PHÂN LOẠI BẰNG PHƯƠNG PHÁP BAYES TỪ SỐ LIỆU RỜI RẠC

Võ Văn Tài[1]

ABSTRACT

The paper represents classification problem by Bayesian method from discrete data through program estimating n – dimenstion probability density function, classifing a new element and calculating Bayes error which are written on Matlab software. The programs are used to for specific applied from real discrete data.

Keywords: Bayes method, Bayes error, classification, probability density function

Title: Classification by Bayesian method from discrete data

TÓM TẮT

Bài báo trình bày bài toán phân loại bằng phương pháp Bayes từ số liệu rời rạc, qua chương trình ước lượng hàm mật độ xác suất, phân loại một phần tử mới và tính sai số Bayes được viết trên phần mềm Matlab. Các chương trình này được sử dụng để thực hiện cho các ứng dụng cụ thể từ số liệu rời rạc thực tế.

Từ khóa: Phương pháp Bayes, sai số Bayes, phân loại, hàm mật độ xác suất

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. Đây là một hướng phát triển quan trọng của nhận dạng không được giám sát của thống kê. Bài toán phân loại được ứng dụng rộng rãi trong nhiều lĩnh vực khác nhau, đặc biệt trong xã hội, sinh học và y học. Hiện tại có ba phương pháp chính được đưa ra để giải quyết bài toán phân loại: phương pháp Fisher, phương pháp hồi quy logistic và phương pháp Bayes [2], [3], [10]. Phương pháp hồi quy logistic được sử dụng phổ biến nhất hiện nay, nhưng nó chỉ áp dụng cho dữ liệu rời rạc và chỉ phân loại cho hai tổng thể. Phương pháp Fisher cũng áp dụng cho dữ liệu rời rạc, mặc dù 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 có thể phân loại cho hai và nhiều hơn hai tổng thể, được xem có nhiều ưu điểm nhất vì nó đã đạt được mục tiêu về mặt lý thuyết cho bài toán phân loại. 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. Một ưu điểm nổi bật của phương pháp này là tính được xác suất sai lầm trong phân loại mà nó được gọi là sai số Bayes. Sai số Bayes đã được chứng minh là xác suất sai lầm nhỏ nhất trong bài toán phân loại. Một số kết quả mới rất có ý nghĩa về phương pháp Bayes đã được trình bày trong những năm gần đây bởi các bài báo [6], [7], [8].

Một cản trở lớn của việc áp dụng thực tế bài toán phân loại bằng phương pháp Bayes trong những lĩnh vực cụ thể là vấn đề tính toán. Phương pháp Bayes dựa trên cơ sở hàm mật độ xác suất đã biết, tuy nhiên số liệu thực tế là số liệu rời rạc,

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

vì vậy để phân loại bằng phương pháp Bayes có ý nghĩa thực tế việc đầu tiên là phải ước lượng hàm mật độ xác suất. Vấn đề tính sai số Bayes, phân loại một phần tử mới còn rất nhiều khó khăn khi gặp số liệu lớn của thực tế. Trong bài viết này, chúng tôi quan tâm đến lý thuyết tính toán các vấn đề liên quan đến phân loại bằng phương pháp Bayes từ số liệu rời rạc. Đặc biệt đưa ra một công thức tương đương của sai số Bayes mà nó rất thuận lợi cho việc tính toán. Các lý thuyết liên quan đến việc tính toán này sẽ được cụ thể hóa bằng các chương trình được viết trên phần mềm Matlab. Các chương trình này sẽ được sử dụng để áp dụng cho bài toán phân loại từ các số liệu rời rạc thực tế trong lĩnh vực sinh học và y học.

2 PHƯƠNG PHÁP BAYES

2.1 Phân loại một phần tử mới

Cho k tổng thể $w_1, w_2, ..., w_k$ có biến quan sát với hàm mật độ xác suất được xác định là $f_1(x), f_2(x), \ldots, f_k(x)$ 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 mới với biến quan sát x bằng phương pháp Bayes như sau:

$$\textit{Nếu } g_{\max}(x) = q_j f_j(x) \textit{ thì xếp phần tử mới vào } w_j, \quad (1)$$

Trong đó:

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

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

2.2 Sai số bayes

a) *Trường hợp hai tổng thể*

Trong trường hợp không quan tâm đến xác suất tiên nghiệm q của w_1, ta có:

$\tau_1 = \mathrm{P}(w_2|w_1) = \int_{R_2^n} q f_1(\boldsymbol{x}) dx$: xác suất phân loại một phần tử vào w_2 khi nó thuộc w_1.

$\tau_2 = \mathrm{P}(w_1|w_2) = \int_{R_1^n} (1-q) f_2(\boldsymbol{x}) dx$: xác suất phân loại một phần tử vào w_1 khi nó thuộc w_2.

Trong đó: $R_1^n = \{x \mid q f_1(x) \geq (1-q) f_2(x)\}$, $R_2^n = \{x \mid q f_1(x) < (1-q) f_2(x)\}$.

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

$$Pe = \tau_1 + \tau_2. \quad (2)$$

Khi quan tâm đến xác suất tiên nghiệm q của w_1 thì τ_1 trở thành $\hat{\tau}_1$ và τ_2 trở thành $\hat{\tau}_2$ với

$$\hat{\tau}_1 = \int_{\hat{R}_2^n} q f_1(x) dx \text{ và } \hat{\tau}_2 = \int_{\hat{R}_1^n} (1-q) f_2(x) dx$$

Trong đó $\hat{R}_1^n = \{x \mid q f_1(x) \geq (1-q) f_2(x)\}$, $\hat{R}_2^n = \{x \mid q f_1(x) < (1-q) f_2(x)\}$.

Đặt $(q) = (q, 1-q)$, khi đó sai số Bayes xác định bởi

$$Pe^{(q)} = \tau_1^* + \tau_2^*. \quad (3)$$

τ_1 và τ_2; $\hat{\tau}_1$ và $\hat{\tau}_2$ được gọi chung là hai thành phần của sai số Bayes.

b) *Trường hợp nhiều hơn hai tổng thể*

Sai số Bayes trong phân loại *k* tổng thể được định nghĩa bởi biểu thức

$$Pe^{(q)}_{1,2,\ldots,k} = \sum_{i=1}^{k} \int_{R^n \setminus R_i^n} q_i f_i d\boldsymbol{x} \tag{4}$$

Để thuận lợi hơn trong tính sai số Bayes, người ta thường tính xác suất của sự phân loại đúng $Pe^{(c)}_{1,2,\ldots,k} = \sum_{i=1}^{k} \int_{R_i^n} q_i f_i d\boldsymbol{x}$, khi đó sai số Bayes sẽ được tính bởi

$$Pe^{(q)}_{1,2,\ldots,k} = 1 - Pc^{(q)}_{1,2,\ldots,k}.$$

3 CÁC CHƯƠNG TRÌNH TÍNH TOÁN CHO PHƯƠNG PHÁP BAYES TỪ SỐ LIỆU RỜI RẠC

3.1 Ước lượng hàm mật độ xác suất

Hiện tại có nhiều phương pháp tham số cũng như phi tham số để ước lượng hàm mật độ xác suất. 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:

$$\hat{f}(x) = \frac{1}{N h_1 h_2 \ldots h_n} \sum_{i=1}^{N} \prod_{j=1}^{n} K_j \left(\frac{x_i - x_{ij}}{h_j} \right), \tag{5}$$

Trong đó:

h_j là tham số trơn cho biến thứ *j*, $h_j > 0$.

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.

Theo [10] có thể chọn nhiều dạng hàm hạt nhân khác nhau như tam giác, hình chữ nhật, song lượng,... Trong bài báo này chúng tôi chọn hàm hạt nhân dạng chuẩn:

$$K(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-x^2/2\right). \tag{6}$$

Có nhiều nghiên cứu về việc chọn tham số trơn, nhưng theo Scott (1992) không có sự lựa chọn nào là tối ưu. Việc chọn tham số trơn quan trọng hơn hàm hạt nhân. Trong bài viết này chúng tôi chọn tham số trơn theo Scott:

$$h_j = \left(\frac{4}{N(n+2)} \right)^{\frac{1}{n+4}} \sigma_j \ . \tag{7}$$

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

Sử dụng phần mềm Mattlab, chúng tôi đã viết các chương trình ước lượng hàm mật độ xác suất *n* chiều như sau:

Chương trình 1: *Chương trình ước lượng hàm mật độ xác suất n chiều*

```
function f=uocluongnc(dl1,dl2,…,dln)
    % dl1, dl2,…,dln lần lượt là n chiều của dữ liệu
     syms x1 x2 … xn
     s = sym('s(x1,x2,…,xn)');
     f = sym('f(x1,x2,…,xn)');

    h1 = std(dl1)*(4/length(dl1)*(n+2))^(1/(n+4));
    h2 = std(dl2)*(4/length(dl2)*(n+2))^(1/(n+4));
                    ……;
     hn = std(dln)*(4/length(dln)*(n+2))^(1/(n+4));
     s = 0;
     for  i= 1:length(dl1)
     s=s+(1/(2*pi)^.5*exp(-(((x1-dl1(1,i))/h1)^2/2)))*
        (1/(2*pi)^.5*exp(-(((x2- dl2(1,i))/h2)^2/2)))*…*
        (1/(2*pi)^.5*exp(-(((xn-dln(1,i))/hn)^2/2)));
    end
     s;
    f = 1/(length(dl1)*h1*h2*…*hn)*s;
```

3.2 Phân loại một phần tử mới

Để phân loại một phần tử mới, theo nguyên tắc (1) đầu tiên chúng ta phải tìm hàm cực đại của các hàm mật độ xác suất. Việc tìm một biểu thức giải tích cụ thể cho hàm cực đại này là một công việc vô cùng phức tạp, ngay cả trường hợp một chiều. Nhưng sử dụng phần mềm Matlab, chúng ta có thể dễ dàng thiết lập chương trình để phân loại một phần tử mới như sau:

Chương trình 2: *Chương trình phân loại một phần tử mới n chiều với k tổng thể*

```
function A=phanloai(f1,f2,...fk,x11,x12,x13,…,x1n)
    syms x1 x2 x3…xn
    f=sym('f(x1,x2,…,xn)');
    f=[f1 f2 … fn];
    y=subs(f,{x1,x2,…,xn},{x11,x12,…,x1n});
    [a,i]= max(y);
    A=[a,i] ;
```

3.3 Tính sai số Bayes

Giả sử $\max_{1\le l\le k}\{q_l f_l(\boldsymbol{x})\} = q_j f_j$ trên miền R_j^n. Sai số Bayes tính theo công thức (4) được tương đương như sau:

$$
\begin{aligned}
Pe^{(q)}_{1,2,\ldots,k} &= \sum_{j=1}^{k} \int_{R^n \setminus R^n_j} q_j f_j(x)dx \\
&= \sum_{j=1}^{k} \left[\int_{R^n} q_j f_j(x)dx - \int_{R^n_j} \max_{1\le l\le k}\{q_l f_l(x)\}dx \right] \\
&= \int_{R^n} \sum_{j=1}^{k} q_j f_j(x)dx - \sum_{j=1}^{k} \int_{R^n_j} \max_{1\le l\le k}\{q_l f_l(x)\}dx \\
&= 1 - \int_{R^n} \max_{1\le l\le k}\{q_l f_l(x)\}dx \\
&= 1 - \int_{R^n} g_{\max}(x)dx
\end{aligned} \tag{8}
$$

So với công thức (4), công thức (8) cho ta một thuận lợi rất lớn trong việc tính toán.

Tuy nhiên, khó khăn lớn nhất của công thức (8) không những là việc xác định hàm $g_{\max}(x)$ mà còn là việc tính tích phân của hàm này trên miền R^n. Trong bài viết này sau khi ước lượng hàm mật độ xác suất theo chương trình 1, chúng tôi tính gần đúng sai số Bayes theo (8) dựa trên việc tính gần đúng tích phân hàm $g_{\max}(x)$ theo phương pháp Moncte Carlo, một phương pháp tính gần đúng tích phân hiệu quả nhất hiện nay. Phương pháp này cụ thể như sau:

Xét tích phân n chiều trên miền V: $I = \int_V f(x)dx,\ x \in R^n$. Khi đó ước lượng $\hat{I}$ của I xác định như sau:

$$\hat{I} = \frac{Mes(V)}{N} \sum_{i=1}^{N} f(x_i) \tag{9}$$

Trong đó x_i là các giá trị lấy ngẫu nhiên độc lập trong miền V,

N là tổng số lần lấy mẫu x_i,

Mes(V) là độ đo của miền V.

Sử dụng việc tính gần đúng tích phân bằng phương pháp Moncte-Carlo, chương trình tính sai số Bayes theo (8) được viết như sau:

Chương trình 3: *Tính sai số Bayes cho k tổng thể n chiều*

```
function h = errorbayes(f1,f2,…,fk)
syms x1 x2 …xn fmax
f = [f1 f2 …fk];
a1 =[random points of the first demension]
a2 = [random points of the second demention]
                    ……..…..
an = [random points of nth demention]
an + 1= rand(1,N)
p = 0;
for i=1:length(a1)
fmax=max(subs(f,{x1,x2,…,xn},{a1(1,i), a2(1,i),…,an(1,i)}));
```

```
f max(subs(f,{x1,x2,…,xn},{a1(1,i), a2(1,i),…,an(1,i)}))<= ap(i)
p = p+1;
end
end
p;
fmax;
gttp = sum(fmax)/(length(a1)^2*(max(a1)-min(a1))*(max(a2)-
min(a2))*… *(max(an)-min(an))*p;
errorb = 1-gttp*q;
h = double(errorb);
```

4 MỘT SỐ ỨNG DỤNG

Bài toán phân loại xuất phát từ nhu cầu của thực tế của nhiều lĩnh vực khác nhau. Ở đây chúng tôi trình bày hai ví dụ trong sinh học và y học để minh họa cho tính ứng dụng của bài toán phân loại bằng phương pháp Bayes. Đây là những ví dụ minh họa cho vô số những ứng dụng thực tế có thể áp dụng trong những lĩnh vực khác. Trong mỗi áp dụng chúng tôi thực hiện bài toán phân loại theo hai cách: Tính toán thủ công từng bước một bằng Excell theo các công thức (5), (6), (7), sau đó sử dụng nguyên tắc (1) để phân loại, đồng thời sử dụng các chương trình đã viết (chương trình 1, chương trình 2) để kiểm chứng kết quả phân loại theo hướng tính thủ công ở trên. Sai số Bayes trong mỗi áp dụng cũng được tính từ chương trình 3.

4.1 Ứng dụng 1

Năm 1990, trong một nghiên cứu tìm hiểu mối liên hệ giữa nguy cơ gãy xương (fx) và mật độ xương cùng một số chỉ số sinh hóa khác như độ tuổi (age), tỉ trọng cơ thể (bmi), mật độ chất khoáng trong xương (bmd), chỉ số hủy xương (ictp), chỉ số tạo xương (pinp). Một nhóm bác sĩ chọn một mẫu gồm 137 người có độ tuổi từ 60 trở lên theo dõi trong 15 năm, ghi nhận bị gãy xương hay không. Số liệu cụ thể được lấy từ bài viết của bác sĩ Nguyễn Văn Tuấn trên trang Webb www.ykhoanet.com. Với số liệu trên chúng ta cần tìm mối quan hệ giữa fx với các biến định lượng age, bmi, bmd, ictp và pinp, để từ đó xác định nếu một người có 4 chỉ số cụ thể, chẳng hạn: x_0 = (**age = 60, bmi = 24.500, bmd = 0.796, ictp = 6.420, pinp = 37.813)** thì kết luận người này có nguy cơ bị gãy xương hay không. Tính xác suất sai lầm trong phân loại này.

i) Tính toán từng bước

- Do không có thông tin ban đầu nên ta giả sử xác suất tiên nghiệm có nguy cơ gãy xương và không có nguy cơ gãy xương đều bằng nhau: $q_i = 1/2$.
- Tính giá trị của $f_i(x_0)$ bằng cách ước lượng $f_i(x)$ theo phương pháp hàm hạt nhân. Cụ thể:

$$f_i(x_0) = \frac{1}{N_i h_{1i} h_{2i} h_{3i} h_{4i} h_{5i}} \sum_{k=1}^{N_i} \prod_{j=1}^{5} K_j \left(\frac{x_{j0} - x_{jk}^{(i)}}{h_{ji}} \right),$$

Trong đó

$x_{jk}^{(i)}$ là phần tử mẫu thứ k, biến thứ j của nhóm thứ i, với $i = 1$ là nhóm có nguy cơ gãy xương, $i = 0$ là nhóm không có nguy cơ gãy xương.

$N_i = N_2 = 137$, $x_{10} = 60$, $x_{20} = 24.500$, $x_{30} = 0.796$, $x_{40} = 6.420$, $x_{50} = 37.810$

$h_{ji}, j = 1, 2, 3, 4, 5$; $i = 1, 2$ được tính từ số liệu mẫu theo công thức (7).

Cụ thể

$h_{11} = 4.6079$, $h_{21} = 32.59010$, $h_{31} = 0.13611$, $h_{41} = 1.17092$, $h_{51} = 12.23564$

$h_{12} = 2.33054$, $h_{22} = 1.89549$, $h_{32} = 0.07498$, $h_{42} = 0.64204$, $h_{52} = 7.36715$.

$K_j(.)$ là hàm hạt nhân dạng chuẩn, được tính bởi (6).

Lập bảng tính trên phần mềm Excell với các tham số cụ thể ở trên, ta có được giá trị cụ thể

$$f_1(x_0) = 2.86085\text{E-}08,\ f_2(x_0) = 3.63723\text{E-}05,$$

$$d_1(x_0) = \frac{1}{2} f_1(x_0) = \frac{1}{2} 2.86085E-08,\ d_2(x_0) = \frac{1}{2} f_2(x_0) = \frac{1}{2} 3.63723E-05.$$

- Vì $d_1(x_0) < d_2(x_0)$ như vậy theo (1) người này được xếp vào nhóm không có nguy cơ gãy xương.

ii) Sử dụng chương trình đã viết

Sử dụng chương trình 1 để ước lượng hàm mật độ xác suất 5 chiều từ 137 số liệu mẫu. Sử dụng chương trình 2 để phân loại một phần tử mới có biến quan sát x_0

ở trên với $k = 2$, $n = 5$, ta có kết quả xuất ra như sau:

ans = 0.012254 0

Trong đó $f_{\max}(x_0) = 0.012254$.

Vậy x_0 thuộc nhóm không có nguy cơ bị gãy xương (nhóm $i = 0$).

Chương trình 3 cũng với $k = 2$, $n = 5$, ta có kết quả xuất ra như sau:

ans = 0.3855

Vậy sai số Bayes hay xác suất sai lầm của phân loại này là 0.3855.

4.2 Ứng dụng 2

Hoa Iris là một loại có giá trị dược liệu, nhưng có nhiều loại khác nhau. Mỗi loại có một giá trị dược liệu khác nhau mà mắt thường không thể phân biệt được. Có 3 loại khó phân biệt và được quan tâm nhiều nhất là Setosa (Se), Versicolor (Ve), Virginica (Vi). Chọn từ mỗi loại 50 phần tử, quan sát 4 biến x_1: độ dài của đài hoa, x_2: độ rộng của đài hoa, x_3: độ dài của cánh hoa, x_4: độ rộng của cánh hoa. Ta có số liệu mẫu được cho trong phụ lục. Sử dụng phương pháp Bayes từ số liệu rời rạc này, xác định cụ thể nếu một hoa Iris có 4 biến cụ thể, chẳng hạn $x_0 = (5\ 3\ 1\ 0.3)$ thì nó sẽ thuộc loại nào. Tính xác suất sai lầm của sự phân loại này.

i) Tính toán từng bước

- Giả sử xác suất tiên nghiệm của 3 nhóm hoa đều bằng nhau: $q_i = 1/3$, $i = 1, 2, 3$.
- Tính giá trị của $f_i(x_0)$ bằng cách ước lượng $f_i(x)$ theo phương pháp hàm hạt nhân. Cụ thể

$$f_i(x_0) = \frac{1}{N_i h_{1i} h_{2i} h_{3i} h_{4i}} \sum_{k=1}^{N_i} \prod_{j=1}^{4} K_j\left(\frac{x_{j0} - x_{jk}^{(i)}}{h_{ji}}\right),$$

Trong đó

$x_{jk}^{(i)}$ là phần tử mẫu thứ k, biến thứ j của nhóm thứ i, gán $i = 1$ cho (Se), $i = 2$ cho (Ve), $i = 3$ cho (Vi).

$N_i = N_2 = N_3 = 50$, $x_{10} = 5$, $x_{20} = 3$, $x_{30} = 1$, $x_{40} = 0.3$.

h_{ji}, $j = 1, 2, 3, 4$; $i = 1, 2, 3, 4$ được tính từ số liệu mẫu theo công thức (7). Cụ thể

$h_{11} = 0.20548$, $h_{21} = 0.22097$, $h_{31} = 0.10123$, $h_{41} = 0.06143$;
$h_{12} = 0.30089$, $h_{22} = 0.18292$, $h_{32} = 0.27393$, $h_{42} = 0.11528$;
$h_{13} = 0.37067$, $h_{23} = 0.18799$, $h_{33} = 0,32172$, $h_{43} = 0.16010$.

Hàm hạt nhân $K_j(.)$ vẫn được chọn dạng chuẩn.

Cũng tính toán từng bước, ta có kết quả

$$d_1(x_0) = \frac{1}{3} f_1(x_0) = 0.16960,\ d_2(x_0) = \frac{1}{3} f_2(x_0) = 0.205239E-35,$$
$$d_3(x_0) = \frac{1}{3} f_3(x_0) = 6.54542E-46.$$

- Vì $d_1(x_0) = \max\{d_1(x_0), d_2(x_0), d_3(x_0)\}$, do đó theo (1) hoa Iris này thuộc nhóm Setosa.

ii) Sử dụng chương trình đã viết

Sử dụng chương trình 1 để ước lượng hàm mật độ xác suất 4 chiều từ 50 số liệu mẫu. Sử dụng chương trình 2 để phân loại một phần tử mới có biến quan sát x_0 ở trên với $k = 3$, $n = 4$, ta có kết quả xuất ra như sau:

ans = 0.16960 1

Trong đó $f_{\max}(x_0) = 0.16960$.

Vậy x_0 cũng được xếp vào nhóm 1, tức hoa Iris này thuộc loại Setosa.

Chương trình 3 với $k = 3$, $n = 4$, ta có kết quả xuất ra như sau:

ans = 0.03200

Vậy sai số Bayes là 0.03200.

5 KẾT LUẬN

Bài báo đã trình bày bài toán phân loại bằng phương pháp Bayes và các vấn đề lý thuyết liên quan đến việc tính toán của phương pháp này từ số liệu rời rạc. Viết các chương trình trên phần mềm Matlab phục vụ cho việc tính toán từ lý thuyết đã nêu. Điều này đã làm cho bài toán phân loại bằng phương pháp Bayes thật sự có ý nghĩa thực tế. Hai ví vụ minh họa cho nhiều ví dụ có thể áp dụng trong lĩnh vực y học và sinh học được khảo sát. Chúng ta tin rằng nếu có đầy đủ số liệu tin cậy và công cụ tính toán đủ mạnh, bài toán phân loại bằng phương pháp Bayes sẽ trở thành một công cụ quan trọng trong nhiều lĩnh vực khác. Để làm được điều này chúng ta cần có sự kết hợp chặt chẽ giữa các nhà khoa học trong lĩnh vực thực hiện, thống kê và công nghệ thông tin.

TÀI LIỆU THAM KHẢO

[1] Devijver. P.A. and Kittler, J., *Pattern recognition, a statistical approach,* Prentice Hall, London, 1982.

[2] Fukunaga, K., *Introduction to statistical pattern recognition*, Academic Press, New York, 1990.

[3] Hand, D.J., *Discriminant and classification,* John Wiley & Sons, New York, 1981.

[4] Hand, D.J. *Kernel discriminant analysis*,1982, Research studies press, Letchworth.

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

[6] Pham–Gia,T. and Turkkan, N., Baysian analysis in the L^1– norm of the mixing proportion using discriminant analysis, *Metrika,* 64(1), 2006, 1–22.

[7] Pham–Gia, T., Turkkan, N. and Bekker, A., Bounds for the Bayes error in clssification: A Bayesian approach using discriminant analysis, *Statistical Methods and Applications,*16, 2006, 7 - 26.

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

[9] Scott, David W. , *Mutivariate density estimation:Theory,practice and visualization visualization,* John Wiley & Son, New York, 1992.

[10] Webb, A., *Statistical pattern recognition*, John Wiley & Sons, New York, 2000.

PHỤ LỤC: DỮ LIỆU CHO ỨNG DỤNG 2

Setosa (Se)				**Versicolor (Ve)**				**Virginica (Vi)**			
x_1	x_2	x_3	x_4	x_1	x_2	x_3	x_4	x_1	x_2	x_3	x_4
5.1	3.5	1.4	0.2	7.0	3.2	4.7	1.4	6.3	3.3	6.0	2.5
4.9	3.0	1.4	0.2	6.4	3.2	4.5	1.5	5.8	2.7	5.1	1.9
4.7	3.2	1.3	0.2	6.9	3.1	4.9	1.5	7.1	3.0	5.9	2.1
4.6	3.1	1.5	0.2	5.5	2.3	4.0	1.3	6.3	2.9	5.6	1.8
5.0	3.6	1.4	0.2	6.5	2.8	4.6	1.5	6.5	3.0	5.8	2.2
5.4	3.9	1.7	0.4	5.7	2.8	4.5	1.3	7.6	3.0	6.6	2.1
4.6	3.4	1.4	0.3	6.3	3.3	4.7	1.6	4.9	2.5	4.5	1.7
5.0	3.4	1.5	0.2	4.9	2.4	3.3	1.0	7.3	2.9	6.3	1.8
4.4	2.9	1.4	0.2	6.6	2.9	4.6	1.3	6.7	2.5	5.8	1.8
4.9	3.1	1.5	0.1	5.2	2.7	3.9	1.4	7.2	3.6	6.1	2.5
5.4	3.7	1.5	0.2	5.0	2.0	3.5	1.0	6.5	3.2	5.1	2.0
4.8	3.4	1.6	0.2	5.9	3.0	4.2	1.5	6.4	2.7	5.3	1.9
4.8	3.0	1.4	0.1	6.0	2.2	4.0	1.0	6.8	3.0	5.5	2.1
4.3	3.0	1.1	0.1	6.1	2.9	4.7	1.4	5.7	2.5	5.0	2.0
5.8	4.0	1.2	0.2	5.6	2.9	3.6	1.3	5.8	2.8	5.1	2.4
5.7	4.4	1.5	0.4	6.7	3.1	4.4	1.4	6.4	3.2	5.3	2.3
5.4	3.9	1.3	0.4	5.6	3.0	4.5	1.5	6.5	3.0	5.5	1.8
5.1	3.5	1.4	0.3	5.8	2.7	4.1	1.0	7.7	3.8	6.7	2.2
5.7	3.8	1.7	0.3	6.2	2.2	4.5	1.5	7.7	2.6	6.9	2.3
5.1	3.8	1.5	0.3	5.6	2.5	3.9	1.1	6.0	2.2	5.0	1.5
5.4	3.4	1.7	0.2	5.9	3.2	4.8	1.8	6.9	3.2	5.7	2.3
5.1	3.7	1.5	0.4	6.1	2.8	4.0	1.3	5.6	2.8	4.9	2.0
4.6	3.6	1.0	0.2	6.3	2.5	4.9	1.5	7.7	2.8	6.7	2.0
5.1	3.3	1.7	0.5	6.1	2.8	4.7	1.2	6.3	2.7	4.9	1.8
4.8	3.4	1.9	0.2	6.4	2.9	4.3	1.3	6.7	3.3	5.7	2.1
5.0	3.0	1.6	0.2	6.6	3.0	4.4	1.4	7.2	3.2	6.0	1.8
5.0	3.4	1.6	0.4	6.8	2.8	4.8	1.4	6.2	2.8	4.8	1.8
5.2	3.5	1.5	0.2	6.7	3.0	5.0	1.7	6.1	3.0	4.9	1.8
5.2	3.4	1.4	0.2	6.0	2.9	4.5	1.5	6.4	2.8	5.6	2.1
4.7	3.2	1.6	0.2	5.7	2.6	3.5	1.0	7.2	3.0	5.8	1.6
4.8	3.1	1.6	0.2	5.5	2.4	3.8	1.1	7.4	2.8	6.1	1.9
5.4	3.4	1.5	0.4	5.5	2.4	3.7	1.0	7.9	3.8	6.4	2.0
5.2	4.1	1.5	0.1	5.8	2.7	3.9	1.2	6.4	2.8	5.6	2.2
5.5	4.2	1.4	0.2	6.0	2.7	5.1	1.6	6.3	2.8	5.1	1.5
4.9	3.1	1.5	0.2	5.4	3.0	4.5	1.5	6.1	2.6	5.6	1.4
5.0	3.2	1.2	0.2	6.0	3.4	4.5	1.6	7.7	3.0	6.1	2.3
5.5	3.5	1.3	0.2	6.7	3.1	4.7	1.5	6.3	3.4	5.6	2.4
4.9	3.6	1.4	0.1	6.3	2.3	4.4	1.3	6.4	3.1	5.5	1.8
4.4	3.0	1.3	0.2	5.6	3.0	4.1	1.3	6.0	3.0	4.8	1.8
5.1	3.4	1.5	0.2	5.5	2.5	4.0	1.3	6.9	3.1	5.4	2.1
5.0	3.5	1.3	0.3	5.5	2.6	4.4	1.2	6.7	3.1	5.6	2.4
4.5	2.3	1.3	0.3	6.1	3.0	4.6	1.4	6.9	3.1	5.1	2.3
4.4	3.2	1.3	0.2	5.8	2.6	4.0	1.2	5.8	2.7	5.1	1.9
5.0	3.5	1.6	0.6	5.0	2.3	3.3	1.0	6.8	3.2	5.9	2.3
5.1	3.8	1.9	0.4	5.6	2.7	4.2	1.3	6.7	3.3	5.7	2.5
4.8	3.0	1.4	0.3	5.7	3.0	4.2	1.2	6.7	3.0	5.2	2.3
5.1	3.8	1.6	0.2	5.7	2.9	4.2	1.3	6.3	2.5	5.0	1.9
4.6	3.2	1.4	0.2	6.2	2.9	4.3	1.3	6.5	3.0	5.2	2.0
5.3	3.7	1.5	0.2	5.1	2.5	3.0	1.1	6.2	3.4	5.4	2.3
5.0	3.3	1.4	0.2	5.7	2.8	4.1	1.3	5.9	3.0	5.1	1.8

x_1: Độ dài của đài hoa, x_2: Độ rộng của đài hoa,
x_3: Độ dài của cánh hoa, x_4: Độ rộng của cánh hoa.