ỨNG DỤNG MỘT SỐ PHƯƠNG PHÁP XÂY DỰNG HÀM PHÂN LOẠI TRONG CẢNH BÁO SỚM NGUY CƠ VỠ NỢ CỦA CÁC NGÂN HÀNG THƯƠNG MẠI CỔ PHẦN VIỆT NAM

Nguyễn Thị Lan*, Đỗ Thị Nhâm, Ngọc Minh Châu, Lê Văn Hỗ

Khoa Công nghệ thông tin, Học viện Nông nghiệp Việt Nam

Tác giả liên hệ: ngtlan@vnua.edu.vn

Ngày gửi bài: 06.03.2018 Ngày chấp nhận: 21.08.2018

TÓM TẮT

Trong nghiên cứu này chúng tôi vận dụng các mô hình thống kê dựa trên phân tích khác biệt đa biến, hồi qui logistic và máy vecto hỗ trợ (SVM) để xây dựng các hàm phân loại nhằm cảnh báo rủi ro sớm cho các ngân hàng thương mại cổ phần (NHTMCP) Việt Nam. Các mô hình được thực hiện trên các nhóm thuộc tính như: khả năng sinh lời, các chỉ số thâm hụt, hiệu quả quản lí tài sản, chất lượng tài sản, mức độ an toàn, nhóm chỉ số tăng trưởng bền vững và tính thanh khoản. Nghiên cứu tính toán độ chính xác của các mô hình nghiên cứu trên cả tập dữ liệu và kiểm tra, ngoài ra còn đưa ra các loai sai lầm loại I, sai lầm loại II mà các mô hình mắc phải

Từ khóa: Ngân hàng thương mại, cảnh báo nguy cơ vỡ nợ, hàm phân loại.

Application of Some Methods for Building Classification Functions in Early Warning of Default Risk for Vietnam Joint Stock Commercial Banks

ABSTRACT

In our study, we used statistical models based on multivariate linear discriminant analysis, logistic regression and SVM methods to construct bank classification functions for early risk warning for Vietnam joint stock commercial banks The models were built on attribute groups such as profitability, deficit indicators, asset management efficiency, asset quality, safety level, sustainable growth rate and liquidity. The study calculates the accuracy of the research models on both data sets and tests, in addition to the types of mistakes of type I, mistakes of type II that models suffer from.

Keywords: Commercial banks, early warning, default risk, classification function.

1. ĐẶT VẤN ĐỀ

Với tư cách là trung gian tài chính, ngân hàng thương mại là loại hình doanh nghiệp kinh doanh đặc thù vì kinh doanh các loại hàng hóa đặc biệt là tiền tệ, vàng bạc, chứng khoán,… và cung ứng các dịch vụ ngân hàng theo quy định của pháp luật. Hiện nay, ở Việt Nam đang có sự phát triển nhanh chóng của hệ thống ngân hàng dẫn đến việc thành lập hàng loạt các ngân hàng và các chi nhánh mới. Hội nhập kinh tế quốc tế đem lại nhiều cơ hội nhưng cũng không ít rủi ro cho hệ thống ngân hàng như: dễ bị phá sản, thiếu vốn để cạnh tranh, thua lỗ và mất thị phần. Việc đánh giá một doanh nghiệp nói chung đã rất khó khăn, phức tạp, đánh giá một ngân hàng với nhiều nét đặc thù riêng còn khó khăn và phức tạp hơn nhiều. Nếu chỉ áp dụng cách đánh giá thông thường dựa trên phân tích báo cáo tài chính sẽ không giúp nhiều cho việc phát hiện sớm nguy cơ vỡ nợ, yếu kém của các ngân hàng, điển hình như hàng loạt vụ sụp đổ của các ngân hàng lớn trên thế giới trong những năm gần đây như Lehman Brothers, Washington Mutual (2008). Tại nước ta, trong năm 2010 và 2011 nhiều tổ chức ngân hàng rơi vào tình trạng mất thanh khoản nghiêm trọng, kết quả cuối năm 2011, một số ngân hàng phải

sáp nhập, hợp nhất (ba ngân hàng Đệ Nhất, Sài Gòn và Tín nghĩa ngân hàng đã hợp nhất và chính thức hoạt động dưới tên Ngân hàng TMCP Sài Gòn kể từ 01/01/2012) và chịu sức ép tái cấu trúc lại để phù hợp với xu hướng hiện tại. Tất cả những vấn đề trên đã không được phản ánh và cảnh báo sớm thông qua các kênh dự báo, phân tích thông thường (Nguyễn Lê Thành, 2012).

Trên thế giới, để giảm thiểu rủi ro, năm 1988, Ủy ban Basel về giám sát ngân hàng ban hành hệ thống đo lường vốn và rủi ro tín dụng với tên thường gọi là hiệp ước Basel 1. Theo yêu cầu của Basel 1, các ngân hàng phải duy trì tỉ lệ vốn bắt buộc trên tổng số tài sản điều chỉnh theo hệ số rủi ro (CAR) ở mức an toàn là 8%. Do những hạn chế của Basel 1, năm 2004 Ủy ban Basel lại giới thiệu phiên bản mới với tên gọi Basel 2, có hiệu lực từ năm 2007 và kết thúc thời gian chuyển đổi đến năm 2010 (Lê Thanh Ngọc và cs., 2015). Từ những năm 70 của thế kỉ trước, mô hình CAMELS (Capital adequacy, Aset Quality, Management, Earnings, Liquidity, Sensitivity to market risk) là hệ thống xếp hạng, giám sát tình hình ngân hàng Mỹ và được coi là chuẩn mực với hầu hết các tổ chức trên toàn thế giới khi đánh giá hiệu quả rủi ro của các ngân hàng nói riêng và các tổ chức tín dụng nói chung. Tuy nhiên nếu chỉ đơn thuần áp dụng mô hình CAMELS để phân tích thì bức tranh đầy đủ về "sức khỏe" của các tổ chức tín dụng sẽ chưa thực sự rõ nét (Nguyễn Lê Thành, 2012). Mà trong phân tích tổ chức tín dụng theo phương pháp hiện đại, ngoài nền tảng cơ bản là các yếu tố tài chính từ kết quả của mô hình CAMELS, cần bổ sung các yếu tố phi tài chính, các yếu tố xuất phát từ quan hệ với đối tác kinh doanh để có cái nhìn toàn diện. Do tính cấp thiết của việc cảnh báo nguy cơ vỡ nợ của các ngân hàng, trên thế giới đã có nhiều tác giả đưa ra các phương pháp khác nhau nhằm phục vụ việc cảnh báo một cách tốt nhất (Aziz & Humayon *et al.*, 2006). Altman (1968) là người đầu tiên sử dụng mô hình thống kê đa biến "Z-core" để tìm sự kết nối giữa các chỉ số tài chính để có thể cảnh báo nguy cơ vỡ nợ. Ohlson (1980) đưa ra mô hình khác: mô hình logistic và mô hình số 8 để cảnh báo. Một số mô hình thống kê, chẳng hạn như phân tích khác biệt, phân tích logistic và hồi qui probit đã được sử dụng bởi các nhà nghiên cứu rủi ro tín dụng (Aziz & Humayon *et al.*, 2006). Một trong những nghiên cứu gần đây ở lĩnh vực này là Lacerda & Moro (2008), họ đã phân tích nguy cơ phá sản của các công ty Bồ Đào Nha với ba phương pháp: hồi qui logistic, phương pháp biệt số và máy vecto hỗ trợ. Gần đây, với sự xuất hiện của cây ra quyết định và mạng noron (Le Cun, 1986), kĩ thuật trí tuệ nhân tạo (AI) được sử dụng rộng rãi cho tín dụng, chúng có hiệu quả vượt trội so với thống kê truyền thống về kết quả đầy hứa hẹn. Mặc dù những mô hình này vấp phải vấn đề cực tiểu địa phương và học quá (Wei & Lichen, 2000). Việc tìm ra các phương pháp mới cải tiến các phương pháp cũ, ứng dụng nhiều loại số liệu là đòi hỏi cấp bách hiện nay trên thế giới.

Ở nước ta việc xây dựng các mô hình cảnh báo vỡ nợ ngân hàng đã được nghiên cứu nhưng chưa đầy đủ, chưa theo diễn biến nợ xấu của các ngân hàng trong một thời kì nhất định (Đặng Huy Ngân, 2016). Một vài tác giả đã đưa ra các bài toán cảnh báo sớm nguy cơ vỡ nợ. Đặng Huy Ngân (2015),đã sử dụng kết hợp phân tích nhân tố và hồi qui logistic để phân loại các ngân hàng thương mại Việt Nam. Nguyễn Quang Dong (2009) đãxếp hạng tín dụng các ngân hàng, tổ chức tài chính Việt Nam bằng phương pháp phân tích tách biệt. Các nghiên cứu trước đây đã xác định các nhân tố tác động đến nguy cơ vỡ nợ, nhưng chúng có phải là nguyên nhân dẫn đến nợ xấu trong thời kì đó không? Hơn nữa các ngân hàng có đặc trưng riêng có ảnh hưởng tới khả năng vỡ nợ hay không? Đề cập tới vấn đề này, Đặng Huy Ngân (2018) đã nghiên cứu xây dựng mô hình cảnh báo nguy cơ vỡ nợ cho các NHTMCP Việt Nam với số liệu mảng, mô hình logit, đồng thời cũng thử nghiệm mô hình mạng noron vào phân loại. Trong nghiên cứu này chúng tôi sẽ vận dụng các mô hình thống kê dựa trên phân tích khác biệt, hồi qui logistic và SVM để xây dựng các hàm phân loại ngân hàng, từ đó đưa ra độ chính xác của các mô hình; giá trị các loại sai lầm loại I, loại II mà mỗi mô hình mắc phải và giá trị *p-value* cho so

sánh hiệu suất các mô hình để từ đó kết luận hiệu suất của chúng có khác biệt nhiều không.

2. PHƯƠNG PHÁP NGHIÊN CỨU

2.1 Nguồn số liệu và biến số

Các số liệu thử nghiệm trong nghiên cứu của chúng tôi đượclấy từ báo cáo tài chính công khai đã được kiểm toán (Bảng cân đối kế toán, bảng báo cáo lưu chuyển tiền tệ, bảng kết quả hoạt động kinh doanh) tại thời điểm cuối năm của các ngân hàng thương mại cổ phần Việt Nam từ năm 2009 đến hết năm 2012, tổng cộng có 136 quan sát (Đặng Huy Ngân, 2018). Các biến trong nghiên cứu gồm:

Bảng 1. Các biến số nghiên cứu đã lựa chọn

Tên biến	Nội dung	Dấu kỳ vọng
Nhóm 1: Khả năng sinh lời		
e1	ROA-Khả năng sinh lời trên tổng tài sản	-
e2	ROE-khả năng sinh lời trên vốn chủ sở hữu	-
e3	Chi phí dự phòng nợ khó đòi + Giảm giá đầu tư chứng khoán/Thu nhập lãi thuần	+
e4	(Lãi thuần - Chi phí hoạt động)/Chi phí hoạt động	-
e5	Lợi nhuận sau thuế/Thu nhập lãi thuần	-
e6	Tổng thu nhập/Tổng tài sản có	-
e7	Tốc độ tăng trưởng thu nhập lãi thuần	-
e8	Tốc độ tăng trưởng lợi nhuận sau thuế	-
e9	Chi phí dự phòng nợ khó đòi/Tổng thu nhập trước dự phòng và thuế	+
e10	Thu nhập từ phí dịch vụ/Tổng thu nhập hoạt động	-
e11	Lãi cận biên thuần	-
Nhóm 2: Các chỉ số thâm hụt		
d1	Tổng nợ/Tài sản có	+
d2	Tổng nợ/Vốn chủ sở hữu	+
d3	Nợ quá hạn/Tổng nợ	+
Nhóm 3: Hiệu quả quản lý tài sản		
m1	Thu nhập lãi thuần/Tài sản cố định	-
m2	(Lợi nhuận trước thuế + Dự phòng)/Chi phí hoạt động	-
m3	Thu nhập lãi thuần/Tổng tài sản có	-
m4	(Lợi nhuận trước thuế + Dư phòng)/Tổng tài sản có	-
Nhóm 4: Chất lượng tài sản		
a1	Dự phòng nợ khó đòi/Nợ khó đòi	+
a2	Dự phòng nợ khó đòi/Dư nợ cho vay	+
a3	Nợ khó đòi/(Vốn chủ sở hữu + Dự phòng nợ khó đòi)	+
a4	Tỷ lệ cho vay/Tài sản sinh lời	+/-
a5	Gửi và cho vay tiền thị trường liên ngân hàng/Tài sản sinh lời	-
a6	Chứng khoán đầu tư và chứng khoán kinh doanh/Tài sản sinh lời	+
a7	Đầu tư góp vốn dài hạn/Tài sản sinh lời	+
Nhóm 5: Mức độ an toàn		
c1	Tăng trưởng vốn chủ sở hữu	-
c2	CAR - tỷ lệ an toàn vốn	-
c3	Vốn chủ sở hữu/Tổng huy động vốn tiền gửi	-
c4	Vốn chủ sở hữu/Tài sản có	-
Nhóm 6: Các chỉ số về tăng trưởng bền vững		
s1	Tốc độ tăng trưởng thu nhập lãi	-
s2	Ln (Tài sản cố định)	+/-
s3	Tốc độ tăng trưởng tài sản (Tolal asser growth)	+/-
s4	Lợi nhuận chưa phân phối/Lợi nhuận sau thuế	-
s5	Lợi nhuận chưa phân phối/Tổng tài sản	-
Nhóm 7: Tính thanh khoản		
11	Tốc độ tăng trưởng tiền gửi	-
12	Tốc độ tăng trưởng các khoản vay	+
13	Các khoản vay thuần/Tiền gửi của khách	+
14	Huy động từ tổ chức kinh tế và dân cư/Tổng huy động	-
15	Huy động trên thị trường liên ngân hàng/Tổng huy động	+
16	Tỷ lệ tài sản lỏng/Tổng tài sản có	-

Bảng 2. Các biến số có khả năng phân biệt các mức nguy cơ

Tên biến	Nội dung
e1	ROA-Khả năng sinh lời trên tổng tài sản
e2	ROE-khả năng sinh lời trên vốn chủ sở hữu
e4	(Lãi thuần - chi phí hoạt động)/Chi phí hoạt động
e7	Tốc độ tăng trưởng thu nhập lãi thuần
e9	Chi phí dự phòng nợ khó đòi/Tổng thu nhập trước dự phòng và thuế
e10	Thu nhập từ phí dịch vụ/Trên tổng thu nhập hoạt động
d3	Nợ quá hạn/Nợ phải trả
m2	(Lợi nhuận trước thuế +Dự phòng)/Chi phí hoạt động
m3	Thu nhập lãi thuần/Tổng tài sản
m4	(Lợi nhuận trước thuế +Dự phòng)/Tổng số tài sản có
a2	Dự phòng nợ khó đòi/Dư nợ cho vay
a3	Nợ khó đòi/(Vốn chủ sở hữu + dự phòng nợ khó đòi)
a4	Tỷ lệ cho vay/Tài sản sinh lời
c1	Tăng trưởng vốn chủ sở hữu
s3	Tốc độ tăng trưởng tài sản
s4	Lợi nhuận chưa phân phối/Lợi nhuận sau thuế
s5	Lợi nhuận chưa phân phối/Tổn tài sản
l4	Huy động từ tổ chức kinh tế và dân cư/Tổng huy động

Biến phụ thuộc: biến có nguy cơ vỡ nợ là biến phụ thuộc Y, Y được gán bằng 1 (nguy cơ vỡ nợ cao) nếu ngân hàng có tỉ lệ nợ xấu từ 3% trở lên. Biến Y được gán bằng 0 (nguy cơ vỡ nợ thấp) nếu tỉ lệ nợ xấu nhỏ hơn 3%. Trong bộ dữ liệu mảng 136 quan sát có 35 quan sát thuộc nhóm nguy cơ vỡ nợ cao và 101 quan sát thuộc nhóm nguy cơ vỡ nợ thấp.

Biến độc lập: Dựa trên nguồn số liệu hiện có, các chỉ tiêu trong mô hình CAMEL và những gợi ý từ các công trình nghiên cứu trước, cũng như hoạt động của các ngân hàng thương mại, nghiên cứu đã được xây dựng, lựa chọn 40 biến số (Bảng 1). Dấu kì vọng (+) tác động cùng chiều, (-) tác động ngược chiều, (+/-) tác động lúc thuận, lúc nghịch.

Từ 40 biến số thuộc 7 nhóm đã được tính toán, tiến hành phân tích phương sai để xác định các biến trong các nhóm có khả năng phân biệt các mức nguy cơ (Đặng Huy Ngân, 2018). Cụ thể còn 18 biến trong bảng 2.

2.2. Phương pháp nghiên cứu

2.2.1. Mô hình phân tích khác biệt tuyến tính

Phân tích khác biệt tuyến tính, có tên tiếng Anh là *Linear Discriminant Analysis* (LDA), là một phương pháp phân loại thống kê cổ điển, được đưa ra bởi Fisher (1936). LDA được sử dụng hiệu quả trong những vấn đề phân loại dữ liệu để tìm kiếm một sự kết hợp tuyến tính của các thuộc tính phân tách hai hay nhiều lớp đối tượng. Kết quả của sự kết hợp có thể được sử dụng như một bộ phận loại tuyến tính (*linear classifier*) hoặc phổ biến hơn để giảm số chiều (*dimensionality reduction*) trước khi phân loại cuối (Hastie *et al.*, 2009; Nguyen Hoang Huy, 2013; Sergio Bacallado, 2017).

a. Mô hình LDA lý thuyết

Trong không gian p chiều, có hai lớp chứa đối tượng, trong bài toán của chúng ta là nhóm có nguy cơ vỡ nợ và nhóm không có nguy cơ vỡ

nợ. Mỗi đối tượng được cho bởi một vecto biểu diễn $\boldsymbol{X} \in \mathbb{R}^p$. Sự phân bố của các đối tượng trong hai lớp đều tuân theo phân bố chuẩn, với tham số vecto trung bình $\boldsymbol{\mu}_0, \boldsymbol{\mu}_1$và cùng ma trận hiệp phương sai Σ. Để phân loại đối tượng về các lớp tương ứng cần dựa vào vecto biểu diễn X, ta giả sử X được biểu diễn như sau: $\boldsymbol{X} \sim \mathcal{N}_p(\boldsymbol{\mu}_0, \boldsymbol{\Sigma})$ hoặc $\boldsymbol{X} \sim \mathcal{N}_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$. Nếu một quan sát X thuộc về lớp k; $k \in \{0,1\}$ thì mật độ của nó là:

$$f_k(\boldsymbol{x}) = (2\pi)^{-\frac{p}{2}}(\det(\boldsymbol{\Sigma}))^{-1/2} exp\left\{-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_k)^{-1}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_k)\right\}$$

Phân lớp Bayes gán X vào lớp 0 nếu:

$$\pi_0 f_0(\boldsymbol{X}) \geq \pi_1 f_1(\boldsymbol{X})$$

Tương đương với $log\frac{\pi_0}{\pi_1} + (\boldsymbol{X} - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1) \geq 0$

Ở đó $\boldsymbol{\mu} = \frac{1}{2}(\boldsymbol{\mu}_0 + \boldsymbol{\mu}_1)$. Hàm phân biệt tuyến tính của X được xác định bởi:

$\delta_F(\boldsymbol{X}) = (\boldsymbol{X} - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)$.

δ_F là giá trị của hàm phân biệt tuyến tính lý thuyết.

b. LDA thực nghiệm

Giả sử có tập dữ liệu huấn luyện: $\{(\boldsymbol{X}^i, k_i), i = 1,2,\dots,m\}$, trong không gian p chiều xét vecto biểu diễn$\boldsymbol{X} \in \mathbb{R}^p$; $\boldsymbol{X} = \left[x_1, x_2, \dots, x_p\right]^T$

Áp dụng phân tích khác biệt tuyến tính cho hai lớp

$\boldsymbol{X}^i$, với $k_i = k, k \in \{0,1\}$ và $i = \overline{1,m}$

Xác định vecto trung bình cho các lớp

$$\widehat{\boldsymbol{\mu}}_0 = \frac{1}{m_1}\sum_{k_i=0}\boldsymbol{X}^i;\ \widehat{\boldsymbol{\mu}}_1 = \frac{1}{m_2}\sum_{k_i=1}\boldsymbol{X}^i$$

trong đó, $m_1 = \#\{i, k_i = 0\}$; $m_2 = \#\{i, k_i = 1\}$; $m_1 + m_2 = m \rightarrow \widehat{\boldsymbol{\mu}} = \frac{\widehat{\boldsymbol{\mu}}_0 + \widehat{\boldsymbol{\mu}}_1}{2}$

Ma trận hiệp phương sai mẫu $\widehat{\boldsymbol{\Sigma}}$ cho các lớp

$$\widehat{\boldsymbol{\Sigma}}_k = \frac{1}{m_k - 1}\sum_{k=k_i}(\boldsymbol{X}^i - \widehat{\boldsymbol{\mu}}_k)(\boldsymbol{X}^i - \widehat{\boldsymbol{\mu}}_k)^T, k = 0,1$$

$$\rightarrow \widehat{\boldsymbol{\Sigma}} = \frac{(m_1 - 1)\widehat{\boldsymbol{\Sigma}}_0 + (m_2 - 1)\widehat{\boldsymbol{\Sigma}}_1}{m_1 + m_2 - 2}$$

Ta xây dựng được hàm phân biệt tuyến tính của $\boldsymbol{X}$ như sau:

$$\hat{\delta}_F(\boldsymbol{X}) = (\boldsymbol{X} - \widehat{\boldsymbol{\mu}})^T\widehat{\boldsymbol{\Sigma}}^{-1}(\widehat{\boldsymbol{\mu}}_0 - \widehat{\boldsymbol{\mu}}_1).$$

Giá trị hàm phân biệt tuyến tính còn gọi là giá trị điểm.

Đặt $\hat{c} = \log\frac{\widehat{\pi}_0}{\widehat{\pi}_1}$, giá trị $\hat{c}$ dùng để phân loại dữ liệu $\boldsymbol{X}^i$ vào lớp tương ứng của nó, gọi là ngưỡng phân loại.

Trong đó: $\widehat{\pi}_0 \approx \frac{m_1}{m}$ là ước lượng của xác suất lớp thứ nhất.

$\widehat{\pi}_1 \approx \frac{m_2}{m}$ là ước lượng của xác suất lớp thứ hai.

- Nếu $\hat{\delta}_F(\boldsymbol{X}) \geq \hat{c} \rightarrow k_i = 0$ (lớp thứ nhất).

- Nếu $\hat{\delta}_F(\boldsymbol{X}) < \hat{c} \rightarrow k_i = 1$ (lớp thứ hai).

Hàm phân biệt tuyến tính thực nghiệm của $\boldsymbol{X}$ được xác định bởi:

$$\hat{\delta}_F(\boldsymbol{X}) = (\boldsymbol{X} - \widehat{\boldsymbol{\mu}})^T\widehat{\boldsymbol{\Sigma}}^{-1}(\widehat{\boldsymbol{\mu}}_0 - \widehat{\boldsymbol{\mu}}_1).$$

2.2.2. Mô hình hồi qui logistic

Trong các mô hình hồi qui truyền thống, biến phụ thuộc và biến độc lập có thể nhận giá trị trên tập số thực. Trong thực tế có rất nhiều trường hợp, một đại lượng chỉ nhận hai giá trị 0 và 1, nhưng nó lại phụ thuộc vào các biến độc lập khác nhận giá trị trên tập số thực. Người ta cần đưa ra một phương trình mô tả mối quan hệ giữa xác suất p để một biến cố A xảy ra với giá trị của các biến độc lập $x_1, x_2, \dots, x_n$. Trong bài toán này biến cố A là biến cố ngân hàng bị vỡ nợ, các biến độc lập là các biến trong bảng 2. Phương trình dạng tuyến tính biểu diễn xác suất p qua một tổ hợp tuyến tính của các biến độc lập thường được nghĩ đến trước tiên. Tuy nhiên, một phương trình tuyến tính như vậy là không hợp lý, vì p chỉ nhận giá trị giới hạn trong đoạn [0,1], trong khi đó tổ hợp tuyến tính của các biến độc lập có thể nhận giá trị bất kỳ trên đường thẳng thực. Nhưng người ta nhận thấy có mối quan hệ chặt chẽ giữa tỉ lệ cược, thành phần $ln(\frac{p}{1-p})$ và các biến độc lập x_i dưới dạng tuyến tính nên đã thiết lập chúng dưới dạng:

$$y = \ln(\frac{p}{1-p}) = \beta_0 + \sum_{i=1}^{n}\beta_i x_{ij} (1)$$

Phương trình (1) được gọi là mô hình hồi qui logistic bội, khi n = 1 ta có mô hình hồi qui logistic đơn. Sử dụng phương pháp hợp lý cực đại, các hệ số β_i trong phương trình (1) có ước lượng là $\widehat{\beta}_i$ được xác định bởi hệ phương trình sau:

$$\sum_{i=1}^{n} p_i = \sum_{i=1}^{n} (1+\exp[-(\widehat{\beta}_0 + \sum_{j}^{k} \widehat{\beta}_j x_{ij})])^{-1}$$
$$\sum_{i=1}^{n} x_i p_i = \sum_{i=1}^{n} x_i (1+\exp[-(\widehat{\beta}_0 + \sum_{j}^{k} \widehat{\beta}_j x_{ij})])' \quad (2)$$

Trong đó p_i nhận giá trị bằng 1 nếu biến cố A xảy ra và nhận giá trị bằng 0 nếu ngược lại: $\widehat{\beta}_i$ là ước lượng của β_i; x_{ij} là dữ liệu thứ j của biến độc lập x_i. Khi tìm được các hệ số của phương trình hồi quy, ta có xác suất thành công của phần tử có biến quan sát $x = (x_1, x_2, \ldots, x_n)$ là:

$$p = \frac{\exp(\widehat{\beta}_0 + \sum_{i=1}^{n} \widehat{\beta}_i x_i)}{1+\exp(\widehat{\beta}_0 + \sum_{i=1}^{n} \widehat{\beta}_i x_i)}$$

Khi đó, nếu $p > 0,5$ thì ta sẽ xếp phần tử này vào lớp xảy ra A, ngược lại ta sẽ xếp nó vào lớp không xảy ra A (Võ Văn Tài và cs.).

2.2.3. Máy vecto hỗ trợ

Máy vecto hỗ trợ (SVM - *Support Vector Machine*) là một khái niệm trong thống kê và khoa học máy tính cho một tập hợp các phương pháp học có giám sát liên quan đến nhau để phân loại và phân tích hồi quy. Thuật toán SVM ban đầu được tìm ra bởi Vapnik (1995) và dạng chuẩn hiện nay sử dụng lề mềm được tìm ra bởi Vapnik và Corte (1995). SVM dạng chuẩn nhận dữ liệu vào và phân loại chúng vào hai lớp khác nhau. Một mô hình SVM là một cách biểu diễn các điểm trong không gian và lựa chọn ranh giới giữa hai lớp sao cho khoảng cách lề từ các ví dụ huấn luyện tới ranh giới là xa nhất có thể. Trong nhiều trường hợp, không thể phân chia các lớp dữ liệu một cách tuyến tính trong một không gian thuộc tính ban đầu. Vì vậy, nhiều khi cần phải ánh xạ các điểm dữ liệu trong không gian ban đầu vào một không gian mới nhiều chiều hơn, để việc phân tách chúng trở nên dễ dàng hơn trong không gian mới. Ánh xạ sử dụng trong SVM chỉ đòi hỏi biết tích vô hướng của các vecto dữ liệu trong không gian mới, tích vô hướng này được xác định bằng một hàm hạt nhân $K(x,y)$ phù hợp. Một sự mô tả đơn giản cho thuật toán SVM được cung cấp dưới đây (Min *et al.*, 2005):

Cho trước một tập huấn luyện $\boldsymbol{D} = \{x_i, y_i\}_{i=1}^{N}$ với đầu vào là các vecto $\boldsymbol{x_i} = (x_i^{(1)}, \ldots, x_i^{(n)})^T \in R^n$ và tập nhãn $y_i \in \{-1, +1\}$, máy phân loại vecto hỗ trợ SVM theo công thức gốc của Vapnik, thỏa mãn các điều kiện sau đây:

$$\begin{cases} \boldsymbol{w}^T \phi(x_i) + b \geq +1 \text{ nếu } y_i = +1 \\ \boldsymbol{w}^T \phi(x_i) + b \leq -1 \text{ nếu } y_i = -1 \end{cases} \quad (3)$$

Điều này tương đương với $y_i[\boldsymbol{w}^T \phi(x_i) + b] \geq 1, i = 1, \ldots, N$ (4)

Ở đó w là vecto trọng số và b là khuynh hướng. Ánh xạ phi tuyến $\phi(.): R^n \to R^{n_k}$ ánh xạ không gian thuộc tính đầu vào đo được vào không gian thuộc tính có số chiều cao, hoặc vô hạn chiều (n_k là số chiều của không gian thuộc tính). Phương trình (3) xác định hai siêu phẳng lề song song (có cùng vecto pháp tuyến) nằm hai bên (theo hướng xác định bởi vecto pháp tuyến) siêu phẳng phân tách $\boldsymbol{w}^T \phi(x) + b = 0$ trong không gian thuộc tính với độ rộng lề giữa hai siêu phẳng bằng $2/(\|\boldsymbol{w}\|^2)$. Hàm phân loại dựa trên giá trị thuộc tính ban đầu xác định bởi:

$$\text{sgn}(\boldsymbol{w}^T \phi(x) + b) \quad (5)$$

Hầu hết các bài toán phân lớp là không phân tách tuyến tính. Vì vậy, một cách tổng quát để tìm vecto trọng số ta sử dụng biến giảm ξ_i để cho phép phân loại sai. Bài toán tối ưu lề trở thành:

$$\min_{\boldsymbol{w},b,\xi} \frac{1}{2} \boldsymbol{w}^T \boldsymbol{w}. + C. \sum_{i=1}^{N} \xi_i \quad (6)$$

Tùy thuộc vào

$$\begin{cases} y_i(\boldsymbol{w}^T \phi(x_i) + b) \geq 1 - \xi_i, i = 1, \ldots, N \\ \xi_i \geq 0 \; i = 1, .., N \end{cases} \quad (7)$$

Ở đó các ξ_i là các biến giảm cần thiết để cho phép phân loại sai mẫu thứ i, và $C \in R^+$ là một siêu tham số điều chỉnh mức độ phân độ sai để cân bằng với độ rộng lề. Từ các điều kiện tối ưu và hàm mục tiêu, thu được một bài toán qui hoạch toàn phương (QP), có thể giải bằng phương pháp nhân tử Lagrange. Hệ số nhân Lagrange α_i tồn tại tương ứng với mỗi mẫu trong dữ liệu huấn luyện. Các mẫu tương ứng với các α_i khác không chính là các vecto hỗ trợ. Khi đó, bài toán trên có thể chuyển đổi thành bài toán đối ngẫu với hàm mục tiêu (8) và ràng buộc (9) như sau:

$$\max_{\alpha} \frac{1}{2}\alpha^T Q\alpha - \boldsymbol{e^T}\alpha \quad (8)$$

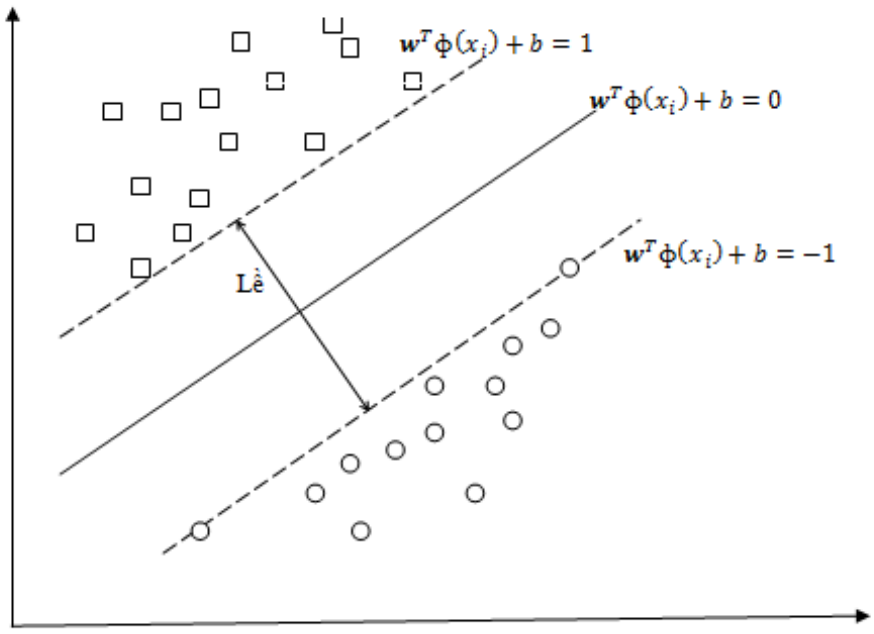
$$\text{Với ràng buộc} \begin{cases} 0 \le \alpha_i \le C, i = 1, \dots, N \\ y^T\alpha = 0 \end{cases} \quad (9)$$

Trong bài toán đối ngẫu ở trên, e là vecto đơn vị gồm toàn số 1, Q là ma trận bán xác định dương, $Q_{ij} = y_i y_j K(x_i, x_j)$ và $K(x_i, x_j) = \phi(x_i)^T\phi(x_j)$ là hạt nhân. Ở đây, các vecto x_i được ánh xạ vào không gian có số chiều cao hơn bởi hàm ϕ. Sau đó xây dựng SVM phân loại cuối cùng như sau:

$$\text{sgn}\left(\sum_i^N \alpha_i y_i K(x, x_i) + b\right) \quad (10)$$

3. KẾT QUẢ VÀ THẢO LUẬN

Để xây dựng hàm phân loại trong việc cảnh báo sớm nguy cơ vỡ nợ của các ngân hàng thương mại cổ phần Việt Nam, nhóm chúng tôi tiến hành xây dựng và cài đặt trên ngôn ngữ lập trình Matlab, sử dụng máy tính Intel Core i3, 2.1 GHz, 2GB Ram. Dữ liệu thử nghiệm được chia một cách ngẫu nhiên nhờ hàm *cvpartition*, thực hiện 5 lần theo k-fold = 5, mỗi lần thành 2 tập con: Một tập dữ liệu huấn luyện chiếm 80% và một tập dữ liệu kiểm tra chiếm 20% trong tổng số 136 dữ liệu của các ngân hàng đối với 2 mô hình phân tích khác biệt và hồi qui SVM, còn đối với hồi qui logistic chúng tôi chia tập dữ liệu lần lượt theo k-fold = 1, 2, 3, 4, 5. Kết quả phân loại thu được đối với các phương pháp phân tích khác biệt tuyến tính (LDA), hồi qui logistic và SVM được thể hiện trong bảng 3.



Hình 1. Ví dụ về siêu phẳng với lề cực đại trong không gian R^2

Bảng 3. Độ chính xác trung bình của các mô hình

Mô hình	Độ chính xác trung bình của dự đoán (%)	
	Dữ liệu huấn luyện	Dữ liệu kiểm tra
LDA	88,26	85,19
Logistic	95,05	83,07
SVM	79,08	80,74

Bảng 4. Giá trị *p-value* cho so sánh hiệu suất của các cặp mô hình

	LDA	Logistic	SVM
LDA		0,6250	0,5000
Logistic			1

Bảng 5. Giá trị các sai lầm của các mô hình

Sai lầm loại I (%)			Sai lầm loại II (%)			Sai lầm chung (%)		
LDA	Logistic	SVM	LDA	Logistic	SVM	LDA	Logistic	SVM
43,33	11,68	84,67	6,48	31	1,86	14,81	16,03	19,26

Độ chính xác của mô hình phân loại được tính bằng tỷ số giữa số mẫu được phân loại đúng trên tổng số mẫu của tập dữ liệu kiểm thử.

Kết quả ở bảng 3 cho thấy trong bài toán này, độ chính xác của ba mô hình LDA, hồi qui logistic, SVM trên tập dữ liệu huấn luyện lần lượt là 88,26; 95,05; 79,08% và trên tập dữ liệu kiểm tra lần lượt là 85,19; 83,70; 80,74%. Như vậy, mô hình hồi qui logistic có độ chính xác cao nhất ở cả tập dữ liệu huấn luyện còn LDA lại là phương pháp có độ chính xác cao nhất ở tập dữ liệu kiểm tra.

Ngoài ra để kiểm định sự khác nhau trong xác suất thành công của ba mô hình,chúng tôi còn đưa ra giá trị p-value trong kiểm định phi tham số χ^2 nhưở bảng 4, theo đó thì hiệu suất dự đoán của các mô hình LDA, Logistic, SVM không quá khác biệt.

Để đánh giá hiệu quả của các mô hình trên thông qua độ chính xác dự báo, chúng tôi còn đánh giá thông qua tỉ lệ phân loại sai hay nói cách khác là các loại sai lầm mắc phải. Bảng 5 chỉ ra tỉ lệ sai lầm loại I, sai lầm loại II và sai lầm nói chung của ba mô hình. Nhắc lại, sai lầm loại I (ER_I)mắc phải khi phân loại một ngân hàng có nguy cơ vỡ nợ thành ngân hàng không có nguy cơ vỡ nợ và sai lầm loại II (ER_I) mắc phải khi phân loại một ngân hàng không có nguy cơ vỡ nợ thành ngân hàng có nguy cơ vỡ nợ. Sai lầm nói chung (ER) mắc phải khi phân loại sai, và nó có công thức tính như sau: $ER = ER_I.P_I + ER_{II}.P_{II}$. Trong đó, P_I, P_{II} lần lượt là xác suất bị phá sản và xác suất không bị phá sản. Trong thực tế, chi phí sai lầm loại I gần như cao hơn sai lầm loại II. Sinkey (1975) đã chỉ ra rằng một tỉ lệ sai lầm chung nhỏ với một sai lầm loại I lớn sẽ hao tốn chi phí nhiều hơn một tỉ lệ sai lầm chung lớn với sai lầm loại I nhỏ.

Theo kết quả ở bảng 5, tỉ lệ các loại sai lầm loại I của phương pháp hồi qui logistic thấp nhất trong baphương pháp, tỉ lệ sai lầm chung thì của LDA là thấp nhất, tiếp đó đến hồi qui logistic và cuối cùng là SVM, cụ thể tỉ lệ sai lầm nói chung của các phương pháp này lần lượt là 14,81%; 16,03% và 19,26%.

Như vậy, từ các kết quả thực nghiệm ở trên, có thể thấy rằng cả ba phương pháp được áp dụng gồmphân tích khác biệt tuyến tính LDA, hồi qui logistic, SVM đều đạt độ chính xác trong dự báo rủi ro của ngân hàng là khá cao (lớn hơn 70%). Trong đó, phương pháp hồi qui logistic và phân tích khác biệt tuyến tính thể hiện được ưu thế hơn so với mô hình máy vecto hỗ trợ SVM, do có độ chính xác cao hơn và tỉ lệ sai lầm thấp hơn mô hình máy vecto hỗ trợ SVM.

4. KẾT LUẬN VÀ KIẾN NGHỊ

Nghiên cứu đề xuất một số phương pháp xây dựng hàm phân loại trong việc cảnh báo sớm nguy cơ vỡ nợ của các ngân hàng thương mại cổ phần Việt Nam, cụ thể ba phương pháp là phân tích khác biệt tuyến tính LDA, hồi qui logistic và SVM. Các kết quả thực nghiệm với mức độ chính xác tương đối tốt, cho thấy việc áp dụng mô hình này trong thực tế là hoàn toàn có triển vọng.

Đây là mô hình có tính ứng dụng cao trong thực tiễn bởi ngân hàng là một trong các tổ chức trung gian tài chính quan trọng của nền kinh tế. Vì thế trong tương lai cần phát triển tiếp ứng dụng này, tiếp tục thu thập thêm nhiều dữ liệu của các ngân hàng khác nhau qua các năm khác nhau, cải tiến phương pháp phân loại SVM, tìm hiểu thêm các phương pháp phân loại khác như mạng nơron, cây quyết định, trí tuệ nhân tạo (AI) để đạt được kết quả phân loại tốt hơn. Đồng thời cần tìm hiểu chuyên sâu hơn nữa mô hình phân loại áp dụng cho các lĩnh vực thực tiễn khác như định giá bất động sản, chuẩn đoán bệnh trong y tế,…

TÀI LIỆU THAM KHẢO

Nguyễn Quang Dong (2009). Xếp hạng tín dụng các ngân hàng, các tổ chức tài chính Việt Nam bằng phương pháp phân tích khác biệt. Đề tài khoa học cấp bộ.

Đặng Huy Ngân (2015). Sử dụng kết hợp phân tích nhân tố và hồi qui Logistic để phân loại các ngân hàng thương mại cổ phần Việt Nam. Kỷ yếu hội thảo khoa học quốc gia "An ninh tài chính tiền tệ của Việt Nam trong bối cảnh hội nhập quốc tế" T7-2015.

Đặng Huy Ngân (2016). Xây dựng mô hình cảnh báo nguy cơ vỡ nợ cho các ngân hàng thương mại cổ phần Việt Nam. Tạp chí Kinh tế & Phát triển. Số đặc biệt, tr. 82-90.

Đặng Huy Ngân (2018). Xây dựng mô hình cảnh báo nguy cơ vỡ nợ cho các ngân hàng thương mại cổ phần Việt Nam. Luận án tiến sĩ Kinh tế học.

Lê Thanh Ngọc, Đặng Trí Dũng và Lê Nguyễn Minh Phương (2015). Mối quan hệ giữa tỉ lệ vốn tự có và rủi ro của ngân hàng thương mại. Tạp chí Phát triển & Hội nhập, 15(35): 54-61.

Nguyễn Nhật Quang (2012). Trí tuệ nhân tạo nâng cao. Viện Công nghệ thông tin và Truyền thông, Đại học Bách Khoa Hà Nội.

Võ Văn Tài, Đổng Yến Nghi (2016). Bài toán phân loại và ứng dụng trong y học. Tạp chí Khoa học, Đại học Cần Thơ, 42: 127-133.

Altman, Edward I. (1968). Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy. Journal of Finance, 23: 589-609.

Barbro Back, Teija Laitinen, Kaisa Sere, Michiel van Wezel (1996). Choosing bankruptcy predictors using discriminant analysis, logit analysis and genetic algorithms. Turku Centre for Computer Science Technical Report No 40, September 1996.

LeCun, Y. (1986). Learning Process in an Asymmetric Threshold Network. Disordered systems and biological and organizations, LesHouches, France, Springer.

Nguyen Hoang Huy (2013). Multi-step linear discriminant Analysis and its applications. Ph.D. thesis. Greifswald University, p. 7.

Hastie, T. và Tibshirani, R., Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer Series in Statistics. Springer Verlag.

Jae H. Min, Young-ChanLee (2005). Bankruptcy prediction using support vector machine with optimal choice of Kernel function parameters. Expert Systems with Applications, 28: 603-614.

M.Adnan Aziz, Humayon (2006). Predicting corporate bankruptcy: where we stand? Corporate governance: The international journal of business in society, 6(1): 18-33.

Soo Y Kim (2011). Prediction of hotel bankruptcy using support vector machine, artificial neural network, logistic regression, and multivariate discriminant analysis. The Service Industries Journal, 31(3): 441-468.

Sergio Bacallado (2017). Data Mining and Analysis. Stanford University. http://web.stanford.edu/class/stats202/content/lectures.html/lec9.pdf. Truy cập ngày 9/1/2018.

Valadimir Vapnik (1995). The nature of statistical learning theory. Springer-Verlag.