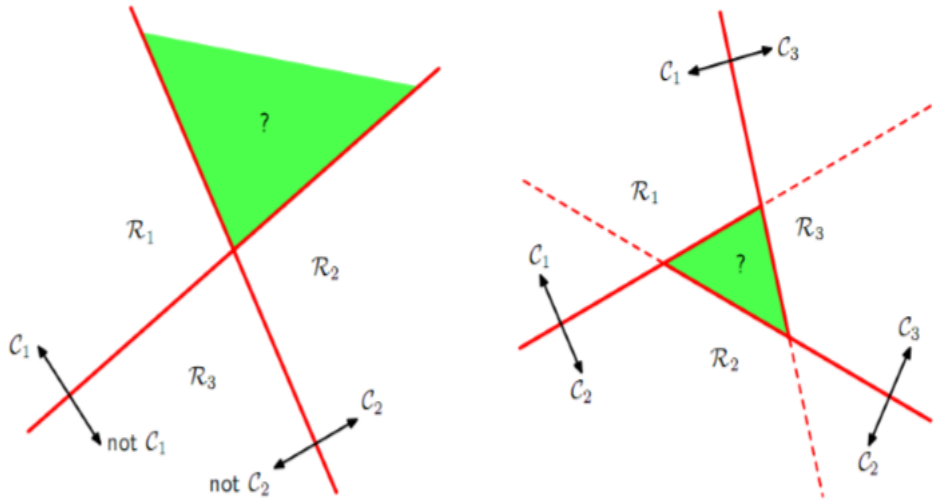
Gaussian RBF:

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

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.

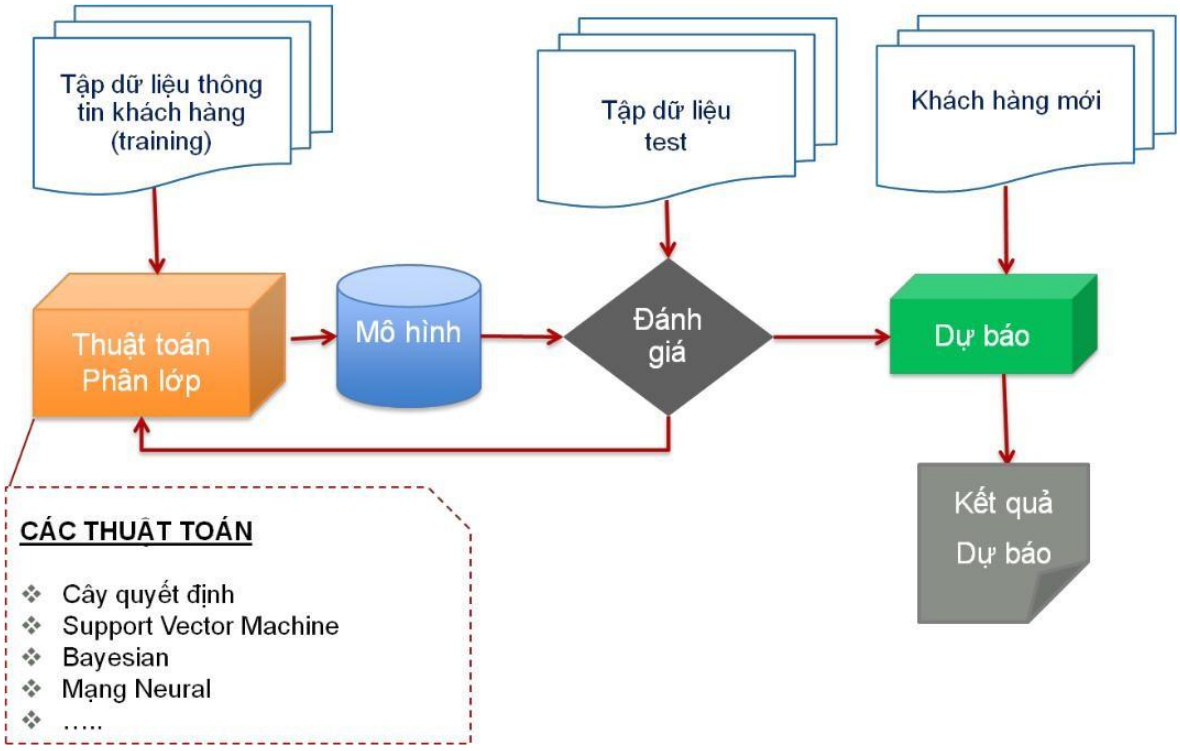
Cả 2 hướng đều dẫn đến vùng mập mờ trong phân lớp (như hình vẽ).

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

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

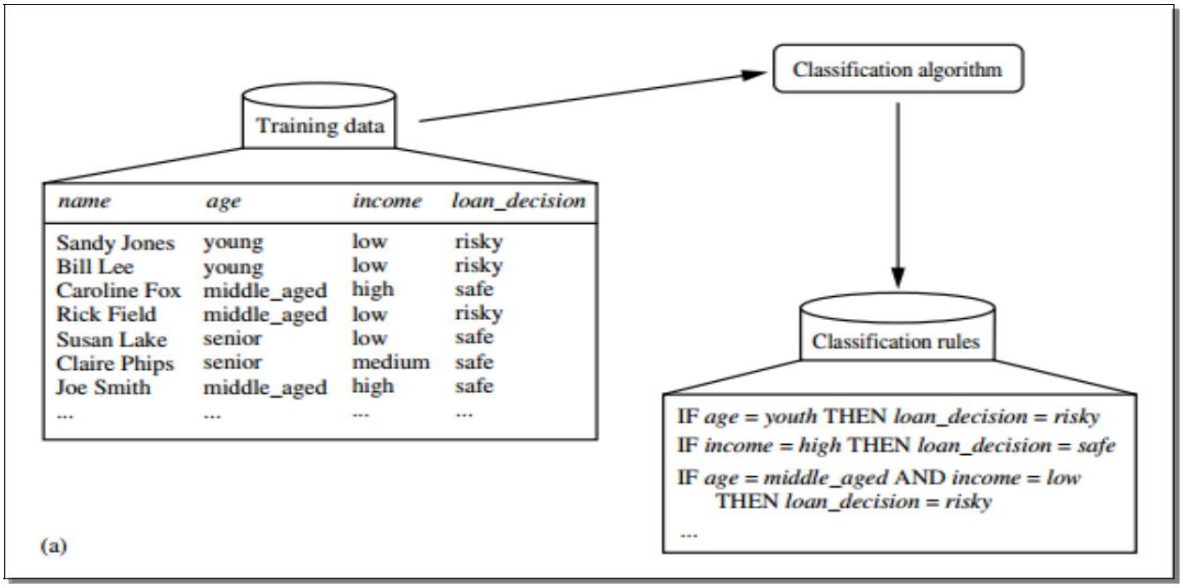
Và một điểm x được gán vào lớp C_k khi $y_k(x) > y_j(x)$ với mọi j ≠k.

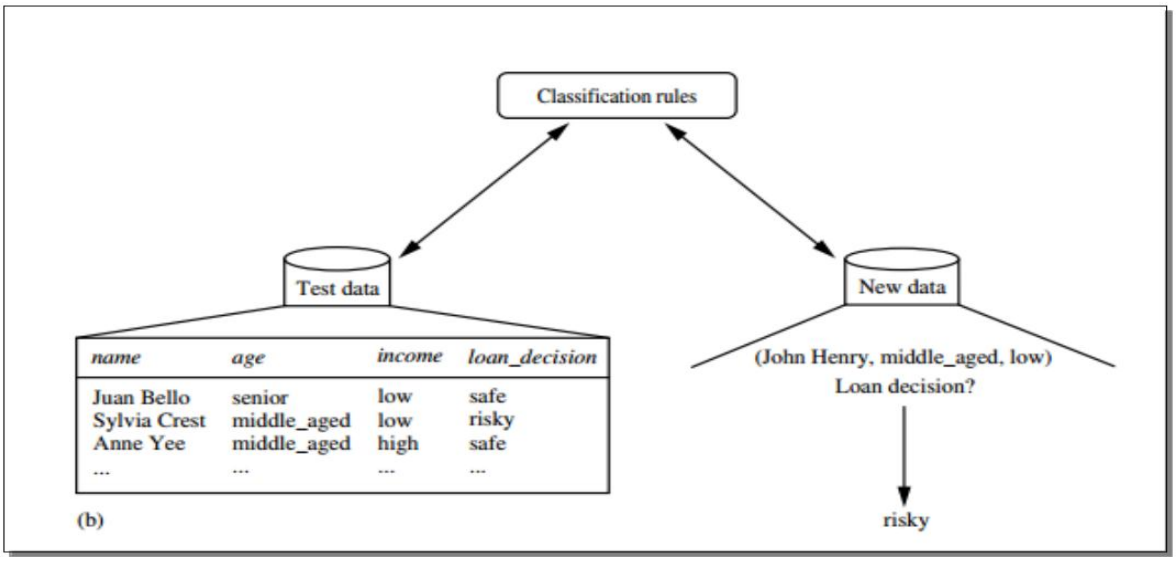
2.4. *Mô hình phân lớp dự báo rủi ro*



Hình 2.12: Mô hình phân lớp dự báo rủi ro

Quá trình phân lớp thực hiện nhiệm vụ xây dựng mô hình các công cụ phân lớp giúp cho việc gán nhãn phân loại cho các dữ liệu. Ví dụ nhãn "An toàn" hoặc "Rủi ro" cho các yêu cầu vay vốn; "Có" hoặc "Không" cho các thông tin thị trường… Các nhãn dùng phân loại được biểu diễn bằng các giá trị rời rạc trong đó việc sắp xếp trùng là không có ý nghĩa. Phân lớp dữ liệu gồm hai quá trình. Trong quá trình thứ nhất một công cụ phân lớp sẽ được xây dựng để xem xét nguồn dữ liệu. Đây là quá trình học, trong đó một thuật toán phân lớp được xây dựng bằng cách phân tích hoặc "học" từ tập dữ liệu huấn luyện được xây dựng sẵn bao gồm nhiều bộ dữ liệu. Một bộ dữ liệu X biểu diễn bằng một vector n chiều, X = (x1, x2, …, xn), đây là các giá trị cụ thể của một tập n thuộc tính của nguồn dữ liệu {A1, A2, …, An}. Mỗi bộ được giả sử rằng nó thuộc về một lớp được định nghĩa trước với các nhãn xác định.





Hình 2.13: Quy trình phân lớp

Có nhiều thuật toán phân lớp đã được nghiên cứu và phát triển như:

- Navie Bayes: nhanh đơn giản
- Supper Vector Machine: Hỗ trợ khai phá dữ liệu text và dữ liệu rộng
- Cây quyết định (Decision tree)
- Mạng nơron
- …

Trong khuôn khổ luận văn có giới hạn nên luận văn trình bày trình bày 2 kỹ thuật phân lớp: Phân lớp sử dụng cây quyết định bằng **thuật toán C4.5 và phân lớp sử dụng SVM.**

2.5. Kết luận chương 2

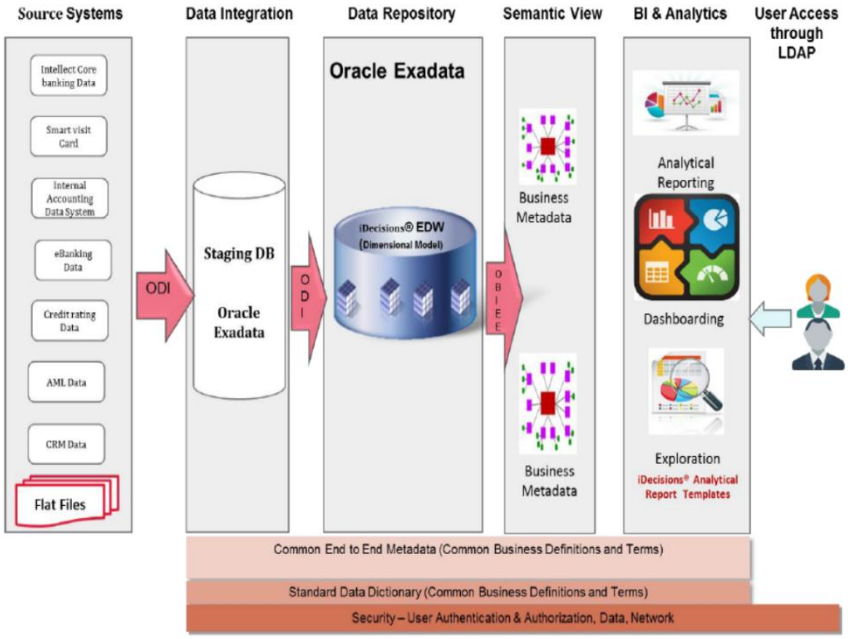
Qua tìm hiểu cơ sở lý thuyết về khai phá dữ liệu và ứng dung thực tiễn của khai phá dữ liệu trong hệ thống các công ty tài chính và ngân hàng cụ thể là áp dụng bài toán phân lớp dự báo rủi ro tín dụng, chúng ta đã hiểu được tầm quan trọng của việc ứng dụng khai phá dữ liệu vào công tác quản lý rủi do 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 như:

- **Model dữ liệu Khách Hàng**
- **Tiền Gửi**
- **Cho Vay**
- **Thẻ**
- **Thanh Toán (Thanh toán trong nước, Thanh toán quốc tế)**
- **Nguồn Vốn**
- **Dịch vụ**
- **Thu Phí**
- **…**



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):** Là vùng dữ liệu được đưa về từ tác hệ hệ thống nguồn hiện có của SHB đã được nêu trên bao gồm hệ thống xếp hạng tín dụng CSS**.**
- **Vùng lư trữ dữ liệu (Data Repository):** Dữ liệu được xử lý sạch tính toán và chuẩn hóa để đưa lên vùng dữ liệu chia theo chủ đề tiếp theo và lượng dữ liệu này sẽ được giải phóng và cuối ngày hôm sau.
- **Vùng dữ liệu theo chủ đề (Semantic Layer):** Tại đay dữ liệu sẽ được chia theo các nghiệp vụ mà ngân hàng SHB hiện tại kinh doanh bao gồm:
    - ✓ *Dữ liệu Khách Hàng*
    - ✓ *Tiền Gửi*
    - ✓ *Cho Vay*
    - ✓ *Thẻ*
    - ✓ *Thanh Toán (Thanh toán trong nước, Thanh toán quốc tế)*
    - ✓ *Nguồn Vốn*
    - ✓ *Dịch vụ*

- ✓ *Thu Phí*
- ✓ ...

Hiện tại giải pháp hệ thống Kho dữ liệu của SHB được xây dựng trên nền tảng của hãng Oracle với hạ tầng là Exadata X7-2 chuyên dụng với khả năng tối ưu cho việc phần tích và lưu trữ dữ liệu. Công cụ tích hợp dữ liệu dùng Oracle Data Intergration để tích hợp dữ liệu để thực hiện chuyển dữ liệu từ nguồn vào các lớp theo thiết kế tổng thể. Do dó hiện tại hệ thống kho dữ liệu của SHB đáp ứng đủ điều kiện để thực hiện bài toán phân lớp dự báo rủi ro được nêu ở Chương số 2. Cùng với lượng dữ liệu lịch sử đã được tổng hợp dài (Trên 5 năm).

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

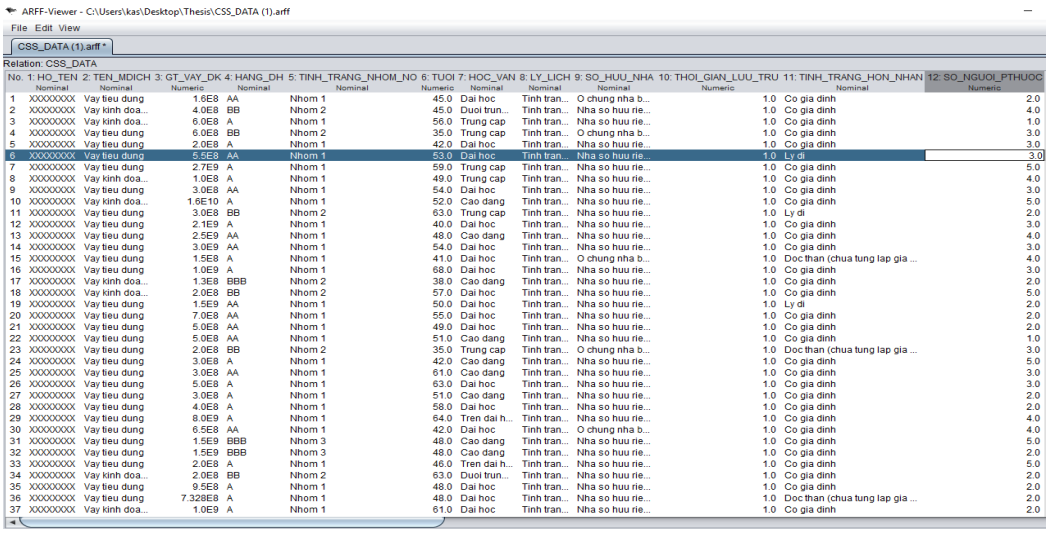
Hiện tại dữ liệu trong hệ thống kho dữ liệu của SHB bao gồm có các thông tin sao kê về các khoản vay của khách hàng và các giao dịch của khách hàng liên quan đến khoản vay như thời gian giải ngân, thời gian đáo hạn, lãi suất, tình trạng nhóm nợ của khách hàng. Với bài toán phân lớp dự báo rủi ro đã đề cập tại chương số 2 thì chỉ cần sử dụng dữ liệu lịch sử về tình trạng nhóm nợ của khách hàng. Dữ liệu tình trạng nhóm nợ của khách hàng được ghi nhận trên 5 giá trị rời rạc tương ứng với 5 nhóm nợ mà khách hàng có thể rơi vào căn cứ trên thời gian khách hàng trả nợ. Trong đó:

- **Nhóm 1:** nhóm nợ đủ tiêu chuẩn, nhóm khách hàng trả trả nợ đúng hạn trước khi tất toán khoản vay.
- **Nhóm 2:** nhóm nợ cần chú ý, nhóm khách hàng đã trả nợ nhưng quá hạn dưới 90 ngày
- **Nhóm 3:** nhóm nợ dưới tiêu chuẩn, bao gồm các khách hàng đã trả nợ nhưng quá hạn từ 90 ngày đến 180 ngày.

- **Nhóm 4:** nhóm nợ nghi ngờ, khách hàng đã trả nợ nhưng quá hạn từ 180 ngày đến 360 ngày. Việc thu hồi nợ từ những khách hàng này rất khó khăn
- **Nhóm 5:** nhóm khách hàng có khả năng mất vốn khi mà nợ quá hạn trên 360 ngày.

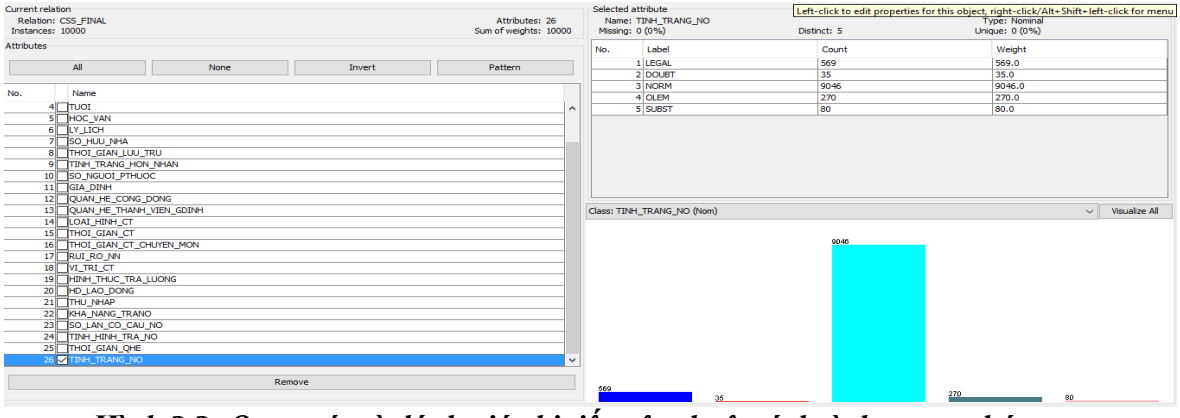
Phạm vi của luận văn cũng như bài toán đã được nêu ở chương số 2 chỉ thực hiện trên tập dữ liệu của khách hàng cá nhân không phải khách hàng cá nhân kinh doanh nên các chỉ tiêu về phương án đầu tư là không có giá trị. Vì vậy trước khi thực hiện thực nghiệm phải loại bỏ các trường không cần thiết này. 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:

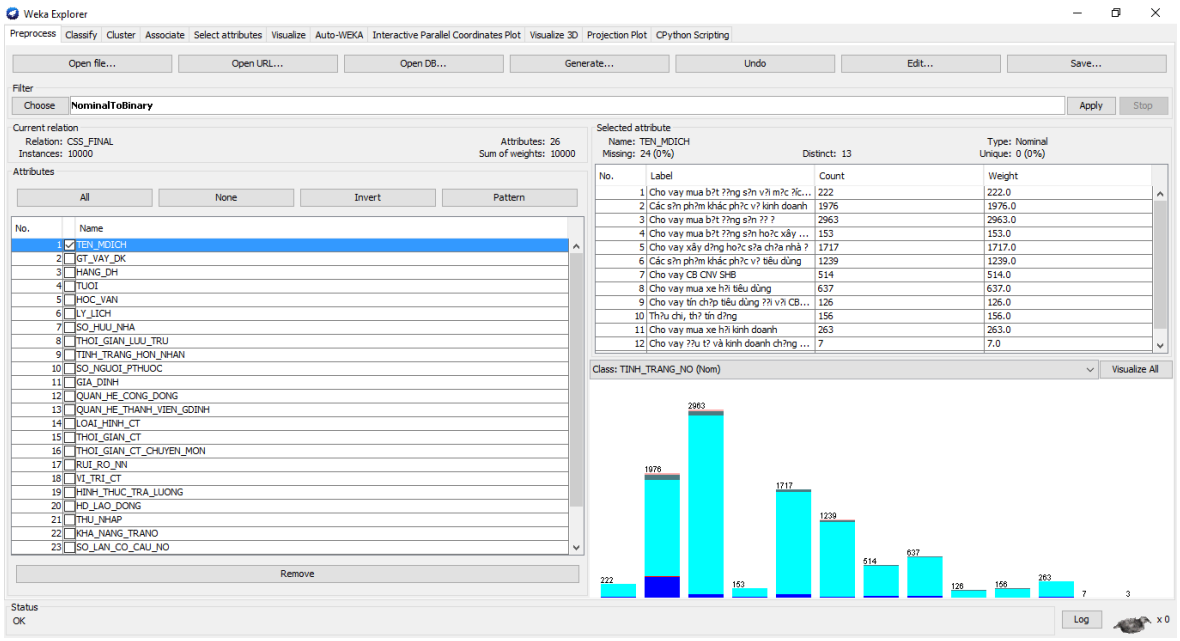


Hình số 3.2: Tập dữ liệu sử dụng làm mẫu.

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:



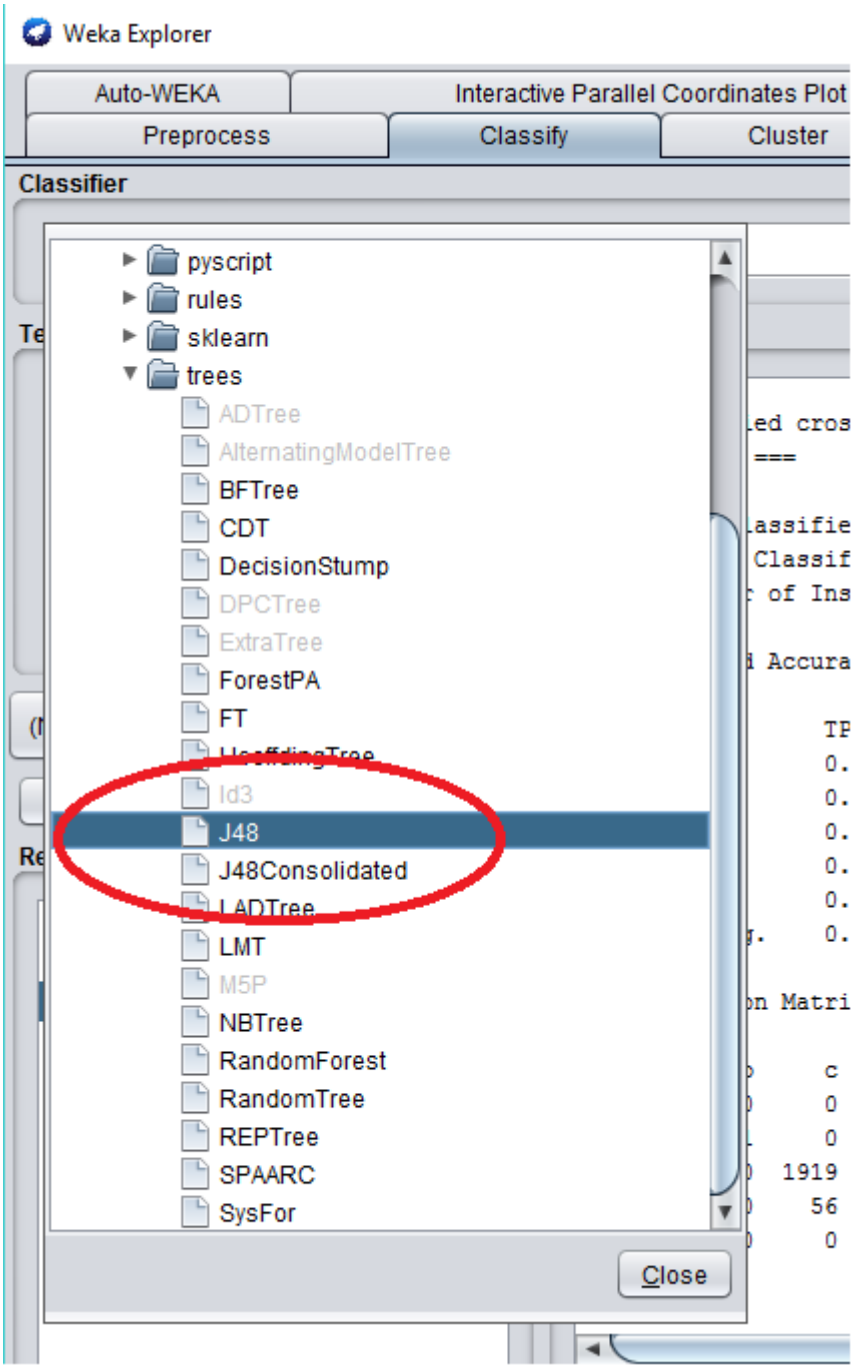
Hình 3.3: Quan sát và đánh giá chi tiết trên thuộc tính tình trạng nhóm nợ.



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.*

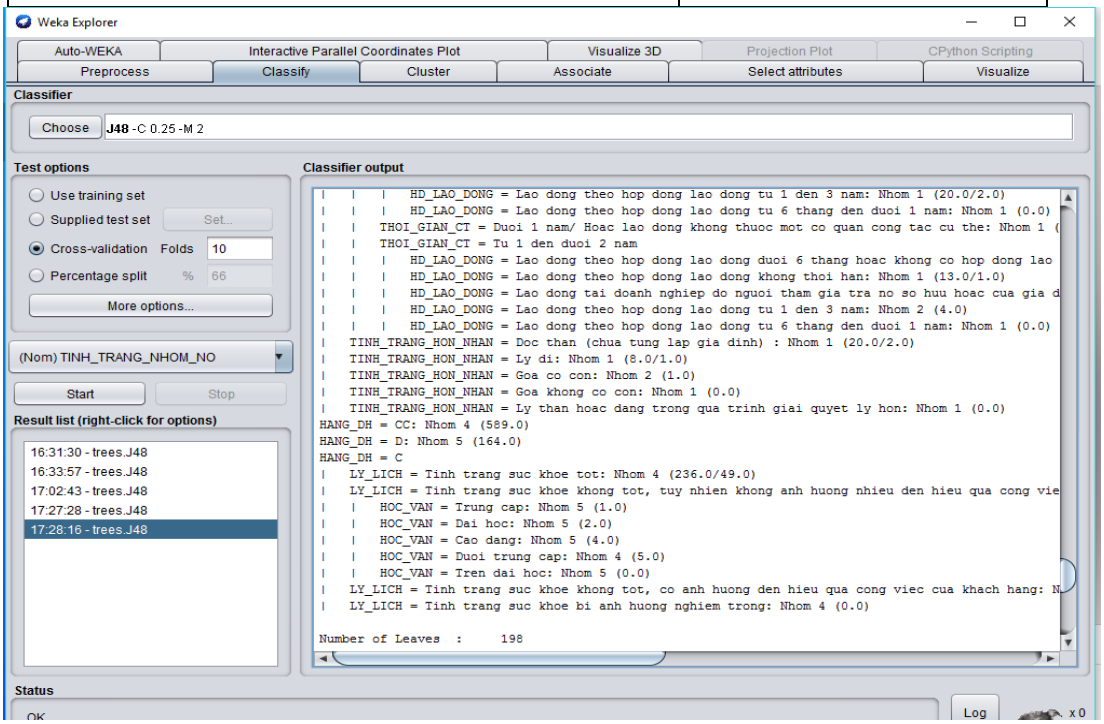
Sau khi điều chỉnh thông số và nghiên cứu ta chọn các giá trị tham số cho kết quả tốt nhất như sau:

- 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

Từ những bảng kết quả trên ta rút ra một số luật (IF - THEN) chú ý như:

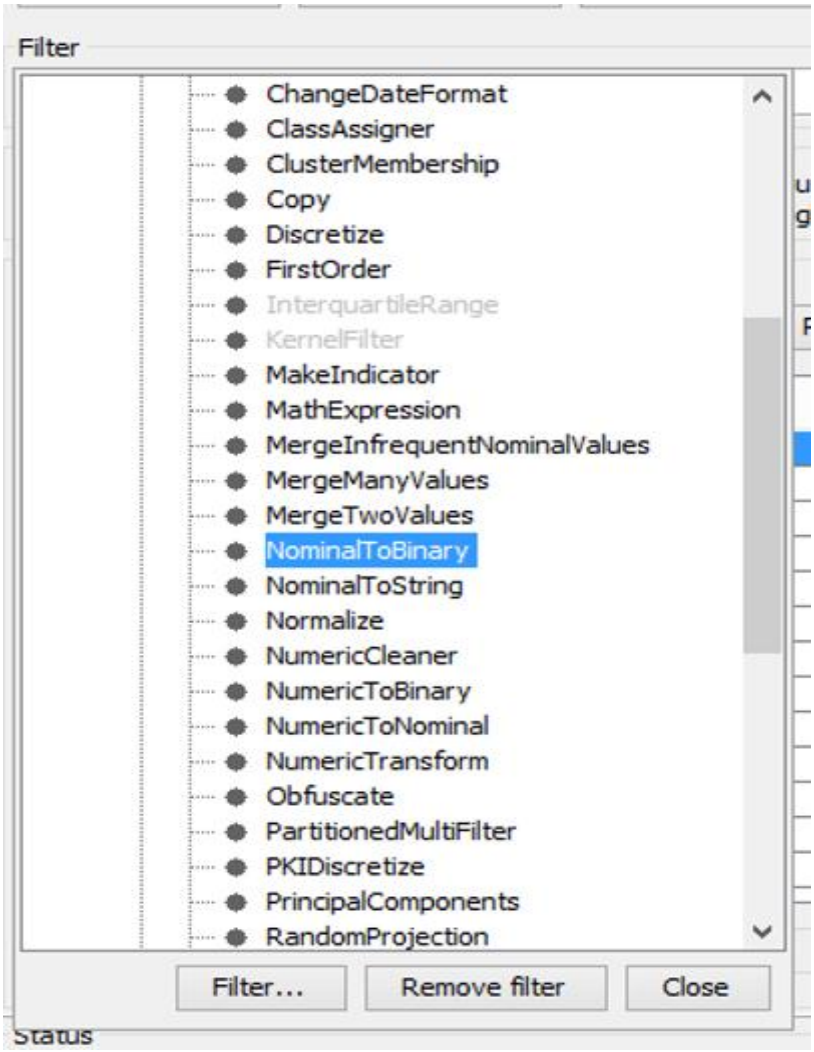
+ If khách hàng "Luôn trả nợ đúng hạn"AND "Học vấn là Đại học " AND "Số tiền vay lớn hơn 50 triệu " AND "Loại hình công ty đang làm việc: Cơ quan nhà nước" AND "Vị trí công tác: Cấp quản lý" AND "Giá trị vay < 140000000 " AND "Thời gian quan hệ với SHB <= 1" AND "Xếp hạng tín dụng AAA" THEN luôn trả đúng hạn (Nhom 1)

+ If khách hàng "Xếp hạng tín dụng C" AND "Tinh trang suc khoe khong tot, tuy nhien khong anh huong nhieu den hieu qua cong viec khach hang" AND "Học vấn trung cấp" THEN Thuộc nhóm nợ không tiếp nhận duyệt hồ sơ (nhóm 5)

c. Phân lớp sử dụng thuật toán SVM

Với cùng tập dữ liệu đã được sử dụng với thuật toán phân lớp cây quyết định áp dụng thuật toán C4.5 tuy nhiên đối với phân lớp áp dụng thuật toán SVM ta phải thêm một bước xử lý dữ liệu đối với các thuộc tính kiểu Nominal. Phải chuyển đổi dữ liệu kiểu Nomial sang kiểu Numeric. Xử lý dữ liệu kiểu Nominal sang kiểu Numeric để phù hợp với bài toán phân lớp SVM bằng cách mở tệp dữ liệu traning và sử dụng chức năng chuyển đổi dữ liệu từ Nominal sang Binary như sau:

Sử dụng chức năng Filter dữ liệu trên Weka Explore chọn Nominal to Binary, Sau khi chuyển đội từ bộ dữ liệu với 24 thuộc tính trở thành bộ dữ liệu có 155 thuộc tính với duy nhất kiểu Numeric để sử dụng trong thuật toán SVM



Hình 3.6: Bộ chuyển đổi từ Nominal sang kiểu Binary

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 =10

Tham số thuật toán: **SVM type= C-SVM (Classification)**

Sử dụng hàm nhân: Gaussian RBF

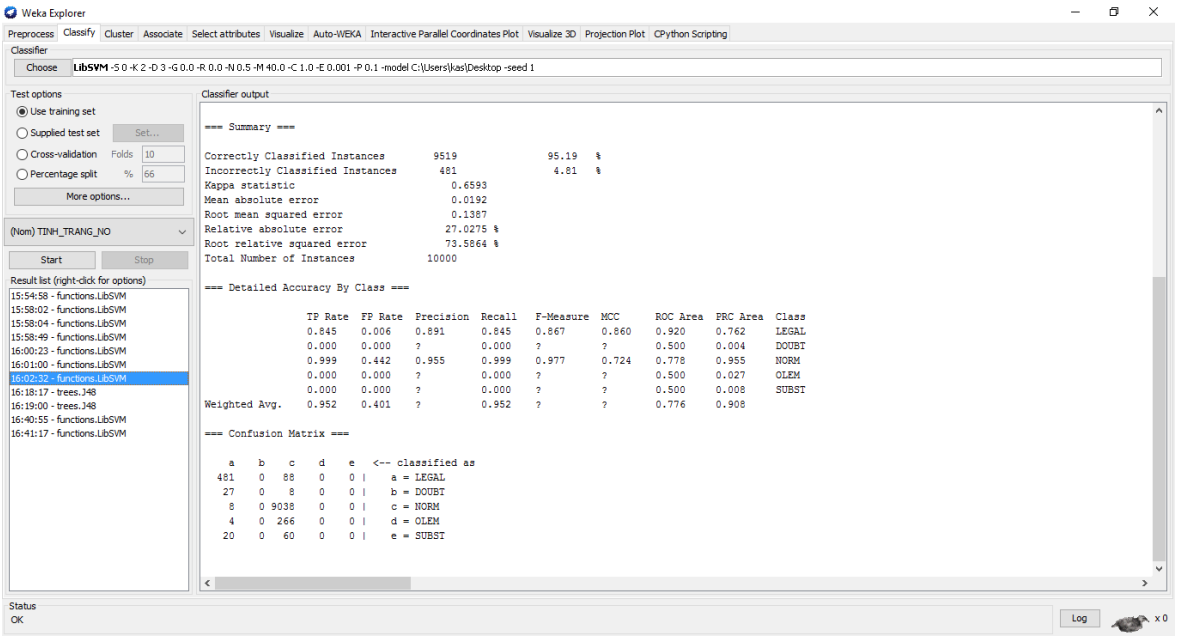
gama=0.5

cost(C) =0.5

Kết quả của xây dựng hình phân lớp bằng cây quyết định trên tập dữ liệu training 10000 mẫu

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**

So sánh 2 thuật toán:

Với mỗi thuật toán đều có kết quả tương tự về tổng số mẫu dự đoán đúng xấp xỉ 63877 mẫu so với tổng 65411 mẫu training tương đương hơn 97% phân lớp đúng

Về thời gian xây dựng mô hình và training thì thuật toán sử dụng cây quyết định áp dụng thuật toán C4.5 cho kết quả vượt trội khi mà chỉ cần sử dụng hơn 2.02 giây để cho kết quả xây dựng mô hình trong khi thuật toán sử dụng SVM thì mất hơn 400 giây để xây dựng mô hình từ tập dữ liệu

Kết quả cho thấy thuật toán SVM cho kết quả phân lớp tốt hơn trong các lớp so với thuật toán cây quyết định. Cụ thể trong kết quả thuật toán SVM có 3 lớp có độ tin cậy tương đối cao phù hợp với phân bố dữ liệu trong khi thuật toán cây quyết định chỉ có 2 lớp có độ tin cậy cao.

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.

DANH MỤC TÀI LIỆU THAM KHẢO

[1] PGS.TS. Nguyễn Văn Hiệu, Ngân hàng VietTinBank, Cơ chế điều chỉnh tự động/bán tự động trong hoạt động rủi ro ,
https://www.vietinbank.vn/web/home/vn/research/14/co-che-dieu-chinh-tu-dongban-tu-dong-trong-quan-tri-rui-ro-tin-dung.html

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

[3] TS. Nguyễn Minh Kiều, Nghiệp vụ ngân hàng, Trường Đại Học Kinh Tế TPHCM. Nhà xuất bản Thống Kê, tháng 12-2005.

[4] Lê Thùy Dương , Khóa luận tốt nghiệp "Bài toán phân lớp văn bản và áp dụng phân lớp dữ liệu tài chính ngân hàng"- Đại học công nghệ, ĐHQGHN, K50

[5] Nguyễn Hà Nam, Giáo trình Khai phá dữ liệu, ĐHQG Hà Nội, 2013.

[6] Quyet-dinh-1253-QD-NHNN-nghiep-vu-phan-tich-xep-hang-tin-dung-doanh-nghiep-178771.

[7] Ngân hàng Nhà nước Việt Nam,http://www.sbv.gov.vn/

[8] Quyết định 493/2005/QĐ-NHNN ngày 22/04/2005 của Thống đốc Ngân hàng Nhà nước

[8] J.Han and M.Kamber, Data Mining, "Concepts and Techniques", Morgan Kaufmann, 3rd Edition, 2011.

[9] Jiawei Han and Micheline Kamber, Data Mining Concepts and Techniques, Second Edition

[10] Carlo Vercellis, Business Intelligence: Data Mining and Optimization for Decision Making, 2009 John Wiley & Sons, Ltd. ISBN: 978-0-470-51138-1

[11] Ron Kohavi, J. Ross Quinlan, Decision Tree Discovery, 1999

[12] Ian H. Witten, Eibe Frank, Data Mining Practical Machine Learning Tools and Techniques, 3rd Edition, Mark A. Hall

[13] S.Prabhu, N.Venkatesan, Data mining and warehouse, New Age International (P) Limited Publishers, 2007

[14] Tom M. Mitchell (1997), Machine Learning, McGraw-Hill

[15] Ron Kohavi, J. Ross Quinlan, Decision Tree Discovery, 1999