SỬ DỤNG HÀM CỰC ĐẠI TRONG BÀI TOÁN NHẬN DẠNG

Võ Văn Tài[(1)], Tô Anh Dũng[(2)]
(1) Trường Đại học Cần Thơ
(2) Trường Đại học Khoa học Tự nhiên, ĐHQG-HCM
(Bài nhận ngày 07 tháng 04 năm 2009, hoàn chỉnh sửa chữa ngày 17 tháng 06 năm 2009)

***TÓM TẮT**: Dựa vào hàm cực đại của các hàm mật độ chúng tôi đã đưa ra một phương pháp mới rất thuận lợi cho bài toán nhận dạng trong các trường hợp khác nhau. Việc tìm hàm cực đại và tính sai số Bayes cũng được khảo sát. Hai chương trình được viết để tính toán cụ thể.*

***Từ khóa**: Hàm cực đại, hàm mật độ xác suất, nhận dạng, sai số Bayes.*

1. GIỚI THIỆU

Nhận dạng một phần tử mới thuộc tổng thể nào trong số *k* tổng thể đã cho là một hướng thống kê có rất nhiều ứng dụng trong thực tế, với nhiều lĩnh vực khác nhau: Nông nghiệp, y học, kinh tế, ... Đặc biệt với sự bùng nổ thông tin hiện nay thì những ứng dụng này ngày càng trở nên đa dạng và cần thiết hơn. Chính vì vậy, ngày càng có nhiều bài toán học nghiên cứu đến vấn đề này.

Bài toán nhận dạng được đặt ra như sau: *Từ một tập hợp gồm n phần tử mà ta biết rõ các phần tử đến từ tổng thể nào trong số k tổng thể, dựa trên n biến quan sát từ mỗi phần tử đưa ra một qui luật để khi có phần tử mới thì biết cách xếp vào tổng thể nào là thích hợp nhất.* Bài toán nhận dạng hiện đang được nhiều nhà toán học quan tâm, tuy nhiên trong việc giải quyết nó, theo sự hiểu biết của chúng tôi nhiều khía cạnh liên quan của bài toán này vẫn chưa có lời giải một cách trọn vẹn. Hiện tại có nhiều phương pháp giải quyết bài toán này trong đó phương pháp Bayes được xem có nhiều ưu điểm nhất vì nó giải quyết được bài toán cho tập dữ liệu bất kỳ và tính được xác suất sai lầm trong nhận dạng. Tuy nhiên trong thực tế tính toán theo phương pháp này còn rất nhiều khó khăn bởi việc xác định hàm mật độ xác suất, việc tính tích phân, việc xác định sai lầm...Trong bài viết này, dựa trên phương pháp Bayes chúng tôi đưa ra một phương pháp, được gọi là phương pháp hàm cực đại rất thuận lợi cho việc lập trình tính toán.

2. PHƯƠNG PHÁP HÀM CỰC ĐẠI TRONG BÀI TOÁN NHẬN DẠNG

2.1. Phương pháp Bayes

Xét hai tổng thể w_1 và w_2 với biến quan sát *x* có hàm mật độ xác suất $f_1(x)$, $f_2(x)$ tương ứng với hai tổng thể đó và xác suất tiên nghiệm q_1 và $q_2 = 1 - q_1$, khi đó một phần tử mới với biến quan sát x_0 được nhận dạng như sau:

Nếu $\frac{f_1(x_0)}{f_2(x_0)} > \frac{q_2}{q_1}$ thì xếp x_0 vào w_1, ngược lại xếp vào w_2. (1)

Khi ta không quan tâm đến xác suất tiên nghiệm hoặc $q_1 = q_2 = \frac{1}{2}$ thì (1) trở thành:

Nếu $f_1(x) > f_2(x)$) thì xếp x_0 vào w_1 ngược lại xếp vào w_2.

Trong trường hợp không quan tâm đến xác suất tiên nghiệm thì xác suất sai lầm khi phân loại phần tử vào tổng thể thứ nhất và thứ hai lần lượt là

$$d = P(w_1 \mid w_2) = \int_{R_1^n} f_2(x)dx,\ \ t = P(w_2 \mid w_1) = \int_{R_2^n} f_1(x)dx$$

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

Xác suất sai lầm trong phân loại này được xác định bởi công thức:

$$Pe_{1,2} = \min\{f_1(x), f_2(x)\} = t + d \quad (2)$$

Khi quan tâm đến xác suất tiên nghiệm q của w_1 thì τ trở thành τ^* và δ trở thành δ^* với:

$$\tau^* = \int_{R_{1n}^*} qf_1(x)dx \text{ và } \delta^* = \int_{R_{2n}^*} (1-q) f_2(x)dx$$

Trong đó $R_{1n}^* = \{x \mid qf_1(x) \geq (1-q) f_2(x)\}$, $R_{2n}^* = \{x \mid qf_1(x) < (1-q) f_2(x)\}$.

Đặt $(q) = (q, 1-q)$, sai số Bayes lúc này là:

$$Pe_{1,2}^{(q)} = \int_{R^n} \min\{qf_1(x), (1-q) f_2(x)\} = t^* + d^* \quad (3)$$

Xét k tổng thể w_i với xác suất tiên nghiệm q_i. Đặt $(q) = (q_1, q_2, ..., q_k)$, khi đó phần tử với biến quan sát x_0 được xếp vào w_i nếu:

$$q_i f_i(x_0) > q_j f_j(x_0) \Leftrightarrow \frac{f_i(x_0)}{f_j(x_0)} > \frac{q_j}{q_i},\ \forall j \neq i \quad (4)$$

Xác suất sai lầm trong nhận dạng này là

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

Trong đó R_i^n là miền mà phần tử mới được xếp vào tổng thể thứ i, $R^n = \bigcup_{i=1}^{k} R_i^n$.

Xác suất sai lầm được tính bởi (2), (3) và (5) đã được chứng minh là xác suất sai lầm nhỏ nhất trong nhận dạng và được gọi là sai số Bayes.

2.2. Phương pháp hàm cực đại

Dựa trên phương pháp Bayes, chúng tôi đề nghị một nguyên tắc nhận dạng phần tử mới x_0 cho k tổng thể với hàm mật độ xác suất $f_i(x)$ và xác suất tiên nghiệm q_i, $\sum_{i=1}^{k} q_i = 1$ như sau:

Nếu $g_{\max}(x_0) = q_j f_j(x_0)$ *thì phân loại* x_0 *vào* w_j. (6)

Trong đó $g_{\max}(x) = \max\{g_1(x), g_2(x), ..., g_k(x)\}$.

Phương pháp nhận dạng trên được gọi là phương pháp hàm cực đại. Phương pháp này vừa đơn giản vừa tổng quát, đặc biệt hiệu quả hơn trong tính toán so với những nguyên tắc đã có. Với nguyên tắc này việc nhận dạng phần tử mới chỉ là vấn đề tìm hàm cực đại của các hàm số

$q_j f_j(x)$, tương đương với những nguyên tắc Bayes bởi vì việc xác định những miền khác nhau cho mục đích nhận dạng của phương pháp Bayes cũng giống như việc xác định những miền khác nhau của định nghĩa $g_{\max}(x)$. Thật vậy, với trường hợp hai tổng thể, những miền khác nhau của R^n nơi $g_{\max}(x)$ nhận giá trị $qf_1(x)$ hoặc $(1-q)f_2(x)$ chính là việc giải bất phương trình $\frac{qf_1(x)}{(1-q)f_2(x)} \geq 1$, hoàn toàn giống như phương pháp Bayes. Trong trường hợp hai tổng thể có phân phối chuẩn, biên nhận dạng cho phương pháp hàm cực đại và phương pháp Bayes đều là tuyến tính hoặc bậc hai. Tương tự cho trường hợp hai tổng thể, khi có nhiều hơn hai tổng thể việc xác định những miền nơi hàm cực đại của các hàm mật độ xác suất $g_{\max}(x)$ nhận giá trị là tương đương miền mà $\frac{q_i f_i(x)}{q_j f_j(x)} \geq 1 \ \forall j \neq i$. Phương pháp Bayes xếp phần tử mới vào tổng thể w_j cũng dựa vào bất đẳng thức này.

Khi ta không quan tâm đến xác suất tiên nghiệm hoặc xác suất tiên nghiệm bằng nhau cho các tổng thể thì nguyên tắc nhận dạng phần tử mới x_0 của (1) trở thành:

Nếu $f_{\max}(x_0) = f_j(x_0)$ *thì phân loại* x_0 *vào* w_j. (7)

Tương tự, khi quan tâm đến xác suất tiên nghiệm, trường hợp này phương pháp hàm cực đại cũng tương đương với phương pháp Bayes.

2.3. Sai số Bayes trong phương pháp hàm cực đại

Giả sử hai tổng thể với hàm mật độ xác suất $f_i(x), i = 1, 2$. Khi không quan tâm đến xác suất tiên nghiệm thì sai số Bayes cho bài toán phân loại và nhận dạng được xác định bởi công thức:

$$Pe_{1,2} = \min\{f_1(x), f_2(x)\} = 2 - \int_{R^n} f_{\max}(x)dx \quad (8)$$

• Xét hai tổng thể có phân phối chuẩn một chiều $N(m_i, s_i^2)$, $i = 1, 2$. Giả sử $m_1 < m_2$.

Nếu $s_1 = s_2$ thì

$$Pe_{1,2} = 2 - \int_{-\infty}^{x_1} f_1(x)dx - \int_{x_1}^{+\infty} f_2(x)dx = 1 + j\left(\frac{x_1 - m_1}{s_1}\right) - j\left(\frac{x_1 - m_2}{s_2}\right)$$

Trong đó $x_1 = \frac{m_1 + m_2}{2}$ và $j(x) = \frac{1}{\sqrt{2p}} \int_0^x e^{-t^2/2}dt$. (9)

Nếu $s_1 \neq s_2$ thì

$$Pe_{1,2} = 2 - \int_{-\infty}^{x_2} f_2(x)dx - \int_{x_3}^{+\infty} f_2(x)dx - \int_{x_2}^{x_3} f_1(x)dx$$

$$= 1 + j\left(\frac{x_2 - m_2}{s_2}\right) - j\left(\frac{x_3 - m_2}{s_2}\right) + j\left(\frac{x_3 - m_1}{s_1}\right) - j\left(\frac{x_2 - m_1}{s_1}\right)$$

Trong đó $x_2 = \dfrac{(m_1 s_2^2 - m_2 s_1^2) - s_1 s_2 \sqrt{(m_1 - m_2)^2 + K}}{s_2^2 - s_1^2}$, (10)

$K = 2(s_2^2 - s_1^2)\ln\left(\dfrac{s_2}{s_1}\right) \geq 0$, $x_3 = \dfrac{(m_1 s_2^2 - m_2 s_1^2) + s_1 s_2 \sqrt{(m_1 - m_2)^2 + K}}{s_2^2 - s_1^2}$ (11)

Đặc biệt khi $m_1 = m_2 = m$.

Nếu $s_1 = s_2$ thì $Pe_{1,2} = 1$.

Nếu $s_1 \neq s_2$ thì

$$Pe_{1,2} = 2 - \int_{-\infty}^{x_4} f_2(x)dx - \int_{x_5}^{+\infty} f_2(x)dx - \int_{x_4}^{x_5} f_1(x)dx$$

$$= 1 + j\left(\frac{x_4 - m}{s_2}\right) - j\left(\frac{x_5 - m}{s_2}\right) + j\left(\frac{x_5 - m}{s_1}\right) - j\left(\frac{x_4 - m}{s_1}\right)$$

Trong đó $x_4 = m - s_1 s_2 \sqrt{E}$, $x_5 = m + s_1 s_2 \sqrt{E}$ với $E = \dfrac{2}{s_2^2 - s_1^2}\left[\ln\left(\dfrac{s_1}{s_2}\right)\right] \geq 0$ (12)

• Xét hai tổng thể của biến *X* có phân phối chuẩn *n* chiều: $N(m_1, \Sigma_1)$ và $N(m_2, \Sigma_2)$.

Giả sử $\Sigma_1 = \Sigma_2 = \Sigma$. Đặt:

$$U = X^T \Sigma^{-1}(m_1 - m_2) - \frac{1}{2}(m_1 - m_2)^T \Sigma^{-1}(m_1 - m_2)$$

Theo Anderson (1984) nếu *X* có phân phối chuẩn $N(m_1, \Sigma)$ thì *U* cũng có phân phối chuẩn $N\left(\frac{1}{2}q^2, q^2\right)$ với $q = \frac{1}{2}(m_1 - m_2)^T \Sigma^{-1}(m_1 - m_2)$. Tương tự nếu *X* có phân phối chuẩn $N(m_2, \Sigma)$ thì *U* cũng có phân phối chuẩn $N\left(-\frac{1}{2}q^2, q^2\right)$. Khi đó nếu không quan tâm đến xác suất tiên nghiệm thì sai số Bayes được xác định $Pe_{1,2} = t + d$ với

$$t = \frac{1}{q\sqrt{2p}} \int_0^{+\infty} \exp\left(-\frac{1}{2q}\left(x + \tfrac{1}{2}q^2\right)^2\right)dx = \frac{1}{\sqrt{2p}} \int_{q/2}^{+\infty} \exp\left(-\frac{1}{2}x^2\right)dx$$

là xác suất sai lầm khi phân loại vào tổng thể thứ nhất, còn

$$d = \frac{1}{q\sqrt{2p}} \int_{-\infty}^{0} \exp\left(-\frac{1}{2q}\left(x - \tfrac{1}{2}q^2\right)^2\right)dx = \frac{1}{\sqrt{2p}} \int_{-\infty}^{-q/2} \exp\left(-\frac{1}{2}x^2\right)dx$$

là xác suất sai lầm khi phân loại vào tổng thể thứ hai.

Khi $\Sigma_1 \neq \Sigma_2$ việc tìm một biểu thức giải tích cho t và d là rất phức tạp và gần như không có ý nghĩa cho việc tính toán cụ thể.

Xét k tổng thể với hàm mật độ xác suất $f_i(x)$ và xác suất tiên nghiệm q_i, i = 1, 2,…, k. Đặt $(q)=(q_1,q_2,...,q_k)$, khi đó sai số Bayes cho bài toán phân loại và nhận dạng được xác định:

$$Pe^{(q)}_{1,2...,k}=\sum_{i<j}\int_{R_i^n\cup R_j^n}\min\{q_i f_i, q_j f_j\}dx=\sum_{j=1}^{k}\sum_{j\neq i}\int_{R_j^n}\min\{q_i f_i, q_j f_j\}dx$$

$$=\sum_{j=1}^{k}\left[\int_{R^n}q_j f_j-\int_{R_j^n}\max\{q_j f_j\}\right]$$

$$=\int_{R^n}\sum_{i=1}^{k}q_j f_j dx-\sum_{j=1}^{k}\int_{R_j^n}\max\{q_j f_j\}dx=1-\int_{R^n}g_{\max}dx$$

Như vậy sai số Bayes được tính thông qua hàm cực đại $g_{\max}(x)$ bởi công thức đơn giản sau:

$$Pe^{(q)}_{1,2,...,k}=1-\int_{R^n}g_{\max}(x)dx \qquad (13)$$

Sai số Bayes với xác suất tiên nghiệm $q_i=\frac{1}{k}$ là

$$Pe^{(1/k)}_{1,2,...,k}=1-\frac{1}{k}\int_{R^n}f_{\max}(x)dx \qquad (14)$$

Việc sử dụng (13) hoặc (14) để tính sai số Bayes cho một thuận lợi rất lớn, đặc biệt trong việc sử dụng các phần mềm toán học để lập trình.

2.4. Hàm cực đại của các hàm mật độ xác suất

Khi biết được hàm mật độ xác suất của các tổng thể thì phương pháp hàm cực đại được xem là sự giải quyết trọn vẹn bài nhận dạng nếu chúng ta xác định được hàm cực đại của các hàm mật độ xác suất. Vì vậy trong phần này chúng ta tập trung tìm hàm cực đại của các hàm mật độ xác suất, đặc biệt các hàm mật độ xác suất thông dụng.

2.4.1. Trường hợp hai hàm mật độ xác suất

Xét hai tổng thể w_1 và w_2 có hàm mật độ xác suất một chiều hoặc nhiều chiều $f_1(x)$ và $f_2(x)$ với xác suất tiên nghiệm tương ứng q và 1– q.

Biên cho sự nhận dạng là $d^{(q)}(x)=qf_1(x)-(1-q)f_2(x)$, lúc này hàm cực đại được xác định:

$$g^{(q)}_{\max}(x)=\begin{cases}qf_1(x) & \text{khi } d^{(q)}(x)\geq 0\\ (1-q)f_2(x) & \text{khi } d^{(q)}(x)<0\end{cases}$$

Khi không quan tâm đến xác suất tiên nghiệm thì biên phân loại trở thành $d(x)=f_1(x)-f_2(x)$. Khi đó hàm cực đại được xác định:

$$f_{\max}(x) = \begin{cases} f_1(x) & \text{khi } d(x) \geq 0 \\ f_2(x) & \text{khi } d(x) < 0 \end{cases}$$

Trong trường hợp một chiều thì biên cho những miền của hàm cực đại là các điểm. Các điểm này cũng chính là ranh giới cho sự phân loại và nhận dạng. Với đa số các hàm mật độ xác suất một chiều thường chỉ có một đỉnh, nên tối đa có 2 giao điểm của hai hàm mật độ xác suất. Giả sử $qf_1(x)$ và $(1-q)f_2(x)$ giao nhau tại một điểm với tọa độ a^* và

$$g_{\max}^{(q)}(x) = \begin{cases} qf_1(x) & \text{khi } x \geq a^* \\ (1-q)f_2(x) & \text{khi } x < a^* \end{cases}$$

Tùy theo giá trị của q mà a^* có thể được xác định, nhưng tổng quát thật không dễ để tìm mối quan hệ giữa a^* và a - giao điểm của $f_1(x)$ và $f_2(x)$.

Trong việc tìm hàm cực đại của các hàm mật độ xác suất một chiều, ngoài phân phối chuẩn, chúng tôi cũng đã đưa ra những kết quả cụ thể cho các trường hợp hàm mật độ xác suất thông dụng một chiều khác như phân phối Gamma, phân phối mũ và phân phối Beta. Cụ thể:

i) $f_1(x)$ và $f_2(x)$ là hàm mật độ xác suất chuẩn một chiều**:**

$$f_i(x) = \frac{1}{s_i\sqrt{2p}} \exp\left[-\frac{1}{2s_i^2}(x-m_i)^2\right], i = 1, 2$$

Trong trường hợp hai trung bình khác nhau, giả sử $m_1 < m_2$:

Nếu $s_1 = s_2 = s$ thì $f_{\max}(x) = \begin{cases} f_1(x) & \text{khi } x < x_1 \\ f_2(x) & \text{khi } x \geq x_1 \end{cases}$

Nếu $s_1 \neq s_2$ thì $f_{\max}(x) = \begin{cases} f_1(x) & \text{khi } x_2 \leq x \leq x_3 \\ f_2(x) & \text{khi } x < x_2 \cup x > x_3 \end{cases}$

Khi $m_1 = m_2$, ta có:

Nếu $s_1 = s_2$ thì $f_{\max}(x) = f_1(x) = f_2(x)$

Nếu $s_1 \neq s_2$ thì $f_{\max}(x) = \begin{cases} f_1(x) & \text{khi } x_4 \leq x \leq x_5 \\ f_2(x) & \text{khi } x < x_4 \cup x > x_5 \end{cases}$

Trong đó x_1, x_2, x_3, x_4 và x_5 được xác định bởi (9), (10), (11) và (12).

ii) $f_1(x)$ và $f_2(x)$ là hàm mật độ xác suất chuẩn n chiều ($n \geq 2$)

Đặt $$d(x) = -\frac{1}{2}x^T\left[(\Sigma_1)^{-1} - (\Sigma_2)^{-1}\right]x + \left[m_1^T(\Sigma_1)^{-1} - m_2^T(\Sigma_2)^{-1}\right]x - k \qquad (15)$$

với $$k = \frac{1}{2}\left[\ln\left(\frac{|\Sigma_1|}{|\Sigma_2|}\right) + m_1^T(\Sigma_1)^{-1}m_1 - m_2^T(\Sigma_2)^{-1}m_2\right]$$

$d(x)$ là biên phân loại của w_1 và w_2. Ta có $d(x)$ là đường bậc 2. Đặt $A = -\frac{1}{2}\left[(\Sigma_1)^{-1} - (\Sigma_2)^{-1}\right]$ thì ta có các trường hợp cụ thể của đường bậc hai:

Nếu det(A) < 0 thì $d(x)$ là hyperbol,
Nếu det(A) = 0 thì $d(x)$ là parabol,
Nếu det(A) > 0 thì $d(x)$ là elip,
ở đây

$$f_{\max}(x) = \begin{cases} f_1(x) & \text{khi } d(x) > 0 \\ f_2(x) & \text{khi } d(x) \le 0 \end{cases}$$

Trong trường hợp $\Sigma_1 = \Sigma_2 = \Sigma$ thì $d(x)$ sẽ trở thành hàm tuyến tính:

$$d(x) = (m_1 - m_2)^T (\Sigma)^{-1} x - \frac{1}{2}(m_1 - m_2)^T (\Sigma)^{-1}(m_1 + m_2) \quad (16)$$

Khi ta quan tâm đến xác suất tiên nghiệm q và 1- q của w_1 và w_2 thì hàm nhận dạng $d(x)$ của (15) và (16) lần lượt trở thành:

$$d^{(q)}(x) = -\frac{1}{2}x^T\left[(\Sigma_1)^{-1} - (\Sigma_2)^{-1}\right]x + \left[m_1^T(\Sigma_1)^{-1} - m_2^T(\Sigma_2)^{-1}\right]x - k - \ln\left(\frac{1-q}{q}\right)$$

$$d^{(q)}(x) = (m_1 - m_2)(\Sigma)^{-1} x - \frac{1}{2}(m_1 - m_2)^T(\Sigma)^{-1}(m_1 + m_2) - \ln\left(\frac{1-q}{q}\right)$$

iii) Hai hàm mật độ xác suất có phân phối mũ trên $(0, +\infty)$:

$$f_i(x) = b_i e^{-b_i x},\ i = 1, 2$$

Giả sử $b_1 > b_2$, ta có:

$$f_{\max}(x) = \begin{cases} f_1(x) & \text{khi } x \ge \dfrac{1}{b_2 - b_1}\ln\left(\dfrac{b_2}{b_1}\right) \\ f_2(x) & \text{khi } x < \dfrac{1}{b_2 - b_1}\ln\left(\dfrac{b_2}{b_1}\right) \end{cases}$$

iv) Khi hai hàm mật độ xác suất có phân phối Beta trên (0; 1):

$$f_i(x) = \frac{1}{B(a_i, b_i)} x^{a_i - 1}(1-x)^{b_i},\ i = 1, 2$$

Hàm cực đại được xác định cụ thể:

$$f_{\max}(x) = \begin{cases} f_1(x) & \text{khi } x^k - x^{k+1} \ge m \\ f_2(x) & \text{khi } x^k - x^{k+1} < m \end{cases}$$

Trong đó $k = \frac{a}{b}$, $m = \sqrt[b]{A} > 0$, $a = a_1 - a_2, b = b_1 - b_2, A = \frac{B(a_1, b_1)}{B(a_2, b_2)}$

Trong trường hợp đặc biệt $a_1 = b_1, a_2 = b_2$ lúc này hàm cực đại trở thành:

$$f_{\max}(x)=\begin{cases} f_1(x) & \text{khi } x\le x_6, x\ge x_7 \\ f_2(x) & \text{khi } x_6<\text{x}<\text{x}_7 \end{cases}$$

trong đó, $x_6=\dfrac{1-\sqrt{1-4m}}{2}$ và $x_7=\dfrac{1+\sqrt{1-4m}}{2}$.

2.4.2. Trường hợp nhiều hơn hai hàm mật độ xác suất

Xét *k* tổng thể w_i =1, 2,…,*k,* với hàm mật độ xác suất $f_i(x)$ và xác suất tiên nghiệm tương ứng q_i, $\sum_{i=1}^{k} q_i=1$. Đặt $(q)=(q_1,q_2,...,q_k)$, $g_i(x)=q_i f_i(x)$.

Biên cho sự phân loại của w_i và w_j là $d_{ij}^{(q)}(x)=q_i f_i(x)-q_j f_j(x)$. Trong đó $d_{ij}^{(q)}(x)>0$ là miền của w_i và ngược lại là miền của w_j. Vì vậy ta có:

$$g_{\max}^{(q)}(x)=\begin{cases} q_1 f_1(x) & \text{khi } d_{1p}^{(q)}(x)>0 \\ q_l f_l(x) & \text{khi } d_{lm}^{(q)}(x)\ge 0\cap d_{nl}^{(q)}(x)\le 0 \\ q_k f_k(x) & \text{khi } d_{qk}^{(q)}(x)<0 \end{cases}$$

Trong đó $p=2,...,k;\ q=1,...,k-1,\ l=2,...,k-1,\ m=l+1,...,n,\ n=1,...,l-1$.

Khi $f_i(x)$ là hàm mật độ xác suất chuẩn *n* chiều, thì $d_{ij}^{(q)}(x)$ có dạng cụ thể:

$$d_{ij}^{(q)}(x)=-\frac{1}{2}x^T\left[(\Sigma_i)^{-1}-(\Sigma_j)^{-1}\right]x+\left[m_i^T(\Sigma_i)^{-1}-m_j^T(\Sigma_j)^{-1}\right]x-k-\ln\left(\frac{q_j}{q_i}\right) \qquad (17)$$

với $k=\dfrac{1}{2}\left[\ln\left(\dfrac{|\Sigma_i|}{|\Sigma_j|}\right)+m_i^T(\Sigma_i)^{-1}m_i-m_j^T(\Sigma_j)^{-1}m_j\right]$.

Ở đây, $d_{ij}^{(q)}(x)$ cũng là đường bậc hai. Đường bậc hai này là hyperbol, parablol hay elip phụ

thuộc vào $-\dfrac{1}{2}\det\left[(\Sigma_i)^{-1}-(\Sigma_j)^{-1}\right]$ lớn hơn 0, bằng 0 hay nhỏ hơn 0.

Trong trường hợp các $\Sigma_i=\Sigma$ với mọi i = 1, 2, …, k thì (17) trở thành:

$$d_{ij}^{(q)}(x)=(m_i-m_j)(\Sigma)^{-1}x-\frac{1}{2}(m_i-m_j)^T(\Sigma)^{-1}(m_i+m_j)-\ln\left(\frac{q_j}{q_i}\right) \qquad (18)$$

$d_{ij}^{(q)}(x)$ lúc này là hàm tuyến tính.

Khi không quan tâm đến xác suất tiên nghiệm thì hàm nhận dạng $d_{ij}^{(q)}(x)$ của (17) và (18) trở thành:

$$d_{ij}(x) = -\frac{1}{2}x^T\left[(\Sigma_i)^{-1} - (\Sigma_j)^{-1}\right]x + \left[m_i^T(\Sigma_i)^{-1} - m_j^T(\Sigma_j)^{-1}\right]x - k$$

$$d_{ij}(x) = (m_i - m_j)(\Sigma)^{-1}x - \frac{1}{2}(m_i - m_j)^T(\Sigma)^{-1}(m_i + m_j)$$

Trong trường hợp $k > 2$, việc xác định biểu thức giải tích cụ thể $f_{\max}(x)$ cũng như $g_{\max}^{(q)}(x)$ cho các hàm mật độ xác suất rất phức tạp. Ngay cả khi xem xét cho các hàm mật độ xác suất chuẩn một chiều vấn đề này cũng không phải là đơn giản. Tuy nhiên, sử dụng các phần mềm toán học như Maple, Mattlab,…bước đầu chúng tôi đã giải quyết được khó khăn này.

3. SỬ DỤNG PHẦN MỀM TOÁN HỌC TRONG BÀI TOÁN NHẬN DẠNG

3.1. Chương trình nhận dạng phần tử mới

Sử dụng nguyên tắc (6) và (7), có thể đưa ra một thuật toán để viết một chương trình nhận dạng phần tử mới. Sau đây chúng tôi minh họa một chương trình được viết bằng phần mềm Maple nhận dạng phần tử mới khi các tổng thể có hàm mật độ xác suất cùng phân phối hai chiều.

Chương trình 1:

```
Nhandang:=proc(L::list(algebraic))
        local  n,u,v,i,d,j,t,l,B,H;n:=nops(L);
        H:={seq(unapply(L[p],x,y),p=1..n-2)};
        u:=L[n-1];v:=L[n];
        for i from 1 to n-2 do
        d[i]:=evalf(H[i](u,v));
        od;
        B:=d[1];t:=H[1](x);
        l:=f[1];[l=t];
        for j from 2 to n-1 do
        if B <d[j] then
        B:=d[j];t:=f[j];l:=H[j](x);
        fi;od;[l=t];
        end:
```

Ở đây, với k tổng thể w_i với hàm mật độ xác suất $f_i(x)$, để nhận dạng phần tử mới $x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$ ta dùng lệnh: Nhandang([f1(x), f2(x), …, fk(x), x1, x2]);

Nhập các hàm số $f_i(x)$ dưới dạng biểu thức trực tiếp trong **Nhandang([ ])** hoặc lệnh gán $f_i(x)$ bên ngoài. Chương trình này dễ dàng thay đổi cho các trường hợp khác nhau của hàm mật độ xác suất một chiều hoặc nhiều chiều. Khi quan tâm đến xác suất tiên nghiệm thì $f_i(x)$ sẽ được thay thế bởi các $g_i(x)$ trong chương trình.

3.2. Chương trình tìm hàm cực đại và tính sai số Bayes

3.2.1. Phân phối một chiều

Xét k hàm số một chiều $g_1(x), g_2(x), ..., g_k(x)$, trong đó $g_i(x) = q_i f_i(x)$, $f_i(x)$ là hàm mật độ xác suất một chiều. Chúng tôi đã đưa ra một thuật toán cụ thể để tìm hàm cực đại $g_{\max}(x)$ và tính sai số Bayes khi nhân dạng. Tuy nhiên do hạn chế của số trang trình bày nên bài viết chỉ trình bày chương trình cụ thể trên phần mềm Maple dựa trên thuật toán đó để tìm $g_{\max}(x)$ và $Pe^{(q)}_{1,2,...,k}$.

Chương trình 2:

```
saiso:=proc(L::list(algebraic))
 local  e,i,j,k,r,s,t,m,n,p,kq,A,C,D,E,F,G,H,S,S1;
 n:=nops(L);
 H:={seq(unapply(L[p],x),p=1..n)};
 A:={seq(H[p],p=1..n)};
 S1:={solve(H[1](x)–H[2](x)=0,x)};
 if nop(H)=2 and nop(S1) = 1 then  e:=S1–0.001;
 if evalf(H[1](f))>evalf(H[1](f)) then
 p[x]:=piecewise(x<S1,H[1](x))
 else p[x]:=piecewise(x<S1,H[2](x));fi;
 else  m:=0;
 for i from 1 to n–1 do
 for j from i+1 to n do
 S:={solve(H[i](x)–H[j](x)=0,x)};
 C:=A minus {H[i],H[j]};
 for k from 1 to nops(S) do
 if max(seq(evalf(C[j](S[k]),25),j=1..nops(C)))<=evalf(H[i](S[k]),25)
 then m:=m+1; D[m]:=S[k]; fi; od; od; od;
 E:=sort([seq(D[p],p=1..m)]);
 F:=[E[1]–1,seq((E[i+1]+E[i])/2,i=1..m–1),E[m]+1];
 kq:=[];
 for r from 1 to nops(F) do
 for s from 1 to n do
 if H[s](F[r])=max(seq(H[p](F[r]),p=1..n)) then
 kq:=[op(kq),H[s]]; fi;od;od;
 p[1]:=piecewise(x<E[1],kq[1](x)):
 for t from 2 to m do
 p[t]:=piecewise(E[t–1]<=x and x<=E[t],kq[t](x),p[t–1]): od:
        unapply(piecewise(x>E[m],kq[m+1](x),p[m]),x);                    fi;
        K:=unapply(piecewise(x>E[m],kq[m+1](x),p[m]),x);
        evalf[5](1-int(K(x),x=-infinity..+infinity));
        end proc:
```

Ở đây,

i) Để tìm sai số Bayes khi phân loại k tổng thể có hàm mật độ xác suất $f_i(x)$, xác suất tiên nghiệm q_i, $g_i(x) = q_i f_i(x)$ ta sử dụng lệnh: **saiso([$g_1(x)$, $g_2(x)$, …, $g_k(x)$]);**

Nhập các hàm số $g_i(x)$ dưới dạng biểu thức trực tiếp trong **saiso ([ ])** hoặc lệnh gán $g_i(x)$ bên ngoài.

ii) Nếu bỏ dòng cuối của chương trình trước *end proc* thì kết quả xuất ra là một hàm số. Hàm này chính là hàm cực đại của các hàm đã cho. Chúng ta có thể đưa chúng vào trong thư viện chương trình của Maple để sử dụng vào các mục đích khác như vẽ đồ thị, tính tích phân…

iii) Đối với các hàm mật độ xác suất chỉ nhận biểu thức trong khoảng (a, b) như hàm mũ, Gamma và Beta thì lệnh giải phương trình tổng quát đổi thành lệnh giải phương trình có điều kiện, nghĩa là lệnh **solve** được thay thế bằng lệnh **fsolve** trong khoảng **(0,+∞)** đối với hàm mũ, Gamma và trong khoảng **(0, 1)** đối với hàm Beta…

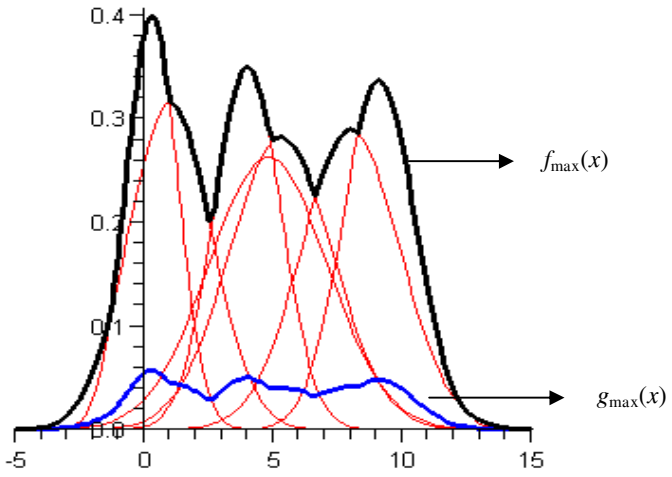
Ví dụ 1. Xét 7 hàm mật độ xác suất có phân phối chuẩn một chiều $N(m_i, s_i^2)$ với các tham số cụ thể:

$m_1 = 0.3, m_2 = 4.0, m_3 = 9.1, m_4 = 1.9, m_5 = 5.3, m_6 = 8, m_7 = 4.8$

$s_1 = 1.0, s_2 = 1.3, s_3 = 1.4, s_4 = 1.6, s_5 = 2, s_6 = 1.9, s_7 = 2.3$

Sử dụng **chương trình 2** đã viết ta có hàm cực đại $f_{\max}(x)$ được viết lại tóm tắt như sau:

$$f_{\max}(x) = \begin{cases} f_1 & \text{khi} \quad -1.2831 < x \le 0.9856 \\ f_2 & \text{khi} \quad 2.5835 < x \le 4.8932 \\ f_3 & \text{khi} \quad 8.2961 < x \le 12.5172 \\ f_4 & \text{khi} \quad -7.8585 < x \le -1.2831 \cup 0.9856 < x \le 2.5835 \\ f_5 & \text{khi} \quad 4.8932 < x \le 6.6485 \\ f_6 & \text{khi} \quad 6.6485 < x \le 8.2961 \cup 12.5171 < x \le 23.3294 \\ f_7 & \text{khi} \quad x \le -7.8585 \cup x > 23.3294 \end{cases}$$



Hình 1. Đồ thị của 7 hàm mật độ xác suất, $f_{\max}(x)$ và $g_{\max}(x)$

Giả sử có 1 phần tử mới với biến quan sát $x_0 = 10$. Áp dụng **chương trình 1** đã viết ta có kết quả:

$$f_3 = \frac{0.0853734721\ 0}{\sqrt{p}} e^{-0.2551020408\ (x-9.1)^2}$$

Nghĩa là phần tử mới này được xếp vào tổng thể thứ 3.

Nếu xác suất tiên nghiệm $q_i = \frac{1}{7}, i = 1,2,..,7$ ta lần lượt có các kết quả:

$$g_{\max}(x) = \begin{cases} g_1 & \text{khi} \quad -1.2831 < x \le 0.9856 \\ g_2 & \text{khi} \quad 2.5835 < x \le 4.8932 \\ g_3 & \text{khi} \quad 8.2961 < x \le 12.5172 \\ g_4 & \text{khi} \quad -7.8585 < x \le -1.2831 \cup 0.9856 < x \le 2.5835 \\ g_5 & \text{khi} \quad 4.8932 < x \le 6.6485 \\ g_6 & \text{khi} \quad 6.6485 < x \le 8.2961 \cup 12.5171 < x \le 23.3294 \\ g_7 & \text{khi} \quad x \le -7.8585 \cup x > 23.3294 \end{cases}$$

$Pe^{(1/7)}_{1,2,...,7} = 0.4722$. Phần tử mới cũng được xếp vào tổng thể thứ ba.

3.2.2. Phân phối chuẩn hai chiều

Khi xét k hàm số trên không gian R^n: $g_1(x), g_2(x),..., g_k(x)$, hàm cực đại $g^{(q)}_{\max}(x) = \max\{g_1(x), g_2(x),..., g_k(x)\}$ xác định trên những miền của R^{n-1} với các biên là $d^{(q)}_{ij}(x) = g_i(x) - g_j(x)$. Đặt $A_{ij} = -\frac{1}{2}\left[(\Sigma_i)^{-1} - (\Sigma_j)^{-1}\right]$, tùy theo giá trị của $\det(A_{ij})$ nhỏ hơn 0, bằng 0 hay lớn hơn 0 mà $d^{(q)}_{ij}(x)$ là hyperbol, parabol hay elip. Khi $k = 2$ chúng tôi đã viết được chương trình cụ thể để tìm $g^{(q)}_{\max}(x)$ cũng như $f_{\max}(x)$. Tuy nhiên khi $k > 3$ việc viết một chương trình trên các phần mềm toán học hiện tại để tìm hàm cực đại cũng như tính sai số Bayes là vô cùng phức tạp. Chúng tôi sẽ tiếp tục nghiên cứu vấn đề này trong thời gian tới. Hiện tại với những hàm số cụ thể cho trước ta có thể xác định hàm $g^{(q)}_{\max}(x)$, dùng phương pháp Moncte Carlo để tính tích phân, từ đó tìm được sai số Bayes.

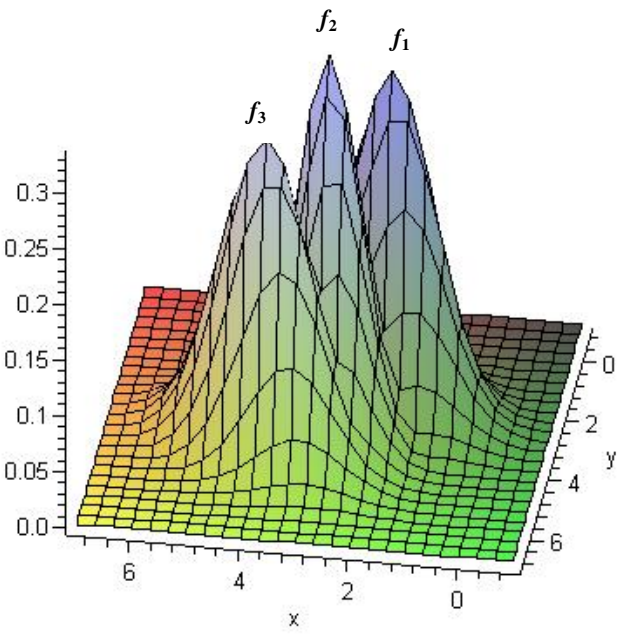
Ví dụ 2. Cho 3 tổng thể w_1, w_2 và w_3 có phân phối chuẩn hai chiều với các tham số cụ thể như sau:

$$\Sigma_1 = \begin{bmatrix} 0.706 & -0.251 \\ -0.251 & 0.507 \end{bmatrix},\ m_1 = \begin{bmatrix} 2 \\ 2 \end{bmatrix},\quad \Sigma_2 = \begin{bmatrix} 0.792 & -0.298 \\ -0.298 & 0.507 \end{bmatrix}$$

$$m_2 = \begin{bmatrix} 4 \\ 4 \end{bmatrix},\ \Sigma_3 = \begin{bmatrix} 0.397 & -0.200 \\ -0.200 & 0.706 \end{bmatrix},\ m_3 = \begin{bmatrix} 4 \\ 4 \end{bmatrix}$$

Hàm cực đại của 3 hàm mật độ xác suất được xác định cụ thể như sau:

$$f_{\max}(x,y) = \begin{cases} f_1(x,y) & \text{khi } (h_1 - y < 0 \cup h_2 - y > 0) \cap (h_3 - y > 0 \cap h_4 - y) < 0 \\ f_2(x,y) & \text{khi } (h_1 - y > 0 \cap h_2 - y < 0) \cap (h_5 - y > 0 \cap h_6 - y) < 0 \\ f_3(x,y) & \text{miền còn lại} \end{cases}$$



Hình 2. Đồ thị của 3 hàm mật độ xác suất và $f_{\max}(x)$

Trong đó:

$$h_1 = -0.0421x - 1.0956 + 1.2787.10^{-10}\sqrt{9.5067.10^{18}x^2 - 9.54027.10^{19}x + 2.61776.10^{21}}$$

$$h_2 = -0.0421x - 1.0956 - 1.2787.10^{-10}\sqrt{9.5067.10^{18}x^2 - 9.54027.10^{19}x + 2.61776.10^{21}}$$

$$h_3 = -0.7292x + 52.2358 + 6.8626.10^{-10}\sqrt{2.5348.10^{18}x^2 - 9.5629.10^{18}x + 4.7005.10^{21}}$$

$$h_4 = -0.7292x + 52.2358 - 6.8626.10^{-10}\sqrt{2.5348.10^{18}x^2 - 9.5629.10^{18}x + 4.7005.10^{21}}$$

$$h_5 = -0.1500x + 7.2805 + 1.0778.10^{-10}\sqrt{1.2354.10^{20}x^2 - 3.5745.10^{20}x + 6.2027.10^{20}}$$

$$h_6 = -0.1500x + 7.2805 + 1.0778.10^{-10}\sqrt{1.2354.10^{20}x^2 - 3.5745.10^{20}x + 6.2027.10^{20}}$$

Nếu có một phần tử mới $x_0 = \begin{bmatrix} 3.5 \\ 4.0 \end{bmatrix}$ cần xếp vào tổng thể nào là thích hợp nhất?

Với **chương trình 1** đã viết ta có kết quả:

$$f_3 = \frac{0.87676}{p}\exp\left(-0.81445(x-4)^2 - 0.97257(x-4)(y-4) - 1.29064(y-4)^2\right)$$

Nghĩa là phần tử mới được xếp vào tổng thể thứ ba.

4. KẾT LUẬN

Hàm cực đại của các hàm mật độ xác suất đã tạo ra một công cụ rất hiệu quả cho bài toán nhận dạng. Khi xem xét các tổng thể có biến quan sát một chiều được biết, bài toán nhận dạng gần như đã được giải quyết trọn vẹn bởi vì với một phần tử mới theo phương pháp hàm cực đại có thể nhận dạng nó một cách dễ dàng và tính được xác suất sai lầm trong nhận dạng đó. Với biến quan sát nhiều chiều việc nhận dạng phần tử mới dễ dàng nhưng việc tính sai lầm còn rất nhiều khó khăn do vấn đề tính tích phân. Chúng tôi sẽ lập trình để tính sai số nhận dạng này trong bài viết tới.

USING MAXIMUM FUNCTION IN DISCRIMINATION ANALYSIS

Vo Van Tai[(1)], To Anh Dung[(2)]
(1) University of Cần Thơ
(2) University of Science, VNU-HCM

***ABSTRACT**: Using maximum function of density functions we provide the new princeple which very advantage to discriminate a element for different situations. Finding maximum function and computing Bayes error are considered. The two programs are written to compute.*

***Key words**: Maximum function, probability density function, discriminant, Bayes error.*

TÀI LIỆU THAM KHẢO

[1]. Anderson, T.W., *A introduction to multivariate statistical analysis,* Wiley, New York, (1984).

[2]. Andrew R. Webb, *Statistical Pattern Recognition*, John Wiley, London, (1999).

[3]. Morris H.Degroot, *Probability and Statistics*, Addison-Wesley, United State, (1986).

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

[5]. Pham–Gia, T., Turkkan, N. and Bekker, A., *Bounds for theBayes error in classification: A baysian approach using discriminant analysis*, Statistical Methods and Applications 16, pp. 7 – 26, (2006).

[6]. Webb, A., *Statistical Pattern Recognition*, 2nd Ed., John Wiley & Sons, New York, (2002).