Tạp chí Khoa học Trường Đại học Cần Thơ
Phần A: Khoa học Tự nhiên, Công nghệ và Môi trường
website: sj.ctu.edu.vn

DOI:10.22144/jvn.2017.015

ĐÁNH GIÁ KHẢ NĂNG TRẢ NỢ VAY CỦA KHÁCH HÀNG BẰNG CÁC PHƯƠNG PHÁP PHÂN LOẠI

Võ Văn Tài, Nguyễn Thị Hồng Dân và Nghiêm Quang Thường
Khoa Khoa học Tự nhiên, Trường Đại học Cần Thơ

Thông tin chung:
Ngày nhận: 06/07/2016
Ngày chấp nhận: 28/04/2017

Title:
Assessing ability of customers in loan repayment by classification methods

Từ khóa:
Ngân hàng, phương pháp Bayes, phân loại, sai lầm, xác suất tiên nghiệm

Keywords:
Bank, Bayesian method, classification, mistake, prior probability

ABSTRACT

This article presents the classification methods and calculable problems in their real application. The article also proposes an algorithm to determine the prior probability in classifying by Bayesian method that is better than existing ones. The application from real data in appraising ability to repay loans of customers is performed by all methods to illustrate for theories and to examine logic of the establishsed algorithm. This application also shows that the proposed approach is more advantage than others and it can be applied for many other domains.

TÓM TẮT

Bài báo trình bày các phương pháp phân loại và những vấn đề tính toán trong áp dụng thực tế của chúng. Bài báo cũng đề nghị một thuật toán xác định xác suất tiên nghiệm trong phân loại bằng phương pháp Bayes tốt hơn các phương pháp khác. Ứng dụng từ số liệu thực tế trong đánh giá khả năng trả nợ vay của khách hàng được thực hiện bằng tất cả các phương pháp để minh họa cho lý thuyết và kiểm tra sự hợp lý của thuật toán được thiết lập. Ứng dụng này cũng cho thấy phương pháp đề nghị có ưu điểm hơn các phương pháp khác và có thể được áp dụng cho nhiều lĩnh vực khác nhau.

Trích dẫn: Võ Văn Tài, Nguyễn Thị Hồng Dân và Nghiêm Quang Thường, 2017. Đánh giá khả năng trả nợ vay của khách hàng bằng các phương pháp phân loại. Tạp chí Khoa học Trường Đại học Cần Thơ. 49a: 110-117.

1 GIỚI THIỆU

Phân loại là xếp một phần tử thích hợp vào các tổng thể đã biết dựa trên các biến quan sát của nó. Hiện nay, các phương pháp chính được sử dụng là Fisher, hồi qui logistic, SVM (Support Vector Machines) và Bayes (Webb, 2000; Tai, 2016). Phương pháp Fisher ra đời sớm nhất, có thể phân loại cho hai hay nhiều hơn hai tổng thể, phương pháp này bị ràng buộc bởi giả thiết ma trận hiệp phương sai của chúng bằng nhau. Phương pháp SVM chỉ phân loại cho hai tổng thể dựa trên số liệu rời rạc. Hiện nay, phương pháp này được áp dụng khá phổ biến trong khai khoáng dữ liệu. Mặc dù được đề xuất muộn nhất và chỉ phân loại cho hai tổng thể, nhưng phương pháp hồi qui logistic đang được sử dụng rất phổ biến hiện nay. Phương pháp Bayes có nhiều ưu điểm, có thể phân loại được cho hai hay nhiều hơn hai tổng thể. Nó không bị ràng buộc bởi các giả thiết phân phối chuẩn và phương sai bằng nhau của các tổng thể. Hai vấn đề chính được quan tâm của phương pháp này là tìm hàm mật độ xác suất từ dữ liệu rời rạc và xác định xác suất tiên nghiệm. Hiện nay, việc nghiên cứu hai vấn đề này không những được sự quan tâm của các nhà thống kê mà còn có sự kết hợp của các nhà khoa học trong lĩnh vực công nghệ thông tin. Vấn đề ước lượng hàm mật độ xác suất đã được thảo luận rất nhiều trong các tổng kết và nghiên cứu, nhiều kết quả đã được áp dụng vào thực tế rất hiệu

quả (Pham-Gia *et al*., 2008; Tai, 2016). Việc xác định xác suất tiên nghiệm thường dựa vào các tổng kết thống kê, kinh nghiệm và tập dữ liệu thực hiện. Các xác suất tiên nghiệm thông thường được đề xuất theo phân phối đều, phương pháp Laplace hoặc tỉ lệ mẫu. Trong bài viết này, dựa vào phân tích chùm mờ, chúng tôi đề xuất thuật toán xác định xác suất tiên nghiệm mà nó được xem là hiệu quả hơn các phương pháp khác khi áp dụng vào thực tế (xác suất sai lầm nhỏ hơn).

Bài toán phân loại đã và đang được áp dụng cho nhiều lĩnh vực khác nhau, đặc biệt trong ngân hàng. Khi khách hàng (cá nhân, doanh nghiệp…) đến vay vốn, cán bộ tín dụng phải có khả năng đánh giá đúng khách hàng và ra quyết định về việc cho hay không cho khách hàng vay. Cán bộ tín dụng cần phải hạn chế sai lầm: Cho vay đối với khách hàng có rủi ro hoặc từ chối cho vay đối với khách hàng tốt. Trong những năm qua, hệ thống ngân hàng Việt Nam phát triển mạnh nhưng nợ xấu cũng tăng cao, tiềm ẩn nhiều rủi ro. Đánh giá khả năng trả nợ của khách hàng là một nhiệm vụ quan trọng đối với các ngân hàng hiện nay. Mỗi khách hàng đến vay vốn tại ngân hàng sẽ được xác định bởi một bộ thông tin (do khách hàng cung cấp, kết hợp với sự điều tra từ cán bộ tín dụng). Thông tin của khách hàng là một véc tơ n chiều gồm các biến định tính và định lượng. Với n biến này, cán bộ tín dụng cần phân loại khách hàng thuộc nhóm nào, từ đó quyết định cho khách hàng vay hay không với mức sai lầm thấp nhất. Kết quả lý thuyết của bài viết này, trong đánh giá khả năng trả nợ vay của khách hàng, hoàn toàn có thể ứng dụng thực hiện tương tự trong nhiều lĩnh vực khác.

Cấu trúc tiếp theo của bài viết như sau: Phần 2 trình bày các phương pháp phân loại và vấn đề xác định xác suất tiên nghiệm bằng phương pháp Bayes. Phần 3 trình bày vấn đề tính toán của các phương pháp, trong đó có vấn đề thiết lập các chương trình trên phần mềm Matlab để hỗ trợ cho các tính toán phức tạp. Phần 4 thực hiện đánh giá khả năng trả nợ vay của khách hàng dựa vào các số liệu thực tế của các doanh nghiệp trên địa bàn thành phố Cần Thơ. Phần cuối cùng là kết luận của bài viết.

2 CÁC PHƯƠNG PHÁP PHÂN LOẠI

2.1 Phương pháp 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, . . ., x_n. 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 [0,1], trong khi đó tổ hợp tuyến tính của các biến độc lập có thể nhận giá trị bất kỳ trên tập số thực. Nhận xét thấy có mối quan hệ chặt chẽ giữa logarit của số chênh, $(\ln(p/(1-p))$, và các biến độc lập x_i dưới dạng tuyến tính nên người ta thiết lập chúng dưới dạng:

$$y = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \sum_{i=1}^{n} \beta_i x_i. \qquad (1)$$

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 mô hình (1) được xác định bởi hệ phương trình sau:

$$\begin{cases} \sum_{i=1}^{n} p_i = \sum_{i=1}^{n} \left(1 + \exp\left[-\left(\widehat{\beta}_0 + \sum_{j=1}^{k} \widehat{\beta}_j x_{ij}\right)\right]\right)^{-1}, \\ \sum_{i=1}^{n} x_i p_i = \sum_{i=1}^{n} x_i \left(1 + \exp\left[-\left(\widehat{\beta}_0 + \sum_{j=1}^{k} \widehat{\beta}_j x_{ij}\right)\right]\right), \end{cases} \qquad (2)$$

trong đó p_i nhận giá trị bằng 1 nếu biến cố A xảy ra và nhận giá trị bằng 0 nếu ngược lại; $\widehat{\beta}$ 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 qui, ta có xác suất thành công của phần tử có biến quan sát $x = (x_1, x_2 \ldots x_n)$ là

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

Khi đó nếu $p > 0.5$ thì ta sẽ xếp phần tử này vào lớp xảy ra A, ngược lại, ta xếp nó vào lớp không xảy ra A.

2.2 Phương pháp Fisher

Xét k tổng thể w_1, w_2, . . ., w_k, $(k \geq 2)$ có véc tơ trung bình μ_i, i = 1, 2,…, k và ma trận hiệp

phương sai của các tổng thể đều bằng nhau $\Sigma_1 = \Sigma_2 = ... = \Sigma_k = \Sigma$. Đặt:

$$d_i(x) = \mu_i^T \Sigma^{-1} x - \frac{1}{2} \mu_i^T \Sigma^{-1} \mu_i. \quad (3)$$

Khi đó một phần tử mới với biến quan sát *x* sẽ được xếp vào w_j nếu:

$$d_j(x) = \max_i \{d_i(x)\}.$$

2.3 Phương pháp Bayes

Cho *k* tổng thể w_1, w_2...w_k có biến quan sát với hàm mật độ xác suất được xác định là $f_1(x)$, $f_2(x)$,…, $f_k(x)$ và xác suất tiên nghiệm cho các tổng thể lần lượt là q_1, q_2, …, q_k với $q_1 + q_2 + …+ q_k = 1$. Ta có nguyên tắc phân loại một phần tử mới với biến quan sát x_0 bằng phương pháp Bayes như sau:

Nếu $g_{max}(x_0) = q_i f_i(x_0)$ thì xếp phần tử mới vào w_i, (4)

trong đó

q_i là xác suất tiên nghiệm của tổng thể thứ *i*,

$g_i(x) = q_i f_i(x)$ và $g_{max}(x) = \max\{g_1(x), g_2(x),…, g_k(x)\}$.

Xác suất sai lầm trong phân loại Bayes được gọi là sai số Bayes và được xác định bởi công thức:

$$Pe^{(q)}_{1,2,...,k} = \sum_{i=1}^{k} \int_{R^n \setminus R_i^n} q_i f_i dx, \quad (5)$$

trong đó *n* là số chiều của biến quan sát.

Từ công thức (5), ta có thể chứng minh được

$$Pe^{(q)}_{1,2,...,k} = 1 - \int_{R^n} \max_{1 \le l \le k} \{q_l f_l(x)\} dx. \quad (6)$$

Sử dụng (6) để tính sai số Bayes cho ta một thuận lợi rất lớn, đặc biệt trong việc sử dụng các phần mềm toán học để lập trình.

2.4 Xác định xác suất tiên nghiệm trong phân loại bằng phương pháp Bayes

a. Vấn đề xác định xác suất tiên nghiệm

Kết quả phân loại một phần tử mới bởi nguyên tắc (4) và sai số Bayes được tính bởi công thức (6) đều phụ thuộc vào xác suất tiên nghiệm. Thông thường có những phương pháp sau để xác định các xác suất tiên nghiệm:

(i) Dựa vào phân phối đều: $q_1 = q_2 = ... = q_c = 1/c.$

(ii) Dựa vào tập mẫu: $q_i = n_i / N$,

(iii) Dựa vào ước lượng Laplace: $q_i = (n_i + 1)/(N + n)$,

trong đó n_i là số các phần tử trong w_i, *n* là số chiều và *N* là số những phần tử của tập mẫu.

Mặc dù có nhiều tác giả đã nghiên cứu về vấn đề này (Inman and Bradley, 1989; Miller, 2011; Bora and Gupta, 2014) nhưng việc tìm một xác suất tiên nghiệm thích hợp cho từng trường hợp cụ thể cho đến nay vẫn là một bài toán chưa có lời giải cuối cùng.

Trong phần này, chúng tôi đề xuất thuật toán tìm xác suất tiên nghiệm mà thực tế kiểm chứng cho ta sai số Bayes nhỏ hơn khi ta sử dụng các xác suất tiên nghiệm vừa đề cập ở trên. Trước khi xem xét thuật toán này, chúng ta tìm hiểu một số khái niệm sau.

b. Khái niệm

Trong không gian *n* chiều, cho *N* tổng thể $N^{(0)} = \{W_1^{(0)}, W_2^{(0)}, ..., W_N^{(0)}\}$ với tập dữ liệu $Z = [z_{ij}]_{nxN}$. Xét ma trận $U = [\mu_{ik}]_{c \times n}$, trong đó μ_{ik} là xác suất khi chúng ta xếp phần tử thứ *k* vào chùm thứ *i*. Trong phân tích chùm không mờ, $\mu_{ik} = 1$ khi phần tử thứ *k* thuộc vào chùm thứ *i*, $\mu_{ik} = 0$ khi phần tử thứ *k* không thuộc chùm thứ *i*. Trong phân tích chùm mờ $\mu_{ik} \in [0,1]$ và phải thỏa những điều kiện sau:

$$\sum_{i=1}^{c} \mu_{ik} = 1, \quad 0 < \sum_{k=1}^{N} \mu_{ik} < N, 1 \le i \le c, 1 \le k \le N.$$

Tập tất cả những ma trận phân vùng mờ cho dữ liệu $[z_{ij}]_{nxN}$, $N \ge 2$ được gọi là không gian phân vùng mờ của *c* chùm:

$$M_{cx} = \left\{ U = [\mu_{ik}]_{cxN} \mid \mu_{ik} \in [0,1], \forall i,k; \sum_{i=1}^{c} \mu_{ik} = 1, \forall k; 0 < \sum_{k=1}^{N} \mu_{ik}, \forall i \right\}.$$

Trong phân tích chùm không mờ, phần tử đại diện chùm được lấy chính là trọng tâm. Khi phân tích chùm mờ, phần tử đại diện chùm thứ *i* được xác định bởi

$$v_i = \frac{\sum_{k=1}^{N} (\mu_{ik})^m z_k}{\sum_{k=1}^{N} (\mu_{ik})^m}, \ 1 \le i \le c. \quad (7)$$

trong đó m là tham số xác định độ mờ.

c. Thuật toán

Thuật toán xác định xác suất tiên nghiệm khi phân loại phần tử x_0 vào c tổng thể được đề nghị gồm các bước như sau:

***Bước* 1**: Chia tập dữ liệu thành c chùm w_1, w_2,…, w_c. Tìm phần tử đại diện của các chùm v_i bởi công thức (7), tính khoảng cách giữa các phần tử của dữ liệu và các v_i (với i = 1, 2,…, c).

***Bước* 2:** Thiết lập ma trận phân vùng ban đầu $U^{(0)} = [\mu_{ij}]_{c \times N+1}$, trong đó N cột đầu tiên là ma trận phân vùng không mờ của các phần tử trong tập dữ liệu khi xếp vào c tổng thể w_1, w_2,..., w_c. Cụ thể $\mu_{ij} = 1$, nếu phần tử thứ j thuộc tổng thể i (với i = 1, 2,..., c) và $\mu_{ij} = 0$ trong trường hợp ngược lại. Cột cuối cùng $N + 1$ là xác suất ban đầu để x_0 xếp vào các chùm w_1, w_2,..., w_c. Ban đầu chúng ta có thể chọn xác suất này bằng nhau.

***Bước* 3:** Tính $D_{ikA}^2 = \|z_k - v_i\|_A^2 = (z_k - v_i)^T A (z_k - v_i)$ là bình phương khoảng cách từ phần tử z_k đến phần tử đại diện chùm thứ i. Cập nhật ma trận phân vùng mới $U^{(1)}$ với

$$\mu_{ik}^{(1)} = \frac{1}{\sum_{j=1}^{c} \left(D_{ikA} / D_{jkA} \right)^{2/(m-1)}}. \quad (8)$$

nếu $D_{ikA} > 0$ cho tất cả i = 1, 2,…, c và $\mu_{ik}^{(1)} = 0$ trong các trường hợp ngược lại.

***Bước 4*:** Tính

$$\left\| U^{(1)} - U^{(0)} \right\| = \max_{ik} \left(\left| \mu_{ik}^{(1)} - \mu_{ik}^{(0)} \right| \right),$$

Lặp lại các bước trên cho đến khi $\left\| U^{(n)} - U^{(n-1)} \right\| < \varepsilon$, khi đó chúng ta sẽ có ma trận phân vùng cuối cùng. Cột cuối cùng của ma trận phân vùng là xác suất tiên nghiệm khi xếp x_0 vào các tổng thể tương ứng.

Trong thuật toán trên, chúng ta cần chú ý những vấn đề sau:

i) ε là một hằng số nhỏ tùy ý. Khi ε càng nhỏ thì vòng lặp thực hiện sẽ càng nhiều. Chúng ta có thể chọn ε = 5% hoặc 1% trong các ứng dụng.

ii) D_{ikA} phụ thuộc vào ma trận A. Khi A là ma trận đơn vị thì D_{ikA} là khoảng cách Euclide. Trong bài báo này, chúng tôi chọn khoảng cách Euclide trong các ứng dụng.

iii) Tham số m đặc trưng cho độ mờ của kết quả phân tích chùm, khi m = 1 phân tích chùm mờ trở thành không mờ, khi m tiến đến vô cùng, xác suất để các phần tử thuộc vào các chùm bằng nhau và bằng 1/c. Hiện tại, chúng ta chưa có phương pháp tối ưu trong xác định m (Yu *et al*., 2004; Thao và Tai, 2016). Việc xác định m một cách cụ thể vẫn thường được thực hiện bằng phương pháp chia lưới (Hall *et al*., 1992). Chúng tôi cũng xác định m theo phương pháp chia lưới.

Trong bài viết này, phương pháp Bayes khi sử dụng các xác suất tiên nghiệm (i), (ii), (iii) và thuật toán đề nghị lần lượt được gọi là BayesU, BayesP, BayesL và BayesC.

2.5 Phương pháp SVM

Cho tập mẫu $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, với x_i thuộc R^n, y_i nhận 2 giá trị $\{-1, 1\}$ với -1 biểu thị lớp I, 1 biểu thị lớp II.

Ta có phương trình siêu phẳng chứa vector $\vec{x}$ trong không gian như sau: $\vec{x}_l \vec{w} + b = 0$.

Đặt

$$f(\vec{x}_l) = sign(\vec{x}_l \vec{w} + b) = \begin{cases} +1 & \text{khi } \vec{x}_l . \vec{w} + b > 0. \\ -1 & \text{khi } \vec{x}_l . \vec{w} + b < 0. \end{cases}$$

Như vậy, $f(\vec{x}_l)$ biểu diễn sự phân lớp của $\vec{x}_l$ vào hai lớp như đã nêu.

Ta xếp $\vec{x}_l$ thuộc lớp I nếu $y_i = +1$ và thuộc lớp II nếu và $y_i = -1$.

3 VẤN ĐỀ TÍNH TOÁN

3.1 Trong phương pháp Fisher, hồi qui logistic và SVM

i) Đối với phương pháp Fisher, do thực tế không có véc tơ trung bình và ma trận hiệp phương sai của tổng thể, nên ta thay thế chúng bằng các ước lượng không chệch từ mẫu. Trong R^n, giả sử chúng ta có k mẫu tương ứng k tổng thể, với mẫu thứ i có kích thước n_i, $\sum_{i=1}^{k} n_i = N$, có ma trận dữ

liệu X_i mà cột thứ j là x_{ij}. Gọi S_i là ma trận hiệp phương sai của tổng thể thứ i. Đặt:

$$\bar{x}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij},$$

$$S_i = \frac{1}{n_i - 1}\sum_{i=1}^{k}(x_{ij} - \bar{x}_i)(x_{ij} - \bar{x}_i)^T,$$

$$S = \frac{\sum_{i=1}^{k}(n_i - 1)S_i}{\sum_{i=1}^{k}(n_i - k)}.$$

Lúc này ta sẽ thay thế μ_i bằng $\bar{x}_i$, Σ bởi S trong công thức (3).

Chúng ta có thể sử dụng các phần mềm thống kê R hoặc SPSS để thực hiện bài toán phân loại bằng phương pháp Fisher.

ii) Để tìm các hệ số của mô hình hồi qui logistic khi có số liệu cụ thể, ta phải giải hệ phương trình (2). Tuy nhiên, việc giải hệ phương trình này thực sự rất phức tạp, vì vậy trong thực hành ta sử dụng các gói hỗ trợ của các phần mềm thống kê như SPSS, R,... để thực hiện. Đối với phương pháp SVM chúng tôi sử dụng phần mềm Weka để thực hiện.

3.2 Trong phương pháp Bayes

i) Trong thực tế, dữ liệu là rời rạc, vì vậy để đảm bảo tính ứng dụng thực tế của phương pháp, đầu tiên chúng ta cần phải ước lượng hàm mật độ xác suất từ dữ liệu rời rạc này. Có nhiều phương pháp ước lượng tham số cũng như phi tham số để thực hiện. Trong bài viết này, chúng tôi sử dụng phương pháp hàm hạt nhân, một phương pháp cho đến hiện tại được đánh giá có nhiều ưu điểm hơn các phương pháp khác. Hàm mật độ n chiều ước lượng bằng phương pháp này có dạng:

$$\hat{f}(x) = \frac{1}{Nh_1h_2...h_n}\sum_{i=1}^{N}\prod_{j=1}^{n} K_j\left(\frac{x_i - x_{ij}}{h_j}\right),$$

trong đó h_j là tham số trơn cho biến thứ j, K_j là hàm hạt nhân của biến thứ j, x_i là chiều thứ i, x_{ij} là số liệu thứ i của biến thứ j, N là số phần tử của mẫu và n là số chiều của dữ liệu.

Có thể chọn nhiều hàm hạt nhân khác nhau như dạng tam giác, hình chữ nhật, song lượng... Trong bài báo này, chúng tôi chọn hàm hạt nhân dạng chuẩn:

$$f(x) = \frac{1}{\sqrt{2\pi}}\exp(-x^2/2).$$

Có nhiều nghiên cứu về việc chọn tham số trơn và cũng chưa có kết luận cuối cùng nào chứng tỏ cách chọn tham số này là thực sự tốt hơn so với cách khác. Trong bài viết này, chúng tôi chọn tham số trơn theo Scott (1992):

$h_j = \left(\frac{4}{N(n+2)}\right)^{\frac{1}{n+4}}\sigma_j$, trong đó σ_j là độ lệch chuẩn mẫu của biến thứ j, n và N lần lượt là số chiều và số phần tử của mẫu.

Các phần mềm thống kê như Matlab, Maple... đã hỗ trợ việc ước lượng hàm mật độ xác suất 1 chiều, tuy nhiên trong trường hợp nhiều chiều chưa có sự hỗ trợ. Trong bài viết này, chúng tôi đã viết chương trình thực hiện trên phần mềm Matlab với hàm hạt nhân và tham số trơn được chọn ở trên.

ii) Dựa vào nguyên tắc (4), chúng tôi cũng đã viết chương trình để phân loại một phần tử mới, chương trình xác định xác suất tiên nghiệm và chương trình tính sai số Bayes, trong đó tích phân được ước lượng theo phương pháp Monte Carlo. Các chương trình này được dùng trong các áp dụng thực tế ở phần 4.

4 ĐÁNH GIÁ KHẢ NĂNG TRẢ NỢ VAY CỦA KHÁCH HÀNG

4.1 Giới thiệu

Trong phần này, dựa trên các số liệu thực tế thu được và lý thuyết đã trình bày, chúng tôi thực hiện việc đánh giá khả năng trả nợ vay của khách hàng trên địa bàn thành phố Cần Thơ. Đối tượng khách hàng được khảo sát là các doanh nghiệp hoạt động trên các lĩnh vực quan trọng: nông nghiệp, công nghiệp và thương mại. Số liệu thực hiện gồm 214 doanh nghiệp, trong đó 143 doanh nghiệp trả nợ được đúng hạn (TN) và 71 không trả nợ được đúng hạn (KTN). Số liệu nghiên cứu được cung cấp bởi cơ quan có trách nhiệm quản lý trên địa bàn thành phố Cần Thơ năm 2013, trong một đề tài nghiên cứu về doanh nghiệp trên địa bàn. Mỗi doanh nghiệp được đánh giá bởi 13 biến theo ý kiến ban đầu của chuyên gia ngân hàng. Các biến cụ thể được cho bởi Bảng 1 như sau:

Bảng 1: Các biến khảo sát trong áp dụng

Xi	Biến khảo sát	Giải thích các biến
X1	Đonbaytaichinh	Tổng nợ/tổng vốn chủ sở hữu
X2	Dongtientudo	Thu nhập giữ lại /tổng tài sản
X3	Roe	Lợi nhuận ròng/vốn chủ sở hữu
X4	Dongtien	(Lợi nhuận ròng + khấu hao)/tổng tài sản
X5	Vonluudong	(Tài sản ngắn hạn - nợ ngắn hạn)/tổng tài sản
X6	Thankhoan	(Tiền + đầu tư ngắn hạn)/nợ ngắn hạn
X7	Loinhuan	Lợi nhuận ròng/tổng tài sản
X8	Knanghoatdong	Doanh thu/tổng tài sản
X9	Qymo	Logarit của tổng tài sản
X10	Kinhnghiem	Số năm hoạt động của doanh nghiệp
X11	Nongnghiep	Ngành nông nghiệp và lâm nghiệp
X12	Congnghiep	Công nghiệp và xây dựng
X13	Thuongmai	Thương mại và dịch vụ

4.2 Phương pháp thực hiện

Từ số liệu, chúng tôi lần lượt thực hiện các bước sau:

i) Xác định các biến có ý nghĩa thống kê với mức ý nghĩa 10% trong đánh giá khả năng trả được nợ vay của các doanh nghiệp qua mô hình hồi qui logistic.

ii) Sử dụng các biến có ý nghĩa đã xác định từ i), kiểm tra sự khác biệt giữa hai nhóm TN và KTN bằng phương pháp Hotelling.

iii) Chia tập dữ liệu thành hai phần: Tập huấn luyện và tập kiểm tra, trong đó 70% số liệu được chọn ngẫu nhiên từ mỗi nhóm (100 doanh nghiệp thuộc nhóm TN và 50 doanh nghiệp thuộc nhóm KTN) được sử dụng cho tập huấn luyện để xác định mô hình tối ưu, 30% dữ liệu còn lại (43 doanh nghiệp thuộc nhóm TN và 21 doanh nghiệp thuộc nhóm KTN) được sử dụng cho tập kiểm tra.

iv) Với tập huấn luyện, chúng tôi sử dụng tất cả các phương pháp phân loại Fisher, logistic, SVM, BayesU, BayesP, BayesL và BayesC để phân loại hai nhóm doanh nghiệp TN và KTN. Trong mỗi phương pháp, xác suất phân loại đúng sẽ được tính để làm tiêu chuẩn lựa chọn mô hình tối ưu.

v) Sử dụng mô hình tối ưu từ mỗi phương pháp đã rút ra từ iv), thực hiện phân loại cho tập kiểm tra, tính tỉ lệ sai lầm khi thực hiện của mỗi phương pháp để so sánh.

Từ dữ liệu rời rạc, phân loại bằng phương pháp Fisher và logistic sẽ được thực hiện bằng phần mềm SPSS. Đối với phương pháp SVM, việc thực hiện được dựa vào phần mềm Weka (http://download.phanmem.com/weka-3.7.8-NM3P98.html). Trong phương pháp Bayes, ước lượng hàm mật độ xác suất từ dữ liệu rời rạc sẽ được thực hiện đầu tiên. Gọi $f_1(x), f_2(x)$ lần lượt là hàm mật độ xác suất ước lượng cho nhóm không trả nợ được và nhóm trả nợ được. Các xác suất tiên nghiệm khác nhau trong phương pháp Bayes sẽ được thực hiện để tìm trường hợp phù hợp nhất (sai số Bayes nhỏ nhất). Gọi $(q_1^{(i)}, q_2^{(i)}), i = 1,2,3,4$ lần lượt là xác suất tiên nghiệm khi sử dụng phân phối đều, phương pháp Laplace, phương pháp tỉ lệ mẫu và phương pháp được đề nghị. Kết quả tối ưu trong thực hiện bằng phương pháp Bayes sẽ được so sánh với các kết quả khi áp dụng các phương pháp logistic, Fisher và SVM.

4.3 Kết quả thực hiện

Khảo sát các biến có ý nghĩa thống kê qua mô hình hồi qui logistic để đánh giá khả năng trả được nợ vay của các doanh nghiệp, ta có bảng tổng kết sau:

Bảng 2: Hệ số và giá trị Sig của các biến trong mô hình logistic

Xi	Hệ số hồi qui	Sig
X1	-2.444	0.003
X2	6.692	0.244
X3	2.566	0.244
X4	2.052	0.034
X5	0.478	0.700
X6	0.340	0.860
X7	4.921	0.093
X8	0.044	0.621
X9	-0.329	0.442
X10	-0.136	0.910
X11	0.009	0.994
X12	-0.007	0.886
X13	2.122	0.360

Bảng 2 cho thấy chỉ có ba biến X1, X4 và X7 có ý nghĩa thống kê, trong đánh giá khả năng trả nợ vay của các doanh nghiệp.

Kiểm định sự khác biệt của 2 nhóm TN và KTN với 3 biến trên bằng phương pháp Hotelling, ta thấy có sự khác biệt của hai nhóm này.

Sử dụng các phương pháp phân loại, với tất cả các trường hợp khác nhau của 1 biến, 2 biến và 3 biến với tất cả các trường hợp của xác suất tiên nghiệm, chúng ta có bảng tổng hợp sau:

Bảng 3: Bảng tổng hợp xác suất phân loại đúng của các phương pháp

Biến	BayesU	BayesP	BayesL	BayesC	Fisher	Logistic	SVM
X1	0.7673	0.8613	0.8128	0.9072	0.8123	0.8115	0.7510
X4	0.7527	0.8290	0.7355	0.8390	0.6205	0.7592	0.7667
X7	0.7213	0.8462	0.7723	0.8759	0.7176	0.7873	0.8520
X1,X4	0.8032	0.8874	0.8576	0.9310	0.7887	0.8017	0.9012
X1,X7	0.8192	0.8832	0.8634	0.9013	0.8337	0.8568	0.8512
X4,X7	0.8017	0.8903	0.8325	0.9281	0.6428	0.8257	0.8833
X1,X4,X7	0.9016	0.9254	0.9012	**0.9517**	0.7657	0.8357	0.9167

Bảng 3 cho thấy việc sử dụng 2 biến X1 và X7 cho kết quả phân loại đúng cao nhất đối với phương pháp Fisher và logistic. Trong khi đó, phương pháp Bayes cho kết quả tốt nhất khi sử dụng 3 biến. Phương pháp Bayes trong các trường hợp luôn cho kết quả tốt và ổn định hơn các phương pháp khác. Hơn nữa, BayesC luôn cho kết quả ổn định và tốt nhất. Đặc biệt BayesC, khi sử dụng 3 biến cho ta kết quả phân loại đúng rất cao (95.17%).

Sử dụng các mô hình tối ưu cho mỗi phương pháp có được từ tập huấn luyện, tiến hành phân loại cho 64 các doanh nghiệp của tập kiểm tra, ta có Bảng 4 tổng kết tỉ lệ phân loại đúng của mỗi phương pháp như sau:

Bảng 4: Tỉ lệ phân loại đúng của các phương pháp với tập kiểm tra

	Số phần tử phân loại sai	Số phần tử phân loại đúng	Tỉ lệ phân loại đúng
Fisher	12	52	0.813
Logistic	11	53	0828
SVM	11	53	0.858
BayesC	8	56	0.975

Một lần nữa BayesC cho kết quả phân loại đúng cao nhất. Theo đánh giá của những người làm trong lĩnh vực tín dụng ngân hàng, kết quả phân loại cho tập huấn luyện và kiểm tra trong trường hợp này là một kết quả tốt. Nó có thể làm kênh tham khảo định lượng trong đánh giá ban đầu, việc chấp nhận cho vay hay không của cán bộ tín dụng.

5 KẾT LUẬN

Bài báo đã trình bày các phương pháp phân loại và vấn đề tính toán của chúng, trong đó đã đề nghị thuật toán xác định xác suất tiên nghiệm trong phân loại bằng phương pháp Bayes. Thuật toán này đã chứng minh ưu điểm, khi làm giảm được xác suất sai lầm phân loại trong tất cả các trường hợp với bộ số liệu thực tế được khảo sát. Bài báo đã xem xét vấn đề tính toán trong áp dụng thực tế của các phương pháp, trong đó đã thiết lập các chương trình để giải quyết vấn đề tính toán của phương pháp Bayes với thuật toán tìm xác suất tiên nghiệm đề nghị. Trong việc cho vay, cán bộ tín dụng phải áp dụng nhiều biện pháp nghiệp vụ định lượng và định tính khác nhau, trong đó theo chúng tôi, việc sử dụng bài toán phân loại là một kênh tham khảo định lượng cần thiết, rất đáng quan tâm. Chúng tôi nghĩ rằng đây là vấn đề thú vị, có tiềm năng ứng dụng rất lớn trong thực tế, không những trong lĩnh vực ngân hàng mà còn nhiều lĩnh vực khác. Trong thời gian tới, chúng tôi sẽ tiếp tục áp dụng cách làm này để phân loại bệnh trong y học.

TÀI LIỆU THAM KHẢO

Bora, D. J. and Gupta, A. K., 2014. Impact of exponent parameter value for the partition matrix on the performance of fuzzy C means algorithm. ArXiv preprint arXiv. 3(3): 1953-1967.

Inman, H. F. and Bradley, E. L., 1989. The overlapping coefficient as a measure of agreement between probability distributions and point estimation of the overlap of two normal densities. Communications in Statistics -Theory Methods. 18(10): 3851-3871.

Hall L. O., Bensaid A.M., Clarke, L.P. and Velthuizen, R.P., 1992. A comparison of neural network and fuzzy clustering techniques in segmenting magnetic resonance images of the brain. IEEE Transactions. 3(5): 672-682.

Miller, G., Inkret, W.C., Little, T.T., Martz, H.F., and Schillaci, M.E., 2011. Bayesian prior probability distributions for internal dosimetry Radiation Protection Dosimetry. 94(4): 347-352.

Pham–Gia, T., Turkkan, N. and Tai, Vovan., 2008. The maximum function in statistical discrimination analysis. Commun. in Stat–Simulation computation. 37(2): 320-336.

Scott, D. W., 1992. Mutivariate density estimation: Theory, practice and visualization. Wiley & Son, New York, 345 pages.

Tai, V.V., 2016. L^1-distance and classification problem by Bayesian. J. Appl. Stat (online first: http://dx.doi.org/10.1080/02664763.2016.1174194).

Thao, N.T., Tai, V.V., 2016. A new approach for determining the prior probabilities in the classification problem by Bayesian method, Adv. Data Anal. Classif. (online first: http://link.springer.com/article/10.1007/s11634-016-0253).

Webb, A., 2000. Statistical pattern recognition. Wiley & Sons, New York, 645 pages.

Yu, J., Qiansheng, C. and Houkuan, H., 2004. Analysis of the weighting exponent in the FCM. IEEE Transactions on Cybernetics. 34(1): 634-639.