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/ctu.jvn.2020.147

PHÂN LOẠI BẰNG PHƯƠNG PHÁP BAYES VÀ ỨNG DỤNG TRONG Y HỌC

Võ Văn Tài[1*], Trần Trung Tín[1], Thái Minh Trọng[1], Châu Ngọc Thơ[1] và Lê Thị Kim Ngọc[2]

[1]*Khoa Khoa học Tự nhiên, Trường Đại học Cần Thơ*

[2]*Khoa Kỹ thuật, Trường Đại học Văn Lang*

**Người chịu trách nhiệm về bài viết: Võ Văn Tài (email: vvtai@ctu.edu.vn)*

Thông tin chung:
Ngày nhận bài: 03/08/2020
Ngày nhận bài sửa: 10/09/2020
Ngày duyệt đăng: 28/12/2020

Title:
Classifying by Bayesian method and applying in medicine

Từ khóa:
Bài toán phân loại, hồi qui logistic, phương pháp Bayes, sai số Bayes

Keywords:
Bayesian method, Bayes error, classification problem, logistic regression

ABSTRACT

This paper is to study the classification problem by Bayesian method in which estimating probability density function, and finding prior probability from real data are considered. The research also solves some complex calculations of this method by the built approximation and Matlab procedure. From the above improvements, an algorithm based on Bayesian method to classify a disease is proposed. This algorithm is applied specifically for a chronic kidney disease at the Can Tho Central General Hospital with real data. The outcome shows that the proposed algorithm has given good result in classifying this disease. Furthermore, this result also illustrates the advantages of the proposed method in comparison with the existing methods which are regularly used recently times.

TÓM TẮT

Bài viết này nghiên cứu bài toán phân loại bằng phương pháp Bayes, trong đó việc ước lượng hàm mật độ xác suất và tìm xác suất tiên nghiệm từ số liệu thực tế được xem xét. Nghiên cứu cũng giải quyết được những tính toán phức tạp của phương pháp này bởi sự xấp xỉ và chương trình Matlab được xây dựng. Từ những cải tiến trên, thuật toán phân loại bệnh bằng phương pháp Bayes được đề xuất. Thuật toán này được áp dụng cụ thể cho một tập dữ liệu thực tế bệnh suy thận mạn tại bệnh viện đa khoa Trung ương Thành phố Cần Thơ. Kết quả cho thấy thuật toán đề nghị đã cho kết quả tốt trong phân loại bệnh này. Kết quả này cũng chứng minh ưu điểm của thuật toán đề xuất so với các thuật toán được áp dụng phổ biến gần đây.

Trích dẫn: Võ Văn Tài, Trần Trung Tín, Thái Minh Trọng, Châu Ngọc Thơ và Lê Thị Kim Ngọc, 2020. Phân loại bằng phương pháp Bayes và ứng dụng trong y học. Tạp chí Khoa học Trường Đại học Cần Thơ. 56(6A): 97-103.

1 GIỚI THIỆU

Phân loại là việc gán một phần tử thích hợp nhất vào các tổng thể đã được biết trước dựa vào các biến quan sát. Nó là một hướng phát triển quan trọng của thống kê nhiều chiều, có vai trò nền tảng trong lĩnh vực khai phá dữ liệu. Bài toán phân loại đã và đang được áp dụng đa dạng trong các lĩnh vực nên hiện tại được rất nhiều nhà thống kê và công nghệ thông tin quan tâm (Cristianini and Shawe, 2000; Pham-Gia *et al*., 2008; Tai et al., 2018). Về mặt lý thuyết,

hiện có bốn phương pháp chính để giải quyết bài toán phân loại: phương pháp Fisher, hồi qui logistic, SVM (super vector machine) và Bayes (Tai, 2017). Phương pháp Fisher ra đời sớm nhất, có thể phân loại cho hai hay nhiều hơn hai tổng thể nhưng phải giả thiết ma trận hiệp phương sai của các tổng thể bằng nhau nên có nhiều hạn chế trong áp dụng thực tế (Tai, 2017). Hiện nay, phương pháp hồi qui logistic được sử dụng phổ biến, nhưng chỉ hiệu quả khi dữ liệu có sự tách rời tốt của các nhóm và biến phụ thuộc là nhị phân (Jan et al., 2010). Phương pháp SVM tận dụng sự phát triển của công nghệ thông tin, xây dựng mô hình dựa trên dữ liệu tập huấn luyện và tập kiểm tra nên đòi hỏi dữ liệu lớn (Cristianini and Shawe, 2000). Phương pháp Bayes được xem có nhiều ưu điểm, có thể phân loại được cho hai hay nhiều hơn hai tổng thể. Phương pháp này cũng không bị ràng buộc bởi các giả thiết phân phối chuẩn và phương sai bằng nhau của các tổng thể. Các kết quả nghiên cứu mới trong những năm gần đây về bài toán phân loại chủ yếu tập trung xung quanh phương pháp Bayes (Tai et al., 2018).

Trong áp dụng thực tế hiện nay, phương pháp Bayes được sử dụng khá hạn chế. Trong hạn chế này, vấn đề xác định xác suất tiên nghiệm, ước lượng hàm mật độ xác suất và sự tính toán phức tạp của phương pháp này là những nguyên nhân chính. Xác suất tiên nghiệm thường được xác định dựa vào kinh nghiệm, sự hiểu biết của người thực hiện, hoặc một tổng kết thống kê trước đó cho vấn đề mà ta cần phân loại. Một số đề xuất dựa vào thống kê cũng được xem xét và áp dụng như phân phối đều, tỉ lệ mẫu, phương pháp Laplace. Tuy nhiên chúng chỉ phù hợp cho từng bộ dữ liệu mà không phải là tất cả (Tai et al., 2018). Bên cạnh xác suất tiên nghiệm, khi thực hiện bài toán phân loại bằng phương pháp Bayes, chúng ta phải ước lượng hàm mật độ xác suất. Mặc dù có nhiều cải tiến trong những năm gần đây, nhưng cho đến nay nó vẫn còn nhiều hạn chế (Thao and Tai, 2017). Ngoài hai vấn đề trên, những phức tạp trong tính toán như tìm hàm cực đại, tính tích phân trong không gian nhiều chiều cũng là cản trở chính trong áp dụng thực tế của phương pháp này (Tai and Pham-Gia, 2010). Dựa trên bài toán phân tích chùm mờ (Pal and Bezdek, 1995; Thao and Tai, 2017) nghiên cứu đề xuất thuật toán xác định xác suất tiên nghiệm phù hợp cho từng bộ dữ liệu và cho từng phần tử cần phân loại. Nghiên cứu này cũng đề nghị phương pháp ước lượng hàm mật độ xác suất từ số liệu rời rạc và áp dụng phương pháp Monte Carlo để giải quyết vấn đề tính toán trong thực tế của phương pháp Bayes.

Hiện nay bài toán phân loại được ứng dụng rộng rãi trong nhiều lĩnh vực khác nhau. Trong y học, bài toán phân loại được ứng dụng theo hai hướng sau:

i) Có k loại bệnh đều được phát hiện dựa vào n biến quan sát định tính hoặc định lượng. Một người có các chỉ số sinh hóa cụ thể, dựa vào các phương pháp phân loại, chúng ta cần kết luận người đó bị bệnh nào trong số k loại bệnh đã biết.

ii) Chúng ta đang quan tâm một loại bệnh cụ thể *B* nào đó của một người. Dựa trên n biến quan sát định tính hoặc định lượng của người này, cần kết luận người này có khả năng bị bệnh *B* hay không.

Cả hai vấn đề này thực chất là việc giải quyết bài toán phân loại cho hai tổng thể và nhiều hơn hai tổng thể. Vì vậy nghiên cứu này sẽ áp dụng những cải tiến trong thực tế của phương pháp Bayes được đề cập ở trên vào lĩnh vực y học.

Phần tiếp theo của bài báo được cấu trúc như sau. Phần 2 trình bày phương pháp Bayes và thuật toán đề nghị trong phân loại bệnh. Phần 3 giải quyết các vấn đề áp dụng thực tế của thuật toán đề nghị. Phần 4 áp dụng thuật toán cho một số liệu thực tế. Phần cuối cùng là kết luận của bài viết.

2 PHƯƠNG PHÁP BAYES VÀ THUẬT TOÁN PHÂN LOẠI BỆNH

2.1 Phương pháp Bayes

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

Nếu $g_{\max}(x_0) = q_j f_j(x_0)$ thì xếp x_0 vào w_j, (1)

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

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

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

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

Từ công thức (2), ta có

$$Pe_{1,2,...,k}^{(q)} = \sum_{j=1}^{k} \int_{R^n \setminus R_j^n} q_j f_j(x)dx$$

$$= \sum_{j=1}^{k} \left[\int_{R^n} q_j f_j(x)dx - \int_{R_j^n} \max_{1\le l\le c} \left\{ q_l f_l(x) \right\} dx \right]$$

$$= \int_{R^n} \sum_{j=1}^{k} q_j f_j(x)dx - \sum_{j=1}^{k} \int_{R_j^n} \max_{1\le l\le k} \left\{ q_l f_l(x) \right\} dx$$

$$= 1 - \int_{R^n} \max_{1\le l\le k} \left\{ q_l f_l(x) \right\} dx. \qquad (3)$$

Sử dụng (3) 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.

2.2 Thuật toán phân loại bệnh

2.2.1 Bài toán

Giả sử ta đang quan tâm một loại bệnh *B*. Bệnh này được chia thành *k* loại w_1, w_2, ..., w_k với những phương pháp hoặc phác đồ điều trị khác nhau. Trên mỗi loại bệnh w_i, chúng ta có N_i phần tử ($n_1 + n_2 + \ldots + n_k = N$). Việc phân thành các loại bệnh dựa trên biến *n* chiều *x* và tập dữ liệu *Z* của các tổng thể. Khi đó, với một người có $x = x_0$, chúng ta cần kết luận người này có bệnh *B* hay không hoặc xếp vào loại bệnh nào, trong số *k* loại đã biết.

2.2.2 Thuật toán

Thuật toán đề nghị bao gồm các bước sau:

Bước 1. Kiểm tra hiện tượng đa cộng tuyến của các biến trên tập dữ liệu và thực hiện việc khắc phục nếu xảy ra hiện tượng này.

Bước 2. Xác định các biến có ý nghĩa thống kê ảnh hưởng đến bệnh *B* qua mô hình hồi qui logistic. Trong áp dụng của nghiên cứu này, mức ý nghĩa 5% được sử dụng để thực hiện.

Bước 3. Ước lượng hàm mật độ xác suất cho mỗi w_i.

Bước 4. Tìm xác suất tiên nghiệm cho các tổng thể.

Bước 5. Tiến hành phân loại theo (1). Đánh giá hiệu quả của phương pháp thực hiện qua xác suất phân loại đúng.

3 MỘT SỐ VẤN ĐỀ LIÊN QUAN ĐẾN ÁP DỤNG

3.1 Thuật toán tìm xác suất tiên nghiệm

3.1.1 Giới thiệu

Kết quả phân loại một phần tử mới bởi nguyên tắc (1) và sai số Bayes được tính bởi (2) đều phụ thuộc vào xác suất tiên nghiệm. Mặc dù có nhiều tác giả đã nghiên cứu về vấn đề này như Inman and Bradley (1989), Miller *et al.* (2011), Tai et al. (2018), 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. Thông thường có những phương pháp sau để xác định các xác suất tiên nghiệm:

(i) Dựa vào phân phối đều: $q_1 = q_2 = ... = q_c = \frac{1}{k}$.

(ii) Dựa vào tập mẫu: $q_i = \frac{n_i}{N}$.

(iii) Dựa vào ước lượng Laplace: $q_i = \frac{n_i + n/k}{N + n}$.

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

Dựa vào bài toán phân tích chùm mờ, nghiên cứu này đề xuất thuật toán tìm xác suất tiên nghiệm.

3.1.2 Khái niệm

Trong không gian *n* chiều, cho *N* phần tử với tập dữ liệu

$$Z = \left[z_{ij} \right]_{n\times N} = \begin{bmatrix} z_{11} & z_{12} & \cdots & z_{1N} \\ z_{21} & z_{22} & \cdots & z_{2N} \\ \cdots & \cdots & \cdots & \cdots \\ z_{n1} & z_{n2} & \cdots & z_{nN} \end{bmatrix}.$$

Tập dữ liệu này được chia thành *k* tổng thể, khi đó

$$U = \left[\mu_{ij} \right]_{k\times N} = \begin{bmatrix} \mu_{11} & \mu_{12} & \cdots & \mu_{1N} \\ \mu_{21} & \mu_{22} & \cdots & \mu_{2N} \\ \cdots & \cdots & \cdots & \cdots \\ \mu_{k1} & \mu_{k2} & \cdots & \mu_{kN} \end{bmatrix}$$

được gọi là ma trận phân vùng mờ với μ_{ij} là xác suất khi xếp phần tử thứ *j* vào chùm thứ *i* (w_i), $\mu_{ij} \in [0,1]$, $\forall 1 \le i \le k$, $1 \le j \le N$. Trong phân tích chùm không mờ, ta có $\mu_{ij} = 1$ khi phần tử thứ *j* thuộc chùm thứ *i* và $\mu_{ij} = 0$ khi phần tử thứ *j* không thuộc chùm thứ *i*.

Tập tất cả những ma trận phân vùng mờ cho dữ liệu $\left[z_{ij} \right]_{n\times N}$, $N \geq 2$ được gọi là không gian phân vùng mờ của *k* chùm:

$$M_{zk} = \left\{ \begin{array}{l} U = [\mu_{ik}]_{k\times N} \mid \mu_{ij} \in [0,1], \forall i, j, \\ \sum\limits_{i=1}^{k} \mu_{ij} = 1, \forall j, \ 0 < \sum\limits_{k=1}^{N} \mu_{ij}, \forall i \end{array} \right\}.$$

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

$$v_i = \frac{\sum\limits_{j=1}^{N} \left(\mu_{ij}\right)^m z_j}{\sum\limits_{k=1}^{N} \left(\mu_{ij}\right)^m}, \ 1 \leq i \leq k, \qquad (4)$$

$m \in [1, +\infty)$ là tham số xác định độ mờ của kết quả phân tích chùm.

Bình phương khoảng cách từ z_j đến v_i được cho bởi

$$D_{ij}^2(z_j, v_i) = \left\| z_j - v_i \right\|^2 = \left(z_j - v_i\right)^T . \left(z_j - v_i\right), \qquad (5)$$

Trong bài viết này, khoảng cách Euclide được áp dụng trong các ứng dụng.

3.1.3 Thuật toán

Bước 1: Tìm phần tử đại diện của mỗi chùm $v_i, i = 1, 2, ..., k$ bởi công thức (4).

Tính các khoảng cách D_{ij} giữa từng phần tử trong chùm w_j tới phần tử đại diện của chùm (v_i) theo công thức (5).

Bước 2: Thiết lập ma trận phân vùng ban đầu như sau:

$$U^{(0)} = \left[\mu_{ij} \right]_{k\times(N+1)} = \begin{bmatrix} \mu_{11} & \mu_{12} & \cdots & \mu_{1N} & \frac{1}{k} \\ \mu_{21} & \mu_{22} & \cdots & \mu_{2N} & \frac{1}{k} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \mu_{k1} & \mu_{k2} & \cdots & \mu_{kN} & \frac{1}{k} \end{bmatrix},$$

trong đó *N* cột đầu tiên là ma trận phân vùng không mờ của các phần tử trong tập dữ liệu khi xếp vào *k* tổng thể $w_1, w_2, ..., w_k$. (Cụ thể $\mu_{ij} = 1$ khi phần tử thứ *j* thuộc chùm thứ *i* và $\mu_{ij} = 0$ nếu phần tử thứ *j* không thuộc chùm thứ *i*). Cột cuối cùng (*N+1*) là xác suất ban đầu để x_0 xếp vào các chùm $w_1, w_2, ..., w_k$. Có thể chọn xác suất này bằng nhau và bằng 1/*k*.

Bước 3: Cập nhật ma trận phân vùng mới $U^{(1)}$ theo nguyên tắc sau:

Nếu $D_{ij} > 0$ cho tất cả *i* = 1, 2,…, *k*, ta tính $\mu_{ij}^{(1)}$ theo công thức

$$\mu_{ij}^{(1)} = \frac{1}{\sum\limits_{h=1}^{k} \left(D_{ij} D_{hj}\right)^{2/(m-1)}}, \text{ với } 1 \leq j \leq N.$$

Nếu tồn tại *i* sao cho $D_{ij} = 0$, ta gán $\mu_{ij}^{(1)} = 0$, những vị trí khác còn lại được gán ngẫu nhiên thỏa $\mu_{ij}^{(1)} \in [0,1], \sum\limits_{i=1}^{k} \mu_{ij}^{(1)} = 1.$

Bước 4: Tính $\left\| U^{(1)} - U^{(0)} \right\| = \max_{ij} \left(\left| \mu_{ij}^{(1)} - \mu_{ij}^{(0)} \right| \right).$

Lặp lại các bước trên *t* lần cho đến khi $\left\| U^{(t)} - U^{(t-1)} \right\| < \varepsilon.$

Khi đó chúng ta sẽ có ma trận phân vùng có cột cuối cùng (cột thứ *N+1*) là xác suất tiên nghiệm khi xếp x_0 vào các tổng thể tương ứng.

Trong thuật toán trên ta có

i) ε là một hằng số nhỏ tùy ý. Khi ε càng nhỏ thì vòng lặp thực hiện sẽ càng nhiều. Nghiên cứu này chọn ε = 1%.

ii) Tham số mờ *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, các xác suất của các phần tử thuộc vào các chùm bằng nhau và bằng 1/*k*. Tuy tham số *m* ảnh hưởng trực tiếp đến kết quả phân tích chùm nhưng làm thế nào để xác định tham số mờ một cách hợp lý là một vấn đề khó. Mặc dù có rất nhiều tác giả nghiên cứu về vấn đề này, nhưng việc xác định *m* một cách cụ thể vẫn

thường được thực hiện bằng phương pháp chia lưới. Theo đó, các ứng dụng thường chọn *m* từ 1 đến 5. Nghiên cứu này chọn *m* = 2 trong tất cả các ví dụ số.

3.2 Một số vấn đề khác

i) Trong thực tế dữ liệu có nhu cầu để thực hiện bài toán phân loại là dữ liệu rời rạc, do đó để bài toán phân loại bằng phương pháp Bayes có tính ứng dụng thực tế, việc đầu tiên phải làm là ước lượng hàm mật độ xác suất từ dữ liệu này. Có nhiều phương pháp tham số cũng như phi tham số để thực hiện việc này. Trong bài viết này, chúng tôi sử dụng phương pháp hàm hạt nhân, một phương pháp cho đến hiện tại có nhiều ưu điểm (Tai and Pham-Gia, 2010; Ghosh *et al.*, 2012; Tai et al., 2018). 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*, và N là số phần tử của mẫu.

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

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

Có nhiều nghiên cứu về việc chọn tham số trơn, nhưng kết luận cuối cùng là không có cách chọn tham số nào thực sự có ưu thế so với các cách khác. Trong nghiên cứu này, tham số trơn theo Thao and Tai (2017) được chọn:

$$h_j = \left(\frac{4}{N\left(n+2\right)} \right)^{\frac{1}{n+4}} \sigma_j,$$

trong đó σ_j là độ lệch chuẩn mẫu của biến thứ *j*.

ii) Để tính sai số Bayes, Tai (2017) đã tìm các biểu thức giải tích cụ thể để xác định trong một số trường hợp đặc biệt của phân phối một chiều cho hai tổng thể. Trong trường hợp nhiều tổng thể một chiều, chúng tôi đã thiết lập chương trình xác định biểu thức giải tích cụ thể hàm cực đại, để từ đó tính tích phân chúng và xác định chính xác sai số Bayes. Khi có nhiều chiều, việc xác định hàm cực đại của các $g_i(x)$ vô cùng phức tạp, ngay cả trường hợp hai tổng thể có phân phối chuẩn. Vì vậy nghiên cứu này, cách tính gần đúng hàm cực đại của các hàm mật độ xác suất bằng phương pháp Monte-Carlo (Jasra *et al.*, 2005) được sử dụng, để từ đó tính sai số Bayes cho trường hợp *k* tổng thể *n* chiều. Một chương trình tính sai số Bayes trên phần mềm Matlab cũng được xây dựng ở đây.

Tất cả các chương trình được đề cập trong phần này được sử dụng để giải quyết hiệu quả bài toán thực tế của phần 4.

4 ÁP DỤNG

4.1 Dữ liệu

Bệnh suy thận mạn là quá trình suy giảm về chức năng, cấu trúc thận, ảnh hưởng rất lớn đến sức khỏe của người bệnh. Bệnh này có nhiều biến chứng nguy hiểm dẫn đến tử vong. Ở Việt Nam, đây là một trong những bệnh phổ biến. Nghiên cứu này sử dụng số liệu thực thực tế được lấy tại Bệnh viện đa khoa Trung ương thành phố Cần Thơ trong năm 2017. Bộ số liệu gồm 259 người bị bệnh thận mạn giai đoạn cuối, trong đó có 237 người được chữa khỏi bệnh và 22 tử vong. Số liệu có 183 biến, trong đó biến phụ thuộc là *Y* (*Y* = 1: Tử vong; *Y* = 0: Khỏi bệnh). Vì số lượng các biến xem xét lớn và mục đích của nghiên cứu chỉ xem xét khía cạnh toán học nên bài viết không trình bày chi tiết các biến. Mục đích của nghiên cứu này là xác định các biến có ý nghĩa thống kê ảnh hưởng đến việc tử vong hoặc khỏi bệnh, từ đó tìm mô hình phân loại tối ưu cho 2 nhóm này. Để thực hiện việc nghiên cứu, chúng tôi chia tập số liệu thành 2 nhóm: tập huấn luyện và tập kiểm tra với tỷ lệ lần lượt là 80% và 20%. Chi tiết vấn đề này được cho bởi Bảng 1.

Bảng 1: Chi tiết các tập số liệu

Số liệu	Toàn bộ	Tập huấn luyện	Tập kiểm tra
Tổng số	259	207	52
Tử vong	22	17	5
Khỏi bệnh	237	190	47

4.2 Kết quả thực hiện

4.2.1 Tập huấn luyện

– Tính hệ số tương quan cặp giữa các biến định lượng, ta có 25 cặp biến có hệ số tương quan từ 0,85 đến 0,95. Điều này cho thấy nếu cùng đưa các cặp biến này vào mô hình thì sẽ xảy ra hiện tượng đa cộng tuyến. Tham khảo thêm ý kiến của bác sĩ, chúng tôi loại bỏ 25 biến. Như vậy các biến còn lại để xem xét là 157.

– Tiến hành phân tích hồi qui logistic với 157 biến, ta nhận được chỉ 4 biến độc lập có ý nghĩa thống kê 5% ảnh hưởng đến biến phụ thuộc *Y*. Chi tiết việc thực hiện trên 4 biến này được cho bởi Bảng 2.

Bảng 2: Trích dẫn phân tích thống kê 4 biến có ý nghĩa 5% ảnh hưởng đến biến *Y*

Biến	Hệ số	SE	Wald	Df	Sig.
Hc	-1,869	0,762	6,017	1	0,014
Cn	-0,003	0,001	8,851	1	0,003
Na	-0,204	0,067	9,218	1	0,002
Ca	2,760	0,851	10,511	1	0,001
Constant	23,188	8,594	7,281	1	0,007

Bảng 2 cho ta thấy có 4 biến có ý nghĩa thống kê lần lượt là HC, Cn, Na, Ca. Do đó nghiên cứu sử dụng 4 biến này để tìm mô hình tối ưu cho bài toán phân loại.

– Thực hiện việc ước lượng hàm mật độ xác suất, tìm xác suất tiên nghiệm trong những trường hợp khác nhau, tiến hành phân loại cho từng trường hợp 1 biến, 2 biến, 3 biến và 4 biến. Trong mỗi trường hợp, phương pháp Bayes với xác suất tiên nghiệm đều (BayesU), dựa vào tập mẫu (BayesT), phương pháp Laplace (BayesL) và thuật toán đề nghị (BayesC) lần lượt được xem xét. Kết quả được tổng kết bởi bảng sau:

Bảng 3: Xác xuất phân loại đúng (%) các trường hợp của phương pháp Bayes

Trường hợp	Biến	Bayes U	Bayes T	Bayes L	Bayes C
1 Biến	Hc	56,60	88,99	68,26	91,23
	Cn	69,50	91,22	70,39	91,80
	Na	63,24	91,78	70,71	92,12
	Ca	63,46	89,24	74,70	92,22
2 biến	Hc, Cn	71,50	84,05	72,56	92,90
	Hc, Na	65,89	84,78	72,49	92,78
	Hc, Ca	66,39	83,46	75,92	93,00
	Cn, Na	72,99	87,88	75,61	93,23
	Cn, Ca	74,00	86,34	77,14	93,29
	Na, Ca	70,98	83,36	77,40	93,17
3 Biến	Hc, Cn, Na	74,35	82,99	76,55	93,30
	Hc, Cn, Ca	87,54	88,74	88,14	94,05
	Hc, Na, Ca	90,09	90,60	90,08	93,21
	Cn, Na, Ca	91,09	93,06	91,29	94,12
4 Biến	Hc, Cn, Na, Ca	78,64	83,39	81,76	95,57

Bảng 3 cho thấy BayesC cho kết quả phân loại đúng rất ổn định và cao nhất (91% - 96%).

Thực hiện phương pháp phân loại logistic, Fisher, SVM với từng trường hợp của biến, lựa chọn trường hợp có xác suất phân loại đúng cao nhất của mỗi phương pháp, sau đó so sánh với trường hợp tốt nhất của phương pháp BayesC (Bảng 3), ta có Bảng 4.

Bảng 4: Xác suất phân loại tối ưu của 4 pháp

Phương pháp	1 biến	2 biến	3 biến	4 Biến
Hồi qui Logistic	91,50	91,93	92,71	92,32
Fisher	69,90	70,64	74,54	72,53
SVM	64,23	79,34	86,13	86,43
Bayes	92,22	92,90	94,12	95,57

Bảng 4 cho ta thấy xác suất phân loại đúng của các phương pháp theo thứ tự tăng dần là Fisher (dưới 75%), SVM (dưới 87%), logistic (dưới 93%) và Bayes (92,2% - 95,6%). Phương pháp BayesC cho kết quả rất tốt và ổn định.

4.2.2 Tập kiểm tra

Sử dụng mô hình tối ưu của mỗi phương pháp được thiết lập từ tập huấn luyện, thực hiện phân loại cho 52 phần tử của tập kiểm tra, ta có kết quả sau:

Bảng 5. Bảng phân loại từng phương pháp tối ưu trên tập kiểm tra

Phương pháp	Số phần tử phân loại đúng	Tỉ lệ (%)
Hồi qui Logistic	48	92,31
Fisher	40	76,92
SVM	49	94,23
Bayes	51	98,08

Kết quả phân loại trong Bảng 5 một lần nữa cho thấy phương pháp BayesC cho kết quả tốt, xác suất phân loại đúng đạt kết quả cao nhất.

5 KẾT LUẬN

Bài báo đã nghiên cứu phương pháp phân loại Bayes với những cải tiến và đề xuất để áp dụng được cho dữ liệu rời rạc của thực tế. Đó là vấn đề xác định xác suất tiên nghiệm, ước lượng hàm mật độ xác suất và tính sai số Bayes. Nghiên cứu đã xây dựng một chương trình trên phần mềm Matlab để thực hiện hiệu quả cho số liệu thực. Từ những cải tiến này, nghiên cứu đã đề xuất được thuật toán phân loại bệnh trong y học. Thuật toán này đã được áp dụng hiệu quả cho một tập dữ liệu thực. Thuật toán đề nghị cũng có thể áp dụng tương tự cho nhiều lĩnh vực khác. Nếu có số liệu đủ lớn và tin cậy, bài toán phân loại có thể trở thành một công cụ quan trọng hỗ trợ ngành y trong nghiên cứu chẩn đoán bệnh. Chúng tôi sẽ tiếp tục nghiên cứu đề xuất các phương pháp để chẩn đoán một số bệnh khác trong thời gian sắp tới dựa vào các số liệu thực tế ở Việt Nam.

TÀI LIỆU THAM KHẢO

Cristianini, S. and Shawe, T. J., 2000. An introduction to support vector machines and other kernel-based learning method. Cambridge University, London. 189 pages.

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. Commun Statist Theory Method. 18(10): 3851–3871.

Jasra, A., Holmes, C. and Stephens, D., 2005. Markov chain Monte Carlo methods and the label switching problem in bayesian mixture modeling. Statistical Science 12: 50–67.

Jan, Y. K., Cheng, C. W. and Shih, Y. H., 2010. Application of logistic regression analysis of home mortgageloan prepayment and default. CIC Express Letters 2: 325–331.

Ghosh, A. K., Chaudhuri, P. and Sengupta, D., 2012. Classification using kernel density estimates. Technometrics 48(1): 377–392.

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

Pal, N. R. and Bezdek, J. C., 1995. On cluster validity for the fuzzy c-means model. IEEE Transactions on Fuzzy Systems 3(3): 370–379.

Pham–Gia, T., Turkkan, T. K. and Tai, V. V., 2008. The maximum function in statistical discrimination analysis. Commun.in Stat–Simulation Computation. 37(2): 320–336.

Tai, V. V. and Pham-Gia, T., 2010. Clustering probability distributions. Journal of Applied Statistics 37(11): 1891–1910.

Tai, V. V., 2017. L^1-distance and classification problem by bayesian method. Journal of Applied Statistics, 44(3): 385–401.

Tai, V. V., Loc T. P. and Ha, C. N., 2018. Classifying two populations by Bayesian method and applications. Communication in Mathematics and Statistics, 7(2): 141 – 161.

Thao, N. T. and Tai, V. V., 2017. Fuzzy clustering of probability density functions. Journal of Applied Statistics, 44(4): 583–601.