TẠP CHÍ
Ứng Dụng Toán Học
Tập 4, Số 2, 2006, 01-16

PHƯƠNG PHÁP THỐNG KÊ XÂY DỰNG MÔ HÌNH ĐỊNH MỨC TÍN NHIỆM KHÁCH HÀNG THỂ NHÂN

VƯƠNG QUÂN HOÀNG
Centre Emile Bernheim, ULB, 21 F.D.Roosevelt, B-1050, Bruxelles

ĐÀO GIA HƯNG
Ngân hàng Techcombank

NGUYỄN VĂN HỮU, TRẦN MINH NGỌC, LÊ HỒNG PHƯƠNG
Khoa Toán-cơ-tin học, trường Đại học KHTN, Đại học Quốc gia Hà Nội

Người giới thiệu: Tống Đình Quỳ

TÓM TẮT. Chúng tôi sử dụng các phương pháp thống kê để xây dựng mô hình định mức tín nhiệm khách hàng thể nhân, tức là đánh giá mức tín nhiệm của một khách hàng tín dụng. Các phương pháp thống kê chính được dùng ở đây là lý thuyết phân lớp và phân tích phân biệt. Chúng tôi minh hoạ phương pháp trên một tập dữ liệu về các khách hàng tín dụng của một ngân hàng thương mại.

Từ khóa: phân lớp, phân tích phân biệt, hồi quy phi tuyến với biến phụ thuộc nhị nguyên, định mức tín nhiệm, kiểm định giả thiết

1 Giới thiệu

Mô hình định mức tín nhiệm thể nhân được đặt ra cách đây 50 năm nhằm xây dựng phương pháp lượng hoá khả năng thanh toán và mức độ tín nhiệm của khách hàng trong giao dịch. Công tác này giúp các ngân hàng và tổ chức tín dụng quyết định có hay không cung cấp các dịch vụ cho khách hàng. Lợi ích của mô hình đem lại rất rõ nét, nổi bật là giảm thiểu chi phí phân tích thông tin (nhất là khi số lượng người sử dụng các dịnh vụ ngân hàng ngày càng lớn); giúp đưa ra quyết định nhanh chóng, chính xác và khách quan; giảm thiểu rủi ro tín dụng, đảm bảo tối đa việc thu hồi tài chính.

Một trong các phương pháp tiếp cận mô hình định mức tín nhiệm khách hàng là giải quyết bài toán phân tích phân biệt, nhận biết hay là xếp một cá thể vào một trong các nhóm khách hàng mà có sự khác nhau tương đối giữa các nhóm. Bài toán phân nhóm một tập hợp được Fisher giới thiệu lần đầu tiên vào năm 1936 khi tiến hành phân loại đặc tính cây Irit dựa trên số liệu về kích thước bên ngoài của cây. David Duran (1941) là người đầu tiên ứng dụng phương pháp đó vào việc phân biệt các

khoản nợ tốt và khoản nợ xấu. Sau đó nhiều công ty tín dụng đã xây dựng các hình thức sơ khai của hệ thống định mức tín nhiệm thể nhân dựa trên các nguyên lý thống kê, và các hệ thống này đã nhanh chóng tỏ rõ sức mạnh của nó trong việc giúp các tổ chức tín dụng ra quyết định. Sự kiện đánh dấu tầm quan trọng của mô hình định mức tín nhiệm thể nhân là việc thông qua đạo luật Cơ Hội Tín Dụng Ngang Bằng ở Mỹ năm 1975-1976, nội dung chủ yếu của đạo luật này là cấm sự phân biệt đối xử trong việc cấp tín dụng trừ khi nó được chứng minh trên cơ sở thống kê.

Có thể hình dung mô hình như sau. Mỗi khách hàng đến giao dịch xin cấp tín dụng sẽ được yêu cầu cung cấp các thông tin bản thân. Thông tin là một vector k-chiều (k dấu hiệu) $X = (X_1, ..., X_k)$ bao gồm các dấu hiệu như tuổi tác, trình độ học vấn, mức thu nhập, tình trạng hôn nhân, chênh lệch thu chi, dư nợ hiện tại,... Phương pháp chúng tôi đề xuất (gọi là phương pháp I) để giải quyết bài toán định mức tín nhiệm thể nhân sẽ bao gồm các bài toán

1. Xác định các dấu hiệu nên đưa vào để lấy thông tin về khách hàng, nên hay không nên đưa vào dấu hiệu nào?

2. Xây dựng thang điểm cho các dấu hiệu.

3. Từ mẫu N khách hàng, phân chia thành các nhóm, chẳng hạn "tốt", "tốt vừa", "xấu",.. Đây chính là nội dung của bài toán phân loại.

4. Với một khách hàng X, xây dựng quy tắc ra quyết định xếp X vào nhóm nào? Và đây chính là nội dung của bài toán phân tích phân biệt.

Chú ý. Chúng ta có thể xét phương pháp khác (sẽ gọi là phương pháp II), mà khác cơ bản phương pháp I: Bài toán 1 và 2 như trên và

3'. Xác định trọng số cho mỗi dấu hiệu, trọng số này đặc trưng cho tầm quan trọng của dấu hiệu đó đối với khả năng thanh toán của khách hàng. Giả sử β_l là trọng số của dấu hiệu X_l, và nếu gọi $s(X)$ là hàm điểm tín dụng của khách hàng $X = (X_1, ..., X_k)$ thì

$$s(X) = \beta_1 X_1 + ... + \beta_k X_k.$$

4'. Xây dựng mô hình ra quyết định tín dụng dựa trên hàm điểm tín dụng $s(X)$.

Với bài toán 1, yêu cầu đầu tiên về các dấu hiệu đưa vào là các dấu hiệu không tương quan với nhau, sau đó là yêu cầu đưa vào các dấu hiệu sao cho đặc trưng được nhiều nhất thông tin về khả năng tín dụng của khách hàng. Sau cùng có thể tính đến các yêu cầu như các dấu hiệu đó giúp khách hàng dễ trả lời, ngân hàng dễ chứng thực tính đúng đắn,... Ví dụ tại ngân hàng Techcombank các dấu hiệu được đưa vào như: tuổi tác, trình độ học vấn, loại hình công việc, mức thu nhập, chênh lệch thu chi, tình trạng hôn nhân, số người sống phụ thuộc, nơi cư trú, thời gian cư trú, phương tiện đi lại, phương tiện thông tin, uy tín trong giao dịch, quan hệ với Techcombank, dư nợ,...

Bài toán thứ 2 sẽ rất quan trọng nếu chúng ta xét phương pháp II bởi nó ảnh hưởng rất nhiều đến hàm điểm tín dụng $s(X)$ và nó đòi hỏi nhiều kỹ thuật phức tạp trong việc lập thang điểm cho mỗi dấu hiệu. Tuy nhiên với phương pháp I, bài toán này có lẽ không đòi hỏi các kỹ thuật tinh tế lắm, bởi ta chỉ cần xác định thang điểm sao cho dẫn

đến sự khác nhau tương đối giữa các nhóm khách hàng mà sẽ được phân lớp trong bài toán 3.

Trong các bài toán được đặt ra trên có thể nói bài toán 3 và bài toán 4 là quan trọng nhất và cũng phức tạp nhất. Trong bài báo này chúng tôi tập trung giải quyết hai bài toán đó.

Cấu trúc bài báo như sau. Mục 2 giải quyết bài toán 3, bài toán phân lớp khách hàng. Mục 3 trình bày lời giải bài toán 4: xây dựng quy tắc đánh giá mức tín nhiệm khách hàng. Mục 4 trình bày các kết quả tính toán từ dữ liệu các khách hàng của ngân hàng Techcombank cùng với một vài nhận xét và bình luận.

2 Phân lớp khách hàng

Xét một mẫu gồm N khách hàng (cá thể), khách hàng thứ i có vector dấu hiệu là $X^{(i)} = (X_{i1}, ..., X_{ik}), i = 1, ..., N$.

Việc phân nhóm các cá thể sẽ được thực hiện dựa trên khái niệm khoảng cách đo sự khác nhau giữa các cá thể, ta sẽ ký hiệu $d(i, j)$ là khoảng cách giữa cá thể thứ i và thứ j dựa trên dấu hiệu $X^{(i)}, X^{(j)}$ tương ứng. Có nhiều định nghĩa cho khoảng cách giữa các cá thể, thường sử dụng các khoảng cách sau:

Khoảng cách Euclide

$$d_1(i, j) = \left\{ \sum_{l=1}^{k} (X_{il} - X_{jl})^2 \right\}^{1/2}.$$

Khoảng cách thống kê

$$d_2(i, j) = \left\{ (X^{(i)} - X^{(j)}) A (X^{(i)} - X^{(j)})^T \right\}^{1/2}$$

trong đó A là một ma trận đối xứng xác định dương cấp N, và thường được chọn là S^{-1} với S là ma trận hiệp phương sai mẫu.

Khoảng cách định tính

$$d_3(i, j) = \frac{1}{1 + s(i, j)}$$

trong đó

$$s(i, j) = \frac{\sum_{l=1}^{k} X_{il} \delta(X_{il} - X_{jl})}{\sum_{l=1}^{k} X_{il} \delta(X_{il} - X_{jl}) + \sum_{l=1}^{k} (1 - \delta(X_{il} - X_{jl}))},$$

với $\delta(x - y) = 1$ nếu $x = y$ và 0 nếu $x \neq y$, là hệ số tương tự đo sự gần nhau của cá thể i và j.

Nhận xét. d_3 không phải là một khoảng cách thông thường, bởi d_3 không thoả mãn tính chất $d(i, i) = 0 \ \forall i$, tuy nhiên nó vẫn đặc trưng cho sự "xa nhau" giữa các cá thể. Khoảng cách d_1 là trường hợp đặc biệt của d_2 khi A là ma trận đơn vị. Với việc chọn ma trận A thích hợp, khoảng cách thống kê d_2 sẽ làm mất đi sự phụ thuộc vào thứ nguyên của các dấu hiệu, tức là d_2 dùng tốt trong trường hợp các dấu hiệu có nhiều thứ nguyên khác nhau. Khoảng cách d_1, d_2 thường được dùng để tính toán cho các dấu hiệu định lượng, còn d_3 được dùng với các dấu hiệu định tính.

Ta ký hiệu

$$D = (d(i,j))_{i,j=1,...,N}$$

là ma trận khoảng cách. Có nhiều phương pháp phân lớp dựa trên ma trận khoảng cách D, như phương pháp phân lớp theo thứ bậc, phương pháp K-trung bình. Theo kinh nghiệm của chúng tôi, trong trường hợp này nên dùng phương pháp K-trung bình, khi đó các nhóm kết quả nhận được sẽ khác nhau tương đối về bản chất, đặc trưng cho các nhóm khách hàng "tốt", "xấu".

Phương pháp K-trung bình được J. B. MacQueen đưa ra năm 1967. Thuật toán có 3 bước

1. Phân chia (ngẫu nhiên) các cá thể vào K nhóm.
2. Tính tâm của từng nhóm. Phân phối lại các cá thể: xếp một cá thể vào nhóm có tâm gần nó nhất. Có nhiều khái niệm tâm của nhóm, và thường là vector trung bình các dấu hiệu của nhóm, còn khoảng cách thường dùng là khoảng cách Euclide.
3. Lặp lại bước 2 cho đến khi không còn sự phân phối lại các cá thể.

Một vấn đề đặt ra là khi nào hai lớp được xem là đủ khác nhau? Hay nói cách khác, chúng ta cần phải thực hiện bài toán kiểm định sự khác nhau giữa các lớp. Xét hai lớp A và B với các cá thể của lớp A là

$$(x_{j1}, ..., x_{jk}), j = 1, ..., n_1$$

và các cá thể của lớp B là

$$(y_{j1}, ..., y_{jk}), j = 1, ..., n_2.$$

Gọi $\overline{X}, \overline{Y}$ lần lượt là tâm của nhóm A và B:

$$\overline{X} = (\overline{x}_1, ..., \overline{x}_k),\ \overline{Y} = (\overline{y}_1, ..., \overline{y}_k)$$

trong đó

$$\overline{x}_l = \frac{1}{n_1}\sum_{j=1}^{n_1} x_{jl},\ \overline{y}_l = \frac{1}{n_2}\sum_{j=1}^{n_2} y_{jl}, l = 1, ..., k.$$

Đặt

$$S^{(1)} = (s_{ij}^{(1)})_{i,j=1,...,k},\ S^{(2)} = (s_{ij}^{(2)})_{i,j=1,...,k}$$

lần lượt là ma trận hiệp phương sai mẫu của hai nhóm, trong đó

$$s_{ij}^{(1)} = \frac{1}{n_1}\sum_{l=1}^{n_1} x_{il}x_{jl} - \overline{x}_i\overline{x}_j,\ s_{ij}^{(2)} = \frac{1}{n_2}\sum_{l=1}^{n_2} y_{il}y_{jl} - \overline{y}_i\overline{y}_j.$$

Xét khoảng cách Hotelling được định nghĩa bởi

$$T^2 = (\overline{X} - \overline{Y})^T S^{-1}(\overline{X} - \overline{Y})$$

trong đó

$$S = \frac{1}{n_1 + n_2}[n_1 S^{(1)} + n_2 S^{(2)}].$$

Người ta chứng minh được rằng nếu hai nhóm A, B là một nhóm thì khi n_1, n_2 lớn T^2 sẽ có phân phối xấp xỉ phân phối χ^2 với k bậc tự do. Từ đó ta có quy tắc sau: **Nếu $T^2 > \chi_k^2(\alpha)$ thì hai lớp A, B được coi là tách biệt nhau một cách có ý nghĩa.**

3 Phân biệt khách hàng

Dựa trên kết quả phân lớp trong mục trên, trong mục này chúng tôi giải quyết bài toán tiếp theo: Với một khách hàng có vector dấu hiệu x, xây dựng quy tắc xếp nhóm cho khách hàng đó. Chúng tôi trình bày hai phương pháp giải quyết bài toán đó trong hai mục tương ứng, Mục 3.1 và Mục 3.2.

3.1 Phương pháp hồi quy với biến phụ thuộc nhị nguyên

Giả sử tập các khách hàng được đánh số $1, 2, \ldots, N$ đã được phân chia thành 2 nhóm A và B. Dấu hiệu X_l nhận giá trị trong tập hữu hạn $E_l = \{e_{l1}, e_{l2}, \ldots, e_{lm_l}\}, l = 1, ..., k$. Nhóm A gồm các khách hàng "tốt", nhóm B gồm các khách hàng "không tốt". Đặt

$$\pi = \frac{\text{số cá thể thuộc nhóm } A}{N}$$

là tỉ lệ khách hàng thuộc nhóm A; $1 - \pi$ là tỉ lệ khách hàng thuộc nhóm B.

Ta có thể dùng biến Z để đặc trưng cho khách hàng thuộc nhóm A hoặc nhóm B:

$$Z = \begin{cases} 1, & \text{nếu khách hàng thuộc nhóm } A, \\ 0, & \text{nếu khách hàng thuộc nhóm } B. \end{cases}$$

Như vậy khách hàng thứ i sẽ có đặc trưng là Z_i với

$$Z_i = \begin{cases} 1, & \text{nếu } i \in A, \\ 0, & \text{nếu } i \in B. \end{cases}$$

Giả sử $x = (x_1, x_2, \ldots, x_k)$ là véc-tơ dấu hiệu của một khách hàng. Ta cần tính xác suất sau:

$$P(Z = 1|X = x) := P(x), \tag{1}$$

đây là xác suất khách hàng có vector dấu hiệu x thuộc nhóm A.

Ta có công thức sau

$$\begin{aligned} P(x) &= \frac{P(Z = 1).P(X = x|\text{cá thể thuộc nhóm } A)}{P(X = x)} \\ &= \frac{\pi P(X = x|A)}{\pi P(X = x|A) + (1 - \pi)P(X = x|B)}, \end{aligned} \tag{2}$$

trong đó kí hiệu

$$P(X = x|A) = P(X = x|\text{cá thể thuộc nhóm } A).$$

Chúng ta có

$$P(Z = 0|X = x) = 1 - P(x).$$

Ta cần ước lượng xác suất $P(x)$ dựa trên mẫu $(Z_i, X^{(i)}), i = 1, 2, \ldots, N$. Người ta thấy rằng $P(x)$ có dạng

$$P(x) = 1 - F(-\beta^T x), \quad \text{với } \beta^T x = \sum_{i=1}^{k} \beta_i x_i, \tag{3}$$

trong đó $F(y)$ là hàm phân bố xác suất nào đó, $\beta = (\beta_1, \ldots, \beta_k)^T$ là các tham số phải ước lượng.

Xét mô hình hồi quy phi tuyến sau đây:

$$Z_i = 1 - F(-\beta^T X^{(i)}) + \epsilon_i, \quad i = 1, 2, \ldots, N, \tag{4}$$

trong đó ϵ_i là sai số ngẫu nhiên với $E\epsilon_i = 0$.

Có thể coi (4) là mô hình thực nghiệm của mô hình lí thuyết sau đây :

$$Z = 1 - F(-\beta^T X) + \epsilon, \quad E\epsilon = 0.$$

Do đó

$$E(Z|X) = P(Z = 1|X) = 1 - F(-\beta^T X).$$

Ta sẽ ước lượng véc-tơ β bằng phương pháp hợp lí cực đại, tức tìm $\widehat{\beta}$ sao cho

$$\log L(\beta) \quad := \quad \sum_{i=1}^{N} \left[Z_i \log(1 - F(-\beta^T X^{(i)})) + (1 - Z_i) \log F(-\beta^T X^{(i)}) \right] \tag{5}$$

đạt giá trị cực đại.

Các hàm phân bố sau đây thường được dùng trong (4) và (5):

- Hàm phân bố chuẩn $F(x) = \dfrac{1}{\sqrt{2\pi}} \displaystyle\int_{-\infty}^{x} e^{-t^2/2} dt$
- Hàm phân bố logistic $F(x) = \dfrac{e^x}{1 + e^x}$
- Hàm phân bố Weibul $F(x) = 1 - \exp(-x)$.

Trong công trình này, chúng tôi sử dụng F là hàm phân bố logistic vì nó thích hợp với các biến rời rạc.

Sau khi tìm được ước lượng $\widehat{\beta}$ của β ta thu được

$$\widehat{P}(x) = 1 - F(-x^T \widehat{\beta}), \tag{6}$$

và

$$\widehat{\epsilon_i} = Z_i - \widehat{P}(X^{(i)}), \quad i = 1, 2, \ldots, N \tag{7}$$

là các phần dư.

Giả sử một phần tử mới có véc-tơ dấu hiệu là X, khi đó ta gán cá thể đó vào lớp A nếu $\widehat{P}(x) > 0.5$ và vào lớp B nếu $\widehat{P}(x) \leq 0.5$.

Mỗi nhóm A và B lại có thể phân thành các nhóm con, ví dụ theo quy tắc sau: Gán phần tử có dấu hiệu X vào

- lớp A_1 nếu $\widehat{P}(x) > 0.8$
- lớp A_2 nếu $0.65 < \widehat{P}(x) \leq 0.8$
- lớp A_3 nếu $0.5 < \widehat{P}(x) \leq 0.65$

- lớp B_1 nếu $0.35 < \widehat{P}(x) \leq 0.5$
- lớp B_2 nếu $0.2 < \widehat{P}(x) \leq 0.35$
- lớp B_3 nếu $0 < \widehat{P}(x) \leq 0.2$

Để đánh giá hiệu năng của quy tắc phân biệt khách hàng, ta tính các đại lượng sau

- Tỷ lệ phân biệt đúng
  - Tỷ lệ cá thể thuộc lớp B với $\widehat{P}(X^{(i)}) \leq 0.5$
  - Tỷ lệ cá thể thuộc lớp A với $\widehat{P}(X^{(i)}) > 0.5$
- Số trung bình các cá thể phân biệt đúng
  - Số trung bình các cá thể thuộc lớp B có $\widehat{P}(X^{(i)}) \leq 0.5$
  - Số trung bình cá thể thuộc lớp A với $\widehat{P}(X^{(i)}) > 0.5$

Để đánh giá sự đóng góp của các biến vào xác suất $P(x) = 1 - F(-\beta^T x)$, ta chú ý rằng nếu $f(x) = F'(x)$ là hàm mật độ của hàm phân bố $F(x)$ thì

$$\frac{\partial P}{\partial x_i} = f(-\beta^T x)\beta_i. \tag{8}$$

Như vậy, nếu $\beta_i > 0$ thì x_i góp phần làm tăng $P(x)$ khi x_i tăng. Ngược lại, nếu $\beta_i < 0$ thì x_i góp phần làm giảm $P(x)$ khi x_i tăng.

Hơn nữa, ta có

$$\frac{\partial P/\partial x_i}{\partial P/\partial x_j} = \frac{\beta_i}{\beta_j}. \tag{9}$$

Do đó tác động của biến x_i sẽ cao hơn tác động của biến x_j nếu $|\beta_i| > |\beta_j|$.

Danh sách các đặc trưng của mỗi khách hàng của Techcombank và các kết quả về ước lượng tham số β và sau đó ước lượng xác suất $P(x)$ cũng như việc đánh giá hiệu năng của quy tắc phân biệt khách hàng được tổng kết trong Mục 4.

3.2 Thuật toán phân biệt khách hàng với các dấu hiệu định tính và định lượng

Giả sử $X^{(i)} = (X_{i1}, \ldots, X_{im})$ là véc-tơ dấu hiệu của khách hàng thứ i, với $i = 1, 2, \ldots, N$, trong đó có r thành phần định tính $X_{i1}, \ldots, X_{ir}$, và có $m - r$ thành phần định lượng $X_{i,r+1}, \ldots, X_{im}$. Kí hiệu lại

$$Y^{(i)} = (X_{i1}, \ldots, X_{ir}) \in E_1 \times \cdots \times E_r \subset \mathbb{R}^r,$$
$$Z^{(i)} = (X_{i,r+1}, \ldots, X_{im}) \in \mathbb{R}^{m-r} = \mathbb{R}^s,$$

trong đó $s = m - r$. Như vậy

$$X^{(i)} = (Y^{(i)}, Z^{(i)}).$$

Vì $Y^{(i)}$ là các dấu hiệu định tính nên tập E_i chỉ gồm một số hữu hạn giá trị

$$\begin{aligned} X_{i1} &\in E_1 = \{e_{11}, \ldots, e_{1m_1}\} \\ X_{i2} &\in E_2 = \{e_{21}, \ldots, e_{2m_2}\} \\ &\vdots \\ X_{ir} &\in E_r = \{e_{r1}, \ldots, e_{rm_r}\} \end{aligned}$$

Giả thiết $Z^{(i)}$ có phân bố chuẩn s chiều, $Z^{(i)} \sim N_s(\mu, \Sigma), \mu \in \mathbb{R}^s; \Sigma$ là ma trận xác định dương cấp $s \times s$. Ta kí hiệu nhóm A (nhóm khách hàng "tốt") gồm các phần tử có chỉ số $1, 2, \ldots, M$; B (nhóm khách hàng "không tốt") gồm các chỉ số $M+1, \ldots, N$. Giả thiết rằng

- $Z^{(i)} \sim N_s(\mu_A, \Sigma)$ nếu cá thể thứ $i \in A$,
- $Z^{(i)} \sim N_s(\mu_B, \Sigma)$ nếu cá thể thứ $i \in B$.

Đặt $\pi = \frac{M}{N}$ là tỉ lệ số các khách hàng thuộc nhóm A. Kí hiệu $Y = (X_1, \ldots, X_r)$ là biến ngẫu nhiên rời rạc bao gồm các dấu hiệu định tính của khách hàng và $Z = (X_{r+1}, \ldots, X_m)$ là các dấu hiệu định lượng của mỗi khách hàng.

Gọi $C(1|2)$ là tổn thất gây ra khi gán một phần tử thuộc nhóm B vào nhóm A, $C(2|1)$ là tổn thất gây ra khi gán một phần tử thuộc nhóm A vào nhóm B. Hai hằng số này được cho trước, chẳng hạn các chuyên gia ngân hàng cho rằng $C(1|2) = C(2|1)$.

Giả sử một khách hàng mới đến đăng kí vay tín dụng có dấu hiệu là $x = (y, z)$, với $y \in E_1 \times \cdots \times E_r, z \in \mathbb{R}^s$. Kí hiệu $P(Y = y|A)$ là xác suất để Y nhận giá trị y với điều kiện là khách hàng thuộc nhóm A và $f(z|Y = y, A), f(z|Y = y, B)$ là mật độ xác suất của thành phần z của véc-tơ dấu hiệu x với điều kiện $Y = y$ và khách hàng thuộc nhóm A, B tương ứng.

Ta giả thiết rằng $f(z|Y = y, A), f(z|Y = y, B)$ không phụ thuộc y, tức là

$$f(z|Y = y, A) = f(z|A), \quad f(z|Y = y, B) = f(z|B),$$

trong đó $f(z|A)$ là mật độ của phân bố chuẩn $N_s(\mu_A, \Sigma)$ và $f(z|B)$ là mật độ của phân bố chuẩn $N_s(\mu_B, \Sigma)$.

Quy tắc phân biệt khách hàng như sau : Gán cá thể có dấu hiệu $x = (y, z)$ vào nhóm A khi và chỉ khi

$$\frac{\pi P(Y = y|A)}{(1-\pi)P(Y = y|B)} \frac{f(Z|A)}{f(Z|B)} \geq \frac{C(1|2)}{C(2|1)}. \tag{10}$$

Vì $\pi P(Y = y|A), (1-\pi)P(Y = y|B), f(z|A), f(z|B)$ là các hàm chưa biết nên ta phải ước lượng chúng bằng cách sau đây.

Đặt $P(y) = P(\text{cá thể } \in A|Y = y)$. $1 - P(y) = P(\text{cá thể } \in B|Y = y)$. Theo công thức xác suất hậu nghiệm

$$P(y) = \frac{\pi P(Y = y|A)}{\pi P(Y = y|A) + (1-\pi)P(Y = y|B)} \tag{11}$$

Đối với các xác suất hậu nghiệm của biến ngẫu nhiên định tính, người ta hay dùng phân bố logistic :

$$P(y) \approx \frac{\exp(\beta_0 + \beta_1 y_1 + \cdots + \beta_r y_r)}{1 + \exp(\beta_0 + \beta_1 y_1 + \cdots + \beta_r y_r)}$$

hoặc

$$u := \ln \frac{P(y)}{1-P(y)} = \ln \frac{\pi P(Y=y|A)}{(1-\pi)P(Y=y|B)} \approx \beta_0 + \beta_1 y_1 + \cdots + \beta_r y_r, \tag{12}$$

tức là ta có quan hệ hồi quy tuyến tính

$$u = \beta_0 + \beta_1 y_1 + \cdots + \beta_r y_r. \tag{13}$$

Để có các số liệu thực nghiệm dùng để ước lượng các hệ số $\beta_i, i = 0, 1, \ldots, r$, ta tiến hành như sau:

Sử dụng hồi quy phi tuyến với biến phụ thuộc nhị nguyên để nhận được các ước lượng $\widehat{\beta}_i, i = 0, 1, \ldots, r$ và sau đó ước lượng $\widehat{P}(y)$ của phân bố hậu nghiệm $P(y)$ (xem (6)), và từ đó ta nhận được ước lượng

$$\widehat{u}(y) = \widehat{\beta}_0 + \widehat{\beta}_1 y_1 + \cdots + \widehat{\beta}_r y_r. \tag{14}$$

Đặt

$$L(z) = \ln \frac{f(z|A)}{f(z|B)} = (\mu_A - \mu_B)^T \Sigma^{-1} z - \frac{1}{2}(\mu_A - \mu_B)^T \Sigma^{-1}(\mu_A + \mu_B).$$

Đại lượng này được ước lượng bởi

$$\widehat{L}(z) = (\widehat{\mu}_A - \widehat{\mu}_B)^T S^{-1} z - \frac{1}{2}(\widehat{\mu}_A - \widehat{\mu}_B)^T S^{-1}(\widehat{\mu}_A + \widehat{\mu}_B), \tag{15}$$

trong đó

$$\widehat{\mu}_A = \frac{1}{M}\sum_{i=1}^{M} Z^{(i)}, \quad \widehat{\mu}_B = \frac{1}{N-M}\sum_{i=M+1}^{N} Z^{(i)}, \tag{16}$$

$$S_A = \frac{1}{M}\sum_{i=1}^{M} {Z^{(i)}}^T Z^{(i)} - \widehat{\mu}_A^T \widehat{\mu}_A,$$

$$S_B = \frac{1}{N-M}\sum_{i=M+1}^{N} {Z^{(i)}}^T Z^{(i)} - \widehat{\mu}_B^T \widehat{\mu}_B,$$

$$S = \frac{1}{N-2}[MS_A + (N-M)S_B]. \tag{17}$$

Như vậy, quy tắc phân biệt là : Gán phần tử có dấu hiệu $x = (y, z)$ vào nhóm A khi và chỉ khi

$$\widehat{u}(y) + \widehat{L}(z) > \ln \frac{C(1|2)}{C(2|1)}, \tag{18}$$

trong đó $\widehat{u}(y)$ cho bởi (14), $\widehat{L}(z)$ cho bởi (15).

4 Kết quả thực hiện

4.1 Kết quả phân lớp

Ngân hàng Techcombank lưu dữ liệu của 1728 khách hàng, mỗi khách hàng trong mẫu này có các dấu hiệu được cho trong Bảng 1. Các dấu hiệu này được mã hóa thành các

Ký hiệu	ý nghĩa
X_{01}	Tuổi tác
X_{02}	Trình độ học vấn
X_{03}	Loại hình công việc
X_{04}	Thời gian công tác
X_{05}	Mức thu nhập hàng tháng
X_{06}	Tình trạng hôn nhân
X_{07}	Nơi cư trú
X_{08}	Thời gian cư trú
X_{09}	Số người sống phụ thuộc
X_{10}	Phương tiện đi lại
X_{11}	Phương tiện thông tin
X_{12}	Chênh lệch thu nhập và chi tiêu
X_{13}	Giá trị tài sản khách hàng
X_{14}	Giá trị các khoản nợ
X_{15}	Quan hệ với Techcombank
X_{16}	Uy tín trong giao dịch

Bảng 1: Các đặc trưng của khách hàng

biến định lượng, việc mã hóa là do phía Ngân hàng đảm nhận. Công tác mã hóa (chia miền giá trị và tính điểm các giá trị cho các biến dấu hiệu) được dựa trên những kỹ thuật chuyên ngành và kinh nghiệm chuyên môn. Quy tắc mã hóa được cho trong Bảng 2 và 3.

Chúng tôi thực hiện bài toán phân lớp dựa trên các dấu hiệu quan trọng nhất, tức là chỉ dùng một số dấu hiệu có ảnh hưởng nhiều nhất đến độ tín nhiệm trong giao dịch của khách hàng để tham gia vào phân lớp. Việc đưa dấu hiệu nào vào phân lớp được chọn với sự tham khảo ý kiến chuyên gia, đó là: trình độ học vấn, loại hình công việc, mức thu nhập hàng tháng, phương tiện đi lại, chệnh lệch thu nhập và chi tiêu, giá trị tài sản khách hàng, giá trị các khoản nợ. Với lý thuyết và thuật toán được trình bày trong Mục 2, chúng tôi thực hiện tính toán bằng phần mềm thống kê SPLUS và được kết quả sau: $N = 1728$ khách hàng được chia thành 2 nhóm: nhóm A có $m = 1374$ khách hàng, nhóm B có $n = 354$ khách hàng. Theo quy tắc mã hóa, khách hàng có điểm càng cao càng thể hiện mức tín nhiệm cao, do đó chúng tôi dễ nhận thấy nhóm A là "nhóm tốt", nhóm B là "nhóm xấu". Khoảng cách Holtelling tính được là

$$T^2_{A,B} = 27,73104$$

trong khi đó $\chi^2_{16}(0.05) = 26,296$. Như vậy $T^2_{A,B} > \chi^2_{16}(0.05)$ nên hai nhóm A, B là khác nhau một cách có ý nghĩa.

Chú ý. Khoảng cách Holtelling trên được tính với sự tham gia của tất cả các dấu hiệu.

4.2 Các hệ số hồi quy

Bảng 4 là kết quả thực hiện hồi quy nhị nguyên logistic trên tập mẫu.

Dấu hiệu		Thang điểm
Tuổi tác	20-25	2
	26-35	3
	36-55	4
	56-60	3
	<20 hoặc >60	1
Trình độ học vấn	Trên đại học	4
	Dại học	3
	Cao đẳng hoặc tương đương	2
	Tú tài hoặc tương đương	1
	Dưới tú tài	0
Loại hình công việc	Không có việc	0
	Dã nghỉ hưu hưởng lương	2
	Lao động phổ thông	2
	Lao động được đào tạo nghề	3
	Diều hành SXKD nhỏ	4
	Cán bộ chuyên viên	4
	Quản lý điều hành	5
	Không thuộc các đối tượng trên	1
Thời gian công tác	Dưới 1 năm	1
	Trên 1 năm	2
Mức thu nhập hàng tháng (triệu đồng)	>5	10
	>4 và <=5	8
	>3 và <=4	6
	>2 và <=3	4
	>1 và <=2	2
	<=1	1
Tình trạng hôn nhân	Dộc thân	2
	Có gia đình	3
	Dã li dị, goá	1
Nơi cư trú	Thuộc sở hữu	3
	ở nhờ cơ quan	2
	Di thuê	1
Thời gian cư trú	Dưới 6 tháng	1
	Trên 6 tháng	2
Số người sống phụ thuộc	0	4
	1	3
	2	2
	3	1
	>=4	0

Bảng 2: Quy tắc mã hóa các dấu hiệu

Dấu hiệu		Thang điểm
Phương tiện đi lại	ôtô con	4
	Công cộng	2
	Xe máy	2
	Khác	1
Phương tiện thông tin	Không dùng điện thoại	0
	Có dùng điện thoại	1
Chênh lệch thu chi (triệu đồng)	<=1	1
	>1 và <=2	2
	>2 và <=3	4
	>3 và <=4	6
	>4 và <=5	8
	>5	10
Giá trị tài sản (triệu đồng)	<=500	1
	>500 và <=1000	2
	>1000 và <=2000	4
	>2000 và <= 3000	6
	>3000	8
Giá trị các khoản nợ (triệu đồng)	>300	0
	>200 và <=300	1
	>100 và <=200	2
	>0 và <=100	3
	0	0
Quan hệ với Techcombank	Chưa	0
	Có	1
Uy tín giao dịch	Dã phát sinh nợ quá hạn	0
	Dã được gia hạn nợ	1
	Trả nợ gốc, lãi đúng hạn	2

Bảng 3: Quy tắc mã hóa các dấu hiệu (tiếp)

4.3 Nhận xét

Ta có một số nhận xét về xác suất $\widehat{P}(x)$.

1. Theo bảng 4, ta có

$$\widehat{P}(x) = \frac{e^{\widehat{\beta}^T x}}{1 + e^{\widehat{\beta}^T x}}$$

trong đó

$$\begin{aligned}\widehat{\beta}^T x = &-1.2382x_1 - 0.5911x_2 - 1.3720x_3 + 3.2401x_5 \\ &- 1.8337x_6 - 8.0706x_7 - 5.3368x_8 - 1.0917x_9 - 1.5085x_{10} \\ &- 18.2826x_{11} + 5.6702x_{12} + 3.595x_{13} - 0.9303x_{14} - 1.4824x_{15}\end{aligned}$$

2. Từ bảng 5, nếu với quy tắc phân biệt khách hàng là "Gán khách hàng có dấu hiệu x vào nhóm A khi và chỉ khi $\widehat{P}(x) > 0.5$" thì tỉ lệ khách hàng được phân biệt đúng trong mẫu 1728 khách hàng là 99.02%, đó là tỉ lệ rất cao.

3. Trong mô hình hồi quy với biến phụ thuộc nhị phân, ta đã loại 2 biến X_4 (thời gian công tác) và X_{16} (uy tín trong giao dịch) ra khỏi mô hình vì hai lí do sau:
   - X_4, X_{16} có sự phụ thuộc tuyến tính với các biến khác
   - Các ước lượng $\widehat{\beta}_4, \widehat{\beta}_{16}$ trong mô hình 16 biến tỏ ra không ổn định.

4. Do tập mẫu gồm 1728 khách hàng đã được phục vụ bởi Techcombank chưa đủ lớn và đã được chọn lựa nên hai nhóm A, B phân biệt khá rõ. Nếu ta mở rộng tập mẫu thì có thể kết quả không còn được hiệu quả như trước.

Variable	Coefficient	Std. Error	z-Statistic	Prob.
X1	-1.2382	0.54726	-2.2625	0.0237
X2	-0.5911	0.45976	-1.2857	0.1986
X3	-1.3720	0.81657	-1.6802	0.0929
X5	3.24010	0.82967	3.90532	0.0001
X6	-1.8337	0.76720	-2.3901	0.0168
X7	-8.0706	2.20437	-3.6612	0.0003
X8	-5.3368	1.51770	-3.5164	0.0004
X9	-1.0917	0.47816	-2.2831	0.0224
X10	-1.5085	0.63631	-2.3706	0.0178
X11	-18.2826	4.6	-3.9745	0.0001
X12	5.67018	1.22706	4.62094	0
X13	3.595	0.8323	4.31934	0
X14	-0.9303	0.42953	-2.1659	0.0303
X15	-1.4824	0.79869	-1.856	0.0634
Mean dependent var	0.79514	S.D. depend. var		0.40372
S.E. of regression	0.06988	Akaike inf. cri.		0.04576
Sum squared resid	8.36966	Schwarz cri.		0.08996
Log likelihood	-25.54	Han-Qui cri.		0.06211
Avg. log likelihood	-0.0148			
Obs with Dep=0	354	Total obs		1728
Obs with Dep=1	1374			

Bảng 4: Các hệ số hồi quy với biến phụ thuộc nhị nguyên

	Estimated Equation			Constant Probability		
	Dep=0	Dep=1	Total	Dep=0	Dep=1	Total
$P(\text{Dep} = 1) \leq C$	347	6	353	0	0	0
$P(\text{Dep} = 1) > C$	7	1368	1375	354	1374	1728
Total	354	1374	1728	354	1374	1728
Correct	347	1368	1715	0	1374	1374
% Correct	98.02	99.56	99.25	0.00	100.00	79.51
% Incorrect	1.98	0.44	0.75	100.00	0.00	20.49
Total Gain*	98.02	-0.44	19.73			
Percent Gain**	98.02	NA	96.66			
	Estimated Equation			Constant Probability		
	Dep=0	Dep=1	Total	Dep=0	Dep=1	Total
$E(\#\text{ of Dep} = 0)$	345.45	7.57	353.01	72.52	281.48	354
$E(\#\text{ of Dep=1})$	8.55	1366.43	1374.99	281.48	1092.52	1374
Total	354	1374	1728	354	1374	1728
Correct	345.45	1366.43	1711.88	72.52	1092.52	1165.04
% Correct	97.58	99.45	99.07	20.49	79.51	67.42
% Incorrect	2.42	0.55	0.93	79.51	20.49	32.58
Total Gain*	77.1	19.94	31.65			
Percent Gain**	96.96	97.31	97.14			

*Change in "% Correct" from default (constant probability) specification
**Percent of incorrect (default) prediction corrected by equation

Bảng 5: Kết quả phân biệt (success cutoff $C = 0.5$)

Lời cảm ơn. Các tác giả xin chân thành cảm ơn TS. Hồ Dăng Phúc cho những ý kiến đóng góp quý báu.

Tài liệu

[1] EMISCOM R&D, *Báo cáo Giai đoạn I: Nghiên cứu khảo sát lý thuyết và thực tiễn đánh giá tín dụng thể nhân*

[2] Nguyễn Văn Hữu, Nguyễn Hữu Dư, *Phân tích thống kê và dự báo*, NXB Đại học Quốc gia HN, 2003.

[3] A. Aggarawal, *Categorical data analysis*, Wiley, New York, 1990. 1.2.1

[4] H. T. Albright, *Construction of a polynomial classifier for consumer loan applications using genetic algorithms*, Department of Systems Engineering, University of Virginia, 1994. 1.2.3

[5] F. Black and M. Scholes, *The pricing of options and corporate liabilities*, Journal of Pocitical Economy, 81:637-654, 1973. 1

[6] M. Boyle, J. N. Crook, R. Hamilton and L. C. Thomas, *Credit scoring and credit control, chapter Methods for credit scroing applied to slow payers*, pages 75-90, Oxford University Press, Oxford, 1992. 1, 1.2.2

[7] L. Breiman, J.H. Friedman, R.A Olshen, and C.J. Stone, *Classification and regression trees*, Wadsworth, Belmont, CA, 1984. 1.2.2

[8] N. Capon, *Credit scoring systems: a critical analysis*, Journal of Marketing, 46:82-91, 1982. 1.2.1

[9] C. Carter and J. Catlett, *Assessing credit card applications using ma- chine learning*, IEEE Expert, 2:71-79, 1987. 1.2.2

[10] R. A. Jonhson, D. W. Wichern, *Applied Multivariate Statistical Analysis*, 1998.

[11] *Credit Scoring and Credit Control*, Edited by L.C. Thomas, J.N. Crook, D.B. Edelman, 1992.

Statistical Method in Development of Credit Scoring System

Abstract. In this paper, we consider the problem of credit scoring for personal customer. The main statistical tools are used to establish the credit scoring system are theory of classification and discrimination. Our method is illustrated on the credit customer dataset of a Trade Bank.

Dịa chỉ tác giả:

Vương Quân Hoàng
Centre Emile Bernheim, ULB, 21 F.D.Roosevelt, B-1050, Bruxelles
Email: qvuong@ulb.ac.be
Đào Gia Hưng
Ngân hàng Techcombank
Nguyễn Văn Hữu, Trần Minh Ngọc và Lê Hồng Phương
Khoa Toán-cơ-tin học, trường Dại học KHTN, DH QGHN
Email:huunv@vnu.edu.vn, ngoctm@vnu.edu.vn, phuonglh@vnu.edu.vn