Tạp chí Khoa học Trường Đại học Cần Thơ
website: sj.ctu.edu.vn

BÀI TOÁN PHÂN LOẠI VÀ ỨNG DỤNG TRONG Y HỌC

Võ Văn Tài và Đồng Yến Nghi
Khoa Khoa học Tự nhiên, Trường Đại học Cần Thơ

Thông tin chung:
Ngày nhận: 23/07/2015
Ngày chấp nhận: 25/02/2016

Title:
Classification problem and applying in medicine

Từ khóa:
Phân loại, Fisher, logistic, Bayes, sai số

Keywords:
Classification, Fisher, logistic, Bayes, error

ABSTRACT

This paper presents methods of classifying Fisher, logistic, SVM, Bayes and their calculating problems. The article also solves real application problem from discrete data of these methods by programs which are built on Matlab software including programs to estimate probability density function, to classify an element and to compute Bayes error. A real application in medicine is presented in detail: Finding the suitable model for assessing hypertensive patient from variables. The result shows that Bayesian method is always the best model. This application is not only as an illustration of the presented theories, to test the logic of the established programs but also to show the application potential of the studied problem.

TÓM TẮT

Bài báo trình bày các phương pháp trong phân loại Fisher, logistic, SVM, Bayes và vấn đề tính toán của chúng. Bài báo cũng giải quyết vấn đề ứng dụng thực tế từ số liệu rời rạc của các phương pháp này bằng các chương trình được xây dựng trên phần mềm Matlab. Đó là chương trình ước lượng hàm mật độ xác suất, phân loại một phần tử và tính sai số Bayes. Một ứng dụng thực tế trong y học được trình bày chi tiết: Tìm mô hình thích hợp trong đánh giá bệnh cao huyết áp từ các biến. Kết quả thực hiện cho thấy phương pháp Bayes luôn cho mô hình tốt nhất. Áp dụng này không những minh họa cho những lý thuyết đã trình bày, kiểm tra sự hợp lý của các chương trình được thiết lập, mà còn cho thấy tiềm năng ứng dụng của vấn đề nghiên cứu.

Trích dẫn: Võ Văn Tài và Đồng Yến Nghi, 2016. Bài toán phân loại và ứng dụng trong y học. Tạp chí Khoa học Trường Đại học Cần Thơ. 42a: 127-133.

1 GIỚI THIỆU

Phân loại là việc gán một phần tử mới thích hợp nhất vào các tổng thể đã được biết trước dựa vào biến quan sát của nó. Hiện tại, ba phương pháp chính được đưa ra để giải quyết bài toán phân loại là: Fisher, logistic và Bayes. Mặc dù được đề xuất muộn nhất và chỉ phân loại cho hai tổng thể, nhưng phương pháp hồi qui logistic được sử dụng rất phổ biến hiện nay. 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. 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. Xác suất sai lầm trong phân loại bằng phương pháp

Bayes được gọi là sai số Bayes. Theo Pham-Gia *et al* (2006), sai số Bayes đã được chứng minh là xác suất sai lầm nhỏ nhất trong bài toán phân loại. Nghiên cứu về sai số Bayes đã được rất nhiều nhà thống kê quan tâm. Một số kết quả mới rất có ý nghĩa về phương pháp Bayes đã được trình bày trong những năm gần đây bởi Pham-Gia *et al* (2006, 2008).

Bài toán phân loại là một hướng phát triển quan trọng của thống kê nhiều chiều. Nó được ứng dụng rộng rãi trong nhiều lĩnh vực khác nhau, đặc biệt trong y học. Cụ thể nó 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.

Đây là một hướng nghiên cứu tiềm năng, được sự quan tâm đặc biệt hiện nay của các nhà khoa học thống kê, công nghệ thông tin và y học trên thế giới. Với những tiến bộ vượt bậc trong việc lưu trữ và xử lý dữ liệu, bài toán phân loại sẽ trở thành một công cụ quan trọng giúp các bác sĩ trong hỗ trợ chẩn đoán bệnh. Ở nước ta, bài toán phân loại chưa được quan tâm nhiều. Do sự hạn chế của phương pháp Fisher, sự phức tạp trong tính toán của phương pháp Bayes, nên các ứng dụng cụ thể chỉ sử dụng phương pháp hồi qui logistic, do đó chưa thể xác định một mô hình phân loại tối ưu trong các ứng dụng cụ thể này. Bài viết này phân tích các phương pháp phân loại, giải quyết vấn đề tính toán để từ đó có thể ứng dụng tìm kiếm mô hình tối ưu trong đánh giá một loại bệnh cụ thể của thực tế.

Cấu trúc của bài báo như sau. Trong phần 2, các phương pháp phân loại được tổng kết. Phần 3 xem xét vấn đề tính toán của các phương pháp này. Phần này cũng thiết lập các chương trình tính toán để hỗ trợ áp dụng thực tế của phương pháp Bayes. Áp dụng thực tế được trình bày trong phần 4 để minh họa lý thuyết và tính ứng dụng của vấn đề nghiên cứu. Cuối cùng là kết luận của bài báo.

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 $x_1, x_2, \ldots, x_n$. 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 đường thẳng thực. Điều đó cho thấy có mối quan hệ chặt chẽ giữa số chênh, thành phần ln(p/(1-p)), và các biến độc lập xi dưới dạng tuyến tính nên người ta thiết lập chúng dưới dạng:

$$y = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \sum_{i=1}^{n}\beta_i x_i. \qquad (2.1)$$

Phương trình (2.1) 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 (2.1) được xác định bởi hệ phương trình sau:

$$\begin{cases} \sum_{i=1}^{n} p_i = \sum_{i=1}^{n}\left(1+\exp\left[-\left(\widehat{\beta}_0 + \sum_{j=1}^{k}\widehat{\beta}_j x_{ij}\right)\right]\right)^{-1} \\ \sum_{i=1}^{n} x_i p_i = \sum_{i=1}^{n} x_i\left(1+\exp\left[-\left(\widehat{\beta}_0 + \sum_{j=1}^{k}\widehat{\beta}_j x_{ij}\right)\right]\right) \end{cases}, \qquad (2.2)$$

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 qui, ta có xác suất thành công của phần tử có biến quan sát $x = (x_1, x_2, \ldots, x_n)$ là:

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

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

$$d_i(x) = \mu_i^T \Sigma^{-1} x - \frac{1}{2}\mu_i^T \Sigma^{-1}\mu_i. \quad (2.3)$$

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 $q_1, q_2, ..., q_k$, $q_1 + q_2 . + ... + q_k = 1$. Ta có nguyên tắc phân loại một phần tử mới với biến quan sát x_0 bằng phương pháp Bayes như sau:

Nếu $g_{\max}(x_0) = q_j f_j(x_0)$ *thì xếp phần tử mới vào* w_j, (2.4)

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:

$$Pe_{1,2,...,k}^{(q)} = \sum_{i=1}^{k}\int_{R^n \setminus R_i^n} q_i f_i dx, \quad (2.5)$$

trong đó n là số chiều của biến quan sát.

Từ công thức (2.5), ta có:

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

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

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

$$= 1 - \int_{R^n}\max_{1\leq l\leq k}\{q_l f_l(x)\}dx. \quad (2.6)$$

Sử dụng (2.6) 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.

Trong trường hợp hai tổng thể, sai số Bayes là vùng chồng lấp giữa $qf_1(x)$ và $(1-q)f_2(x)$, do đó nó có thể tính bởi:

$$Pe_{1,2}^{(q,1-q)} = \int_{R^n}\min\{qf_1(x), (1-q)f_2(x)\}dx. \quad (2.7)$$

3 VẤN ĐỀ TÍNH TOÁN

3.1 Trong phương pháp Fisher và hồi qui logistic

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:

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

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

Lúc này ta sẽ thay thế μ_i bằng $\overline{x}_i$, Σ bởi S_{pooled} trong công thức (2.3).

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.

Để tìm các hệ số của mô hình hồi qui khi có số liệu cụ thể, ta phải giải phương trình (2.2). Tuy nhiên, việc giải này thực sự rất phức tạp, vì vậy trong thực tế ta chỉ sử dụng các gói hỗ trợ

trong những phần mềm thống kê như SPSS, R,... để thực hiện.

3.2 Trong phương pháp Bayes

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 rời rạc. 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 nhất. 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.

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

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

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 bài viết, chúng tôi chọn tham số trơn theo Scott (1992):

$h_j = \left(\frac{4}{N(n+2)}\right)^{\frac{1}{n+4}}\sigma_j$, σ_j là độ lệch chuẩn mẫu của biến thứ j.

Các phần mềm thống kê như Mattlab, 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, chúng tôi đã viết chương trình thực hiện trong trường hợp này với hàm hạt nhân và tham số trơn được chọn ở trên. Khi có các hàm mật độ xác suất ước lượng, dựa vào (2.4) chúng tôi đã viết chương trình để phân loại một phần tử mới.

Kết quả phân loại một phần tử mới bởi nguyên tắc (2.4) và sai số Bayes được tính bởi (2.6) đều phụ thuộc vào xác suất tiên nghiệm. Mặc dù có nhiều tác giả đã nghiên cứu về vấn đề này như McLachlan và Basford (1988), Inman (1989), Miller (2001) 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:

(a) Dựa vào phân

(b) Dựa vào tập mẫu: phối đều: $q_1 = q_2 = ... = q_c = \frac{1}{c}$. $q_i = \frac{n_i}{N}$.

(c) Dựa vào ước lượng Laplace: $q_i = \frac{n_i + 1}{N + n}$.

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.

Mặt dù về mặt lý thuyết, chúng ta chưa khẳng định việc chọn xác suất tiên nghiệm theo phương pháp nào là hợp lý, tuy nhiên các ứng dụng thực tế cho thấy việc chọn theo (b) thường cho kết quả tốt nhất.

Để tính sai số Bayes, Pham-Gia *et al* (2006) 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 (xem Pham-Gia *et al.,* 2008). Chúng tôi sử dụng cách tính gần đúng hàm cực đại của các hàm mật độ xác suất bằng phương pháp Monte-Carlo, 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 thiết lập 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 CỤ THỂ TRONG Y HỌC

4.1 Giới thiệu

Chúng tôi lấy một bộ số liệu thực tế để minh họa cho việc ứng dụng các mô hình phân loại trong đánh giá một loại bệnh. Cụ thể ứng dụng thực hiện phân loại bệnh cao huyết áp dựa vào số liệu thứ cấp thu được từ tổng kết của ngành y tế tỉnh Sóc Trăng từ tháng 12/2011 đến 12/2013. Số liệu mẫu gồm 54 người bệnh và 96 người không bệnh. Các biến sinh hóa ban đầu được chọn để đưa vào mô hình phân loại theo ý kiến của chuyên gia y tế. Cụ thể các biến được khảo sát gồm:

Bảng 1: Các biến độc lập được khảo sát.

STT	Chỉ tiêu	Thang đo	Ký hiệu
1	Tuổi	Tuổi	TU
2	Giới tính	0: Nữ - 1:Nam	GT
3	Trọng lượng	Kg	TL
4	Chiều cao	Centimet (cm)	CC
5	Huyết áp trung bình	MmHg	HA
6	Tỉ trọng cơ thể	Kg/m^2	BMI
7	Nhịp tim	Lần/phút	NT

4.2 Phương pháp thực hiện

Với số liệu đã có, bài viết thực hiện phân tích xem các biến độc lập có tương quan với nhau không để tránh trường hợp đa cộng tuyến. Sau khi loại bỏ bớt biến để không còn đa cộng tuyến, bài viết xây dựng mô hình logistic, lựa chọn các biến có ý nghĩa thống kê ở mức 5%. Với các biến đã lựa chọn tiến hành đánh giá khả năng mắc bệnh tăng huyết áp của bệnh nhân theo 3 phương pháp: Fisher, logistic và Bayes. Mỗi phương pháp sẽ tiến hành đánh giá tính hợp lý giữa biến phụ thuộc với tất cả các biến độc lập để lựa chọn mô hình phù hợp nhất. Cuối cùng chúng ta sẽ cho những nhận xét về các phương pháp đã thực hiện để lựa chọn được phương pháp phù hợp nhất. Các tính toán trong phương pháp Fisher và logistic được thực hiện trên phần mềm SPSS. Những tính toán trong phương pháp Bayes bao gồm việc ước lượng hàm mật độ xác suất, phân loại phần tử mới, tính sai số Bayes đều dựa trên các chương trình đã viết trên Matlab với xác suất tiên nghiệm được chọn theo b)

4.3 Kết quả thực hiện

Từ số liệu, tính hệ số tương quan giữa các biến ta có kết quả:

Bảng 2: Hệ số tương quan cặp của biến đưa vào mô hình

		TU	TL	CC	HA	BMI	NT
TU	Pearson Correlation	1	-0.327**	-0.160*	.235**	-0.281**	-0.768**
	Sig. (2-tailed)		0.000	0.050	.004	0.000	0.000
	N	150	150	150	150	150	150
TL	Pearson Correlation	-0.327**	1	0.583**	.054	**0.816****	0.241**
	Sig. (2-tailed)	0.000		0.000	.515	0.000	0.003
	N	150	150	150	150	150	150
CC	Pearson Correlation	-0.160*	0.583**	1	.107	0.015	0.175*
	Sig. (2-tailed)	0.050	0.000		.194	0.858	0.032
	N	150	150	150	150	150	150
HA	Pearson Correlation	0.235**	0.054	0.107	1	0.006	-0.168*
	Sig. (2-tailed)	0.004	0.515	0.194		0.937	0.040
	N	150	150	150	150	150	150
BMI	Pearson Correlation	-0.281**	0.816**	0.015	0.006	1	0.173*
	Sig. (2-tailed)	0.000	0.000	0.858	0.937		0.034
	N	150	150	150	150	150	150
NT	Pearson Correlation	-0.768**	0.241**	0.175*	-0.168*	0.173*	1
	Sig. (2-tailed)	0.000	0.003	0.032	0.040	0.034	
	N	150	150	150	150	150	150

Bảng 2 cho ta thấy hai biến TL và BMI có sự tương quan chặt chẽ với nhau, vì vậy trong thực hiện bài toán phân loại chúng ta bỏ 1 biến. Vì biến BMI chứa biến TL nên chúng tôi loại biến TL. Tiến hành phân tích hồi qui logistic, ta có kết quả xử lý được cho bởi bảng tổng hợp sau:

Bảng 3: Bảng phân tích hồi qui logistic cho 6 biến

	B	S.E.	Wald	Df	Sig.	Exp(B)
TU	-0.172	0.112	2.349	1	0.125	0.842
GT	-1.691	1.540	1.207	1	0.272	0.184
CC	0.099	0.128	0.603	1	0.437	1.104
HA	0.312	0.076	16.906	1	0.000	1.366
BMI	0.436	0.358	1.487	1	0.223	1.547
NT	-0.487	0.233	4.373	1	0.037	0.615
Constant	-14.20	30.407	0.218	1	0.640	0.000

Bảng 3 cho ta thấy chỉ có 2 biến HA và NT có ý nghĩa thống kê 5% khi đưa vào mô hình, các biến còn lại không có ý nghĩa ở mức này. Điều này cho thấy các biến không đóng vai trò quan trọng đối với khả năng mắc bệnh cao huyết áp, tuy nhiên theo ý kiến tham khảo từ chuyên gia, biến TU cũng có khả năng ảnh hưởng đến khả năng tăng huyết áp. Chính vì vậy, chúng tôi sử dụng 3 biến này để thực hiện bài toán phân loại.

Thực hiện phương pháp Fisher, logistic, Bayes trong 3 trường hợp một biến, 3 trường hợp hai biến và 1 trường hợp ba biến để tìm mô hình có xác suất phân loại tốt nhất ta được bảng tổng hợp sau:

Bảng 4: Bảng tổng hợp khả năng phân loại đúng (%) của 3 phương pháp.

Phương pháp	1 biến			2 biến			2 biến
	HA	NT	TU	HA,NT	HA,TU	NT,TU	HA, NT, TU
Hồi qui logistic	95,7	65,3	64,7	98,7	98,7	66,0	98,0
Fisher	95,7	62,0	64,0	98,7	96,3	62,7	97,2
Bayes	98,5	67,2	70,5	98,7	98,0	70,7	98,9

Bảng 4 cho ta những nhận xét sau:

Hai biến NT và TU cũng ảnh hưởng đến bệnh cao huyết áp, tuy nhiên chúng ở mức thấp. Nếu sử dụng 1 biến thì biến HA là yếu tố quyết định. Việc kết hợp hai biến (HA,TU) và 3 biến (HA, NT, TU) cũng cho một chẩn đoán tốt bệnh cao huyết áp.

Ma trận hiệp phương sai của hai nhóm gần xấp xỉ nhau nên kết quả khảo sát của các phương pháp không có sự khác biệt nhiều. Tuy nhiên, phương pháp hồi qui logistic và Bayes có ưu thế hơn. Trong đó, sử dụng phương pháp Bayes với 3 biến sẽ cho kết quả phân loại cao nhất.

Trong từng trường hợp tối ưu, khả năng phân loại đúng rất cao, vì vậy chúng ta có thể sử dụng kết quả này cho thực tế. Để đề phòng bệnh cao huyết áp, chúng ta cần quan tâm đến tuổi tác, kiểm soát huyết áp trung bình và nhịp tim. Các biến này cũng dùng để dự báo người bị tăng huyết áp với xác suất rất cao.

5 KẾT LUẬN

Bài báo đã tổng kết các phương pháp phân loại và khảo sát vấn đề tính toán. Các chương trình được viết bằng Matlab phục vụ cho việc tính toán hiệu quả các áp dụng từ số liệu của thực tế. Bài toán áp dụng trong y học từ số liệu thực tế được trình bày, minh chứng cho tiềm năng ứng dụng của bài toán phân loại trong lĩnh vực này cũng như những lĩnh vực khác. Nếu có đầy đủ số liệu tin cậy và công cụ tính toán đủ mạnh, bài toán phân loại sẽ trở thành một công cụ quan trọng giúp ngành y trong nghiên cứu chẩn đoán bệnh. Chúng tôi sẽ tiếp tục nghiên cứu chẩn đoán một số bệnh khác trong thời gian sắp tới dựa vào các số liệu thực tế ở Việt Nam.

LỜI CẢM TẠ

Chúng tôi trân trọng cảm ơn Phòng y tế huyện Vĩnh Châu, tỉnh Sóc Trăng đã cung cấp số liệu và có những ý kiến đóng góp cho việc thực hiện. Xin trân trọng cảm ơn các phản biện đã cho những ý kiến quý báo để bài viết được hoàn thiện hơn.

TÀI LIỆU THAM KHẢO

Devijver P. A and Kittler K., 1982. Pattern recognition, a statistical approach. Prentice Hall, London.

Fukunaga K., 1990. Introduction to statistical pattern recognition. Academic Press, New York.

Hand D. J., 1982. Discriminant and classification. John Wiley & Sons, New York.

Hand D. J., 1982. Kernel discriminant analysis. Letchworth, London.

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: 3851-3871.

Martinez W. L. and Martinez A. R., 2008. Computational statistics handbook with Matlab. Chapman & Hall/CRC, Boca Raton.

McLachlan G. J. and Basford K., 1988. Mixture models. Marcel Dekker, New York.

G.Miller et al, 2011. Baysian prior probability distributions for internal dosimetry. Radiat. Prot. Dosim. 94(4): 347-352.

Pham-Gia T. et al, 2006. Baysian analysis in the L1– norm of the mixing proportion using discriminant analysis. Metrika. 64(1):1-22.

Pham–Gia T. et al, 2006. Bounds for the Bayes error in classification: A Bayesian approach using discriminant analysis. Statistical Methods and Applications. 16: 7- 26.

Pham–Gia T. et al, 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. John Wiley & Son, New York.

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