# MỘT PHƯƠNG PHÁP ĐỊNH VỊ ĐỐI TƯỢNG DỰA TRÊN PHÂN LỚP CÓ GIÁM SÁT

Nguyễn Hồng Thủy, Hồ Văn Canh[1*], Lê Nhật Thăng[2]

***Tóm tắt:*** *Mục đích của bài báo là nghiên cứu và đề xuất một phương pháp dò tìm một đối tượng nào đó trên cơ sở phân lớp có giám sát. Để giải quyết bài toán đặt ra, bài báo đề xuất hai bổ đề được phát triển từ bổ đề của Lý thuyết Thông tin và Thống kê toán và đề xuất một số gợi ý được ứng dụng trong thực hành.*

**Từ khóa:** Phân lớp; Nhận dạng ngôn ngữ; Đặc trưng.

## 1. MỞ ĐẦU

Để định vị một đối tượng nào đó (người, động vật hoặc bất cứ vật thể nào) chúng ta trước hết cần biết được các đặc điểm hay gọi là đặc trưng (characteristic) của đối tượng đó như: Họ và tên, quê quán, ngày tháng năm sinh, tên thường gọi, nghề nghiệp, thường sử dụng thiết bị liên lạc gì ? v.v. Nhưng để xác định được một đối tượng, trước hết ta phải phân lớp đối tượng dựa trên đặc trưng của chúng. Bài toán phân lớp được mô tả như sau:

Cho trước một tập hợp hữu hạn $\Omega$ các đối tượng, mỗi đối tượng gồm n đặc trưng. Như vậy ta có thể coi $\Omega$ là một tập con trong không gian Euclide n-chiều $R^n$.

Giả sử trên cơ sở nào đó ta có $y \in R^n$. Hãy xác định xem có tồn tại một $x \in \Omega$ mà y = x hay không ? Ở đây, ta hiểu khái niệm "y = x" theo nghĩa xác suất.

Đây là một bài toán rất lý thú và đã có kết quả được áp dụng trong thực tiễn, đặc biệt là trong lĩnh vực An ninh Quốc gia.

## 2. MỘT SỐ KẾT QUẢ ĐÃ ĐƯỢC CÔNG BỐ

Bài toán phân lớp các đối tượng được ứng dụng nhiều trong nhiều lĩnh vực, đặc biệt là nhận dạng ngôn ngữ tự nhiên (xem [1], [2], [3], [4], [8], [14],v.v.). Các phương pháp nêu trên được các tác giả dựa trên các kỹ thuật sau đây:

- Phân lớp trên cơ sở mô hình xích Markov ẩn (HMM),
- Phân lớp dựa trên mạng Neural nhân tạo (ANN - Artificial Neural Network),
- Phân lớp dựa trên máy học vector hỗ trợ (SVMs),
- Phân lớp dựa trên mô hình Gaussian hỗn hợp (GMMs),
- Phân lớp dựa trên cơ sở lượng tử hóa vector (VQ).

Các phương pháp phân lớp nêu trên khi ứng dụng vào nhận dạng các ngôn ngữ tự nhiên đều cho kết quả đúng khoảng 65 - 85%. Hạn chế của các phương pháp phân lớp nêu trên là độ dài đầu vào để kiểm tra phải đủ lớn (độ dài đầu vào thường là trên 1000 bít). Ngay cả việc phân lớp ở [9] được cho là tối ưu nhưng mẫu đầu vào để kiểm tra cũng cần có kích cỡ là 800 bít trở lên (tương ứng 100 ký tự la tinh). Trong thực tế, nhiều trường hợp, đầu vào của thuật toán là dãy chỉ khoảng 64 bít hoặc ít hơn. Năm 2017, các tác giả Phạm Anh Phương và Quách Hải Thọ đã đề xuất một phương pháp phân lớp trên cơ sở ứng dụng lý thuyết tập mờ (Fuzzy set theory) đã cho kết quả khá (xem [15]). Trong các kết quả đã được trình bày ở trên, đáng chú ý là phân lớp theo khoảng cách và phân lớp dựa trên lượng tử hóa vector (VQ). Tuy nhiên, các phương pháp đó vẫn không xét đến sai số trong phân lớp. Phần nội dung tiếp sau của bài báo này, các tác giả đề xuất một phương pháp phân lớp có giám sát nhằm khắc phục được hai nhược điểm vừa trình bày ở trên. Đó là:

- Độ dài mẫu đầu vào ngắn (cỡ 8 ký tự la tinh trở lên), và
- Sai số của phân lớp về trung bình là cực tiểu.

## 3. NỘI DUNG CẦN GIẢI QUYẾT

Để rõ hơn vấn đề này, ta xét bài toán phân lớp tổng quát nhất như sau:

Cho một tập hợp $\Omega$ hữu hạn tùy ý. Mỗi $x \in \Omega$ được gọi là một đối tượng (object) hay về mặt toán học, x được gọi là một phần tử (element) trong tập hợp $\Omega$. Mỗi phần tử được thể hiện bởi các đặc trưng (characteristic) của nó. Như vậy, các phần tử khác nhau sẽ có các đặc trưng tương ứng không giống nhau. Để dễ dàng cho việc xây dựng phương pháp phân lớp (classification) các đối tượng của $\Omega$, ta giả sử mỗi đối tượng được mô tả bởi n đặc trưng. Như vậy, ta có thể xem $\Omega$ như là một tập hợp con trong không gian Euclide n chiều (được ký hiệu là $R^n$) tức là $\Omega \subset R^n$. Bài toán đặt ra là hãy phân hoạch $\Omega$ thành k lớp: $A_1, A_2, \ldots, A_k$ với $A_i \neq \varnothing$, i = 1, 2,…, k sao cho:

1/ $A_i \cap A_j = \varnothing$, $i \neq j$, i, j = 1, 2,…, k

2/ $\bigcup_{i=1}^{k} A_i = \Omega$

Rõ ràng là có nhiều cách phân hoạch (partition) $\Omega$ thỏa mãn các điều kiện đã nêu. Song, dù phân hoạch bằng cách nào cũng đều xảy ra hai trường hợp:

<u>Trường hợp 1</u>: Đối tượng $x \in \Omega$, thực tế là $x \in A_i$ nhưng lại gán cho $x \in A_j$, $j \neq i$.

<u>Trường hợp 2</u>: $x \in A_j$ nhưng ta lại gán cho $x \in A_i$, $i \neq j$.

Trường hợp 1 xảy ra thì ta nói đã mắc sai lầm loại 1, trường hợp 2 xảy ra thì ta đã mắc sai lầm loại 2. Xác suất mắc sai lầm loại 1 ta ký hiệu là $\alpha$ $(0 \leq \alpha \leq 1)$ và xác suất mắc phải sai lầm loại 2 được ký hiệu là $\beta$ $(0 \leq \beta \leq 1)$. $\alpha$ là xác suất bác bỏ giả thiết đúng còn $\beta$ là xác suất chấp nhận giả thiết sai. Dù với thuật toán phân lớp nào cũng không thể triệt tiêu được cả hai loại sai lầm nêu trên. Trong thực tế người ta muốn cố định xác suất sai lầm loại 1, $\alpha$ và xây dựng thuật toán làm cực tiểu hóa sai lầm loại 2, $\beta$. Như vậy một thuật toán được cho là tối ưu là thuật toán làm cho tổn thất trung bình của cả hai sai lầm là bé nhất có thể.

Bổ đề sau đây nhằm giải quyết bài toán đặt ra:

Trước hết ta ký hiệu $z_{ij}$ là tổn thất khi đối tượng x thực tế là $x \in A_i$ nhưng ta lại quyết định $x \in A_j$, $j \neq i$. Rõ ràng rằng $z_{ii} = 0$ $\forall$ i = 1, 2,…, k (k là số lớp). Trái lại, $z_{ij} > 0$ với $i \neq j$.

Nếu đối tượng $x \in A_i$ thì tổn thất trung bình có điều kiện với $x \in A_i$ là:

$$L_i = \sum_{j=1}^{k} z_{ij} \int_{A_j} f_i(x) d\mu(x) \qquad (1)$$

Trong đó $\{f_i(x)\}_{i=\overline{1,k}}$ là k hàm mật độ xác suất của họ phân bố chuẩn $N(\mu_i, \nu_i)$ với i = 1, 2,…, k. (ở đây μ là độ đo σ- hữu hạn trên không gian các tập con của $\Omega$).

Tiếp theo, ta ký hiệu $\pi_i$ là xác suất để đối tượng $x \in A_i$, tức là $\pi_i = P\{x \in A_i\}$ và giả thiết $\pi_i > 0$, i = 1, 2,…, k.

Do đó giá trị trung bình không điều kiện của tổn thất khi phân lớp $\Omega = \bigcup_{i=1}^{k} A_i$ là:

$$L = \sum_{i=1}^{k} \pi_i L_i \tag{2}$$

Đặt $q_j(x) = \sum_{i=1}^{k} \pi_i z_{ij} f_i(x)$, j = 1, 2, …, k. (3)

Từ (1), (2) và (3) suy ra rằng:

$$L = \sum_{j=1}^{k} \int_{A_j} q_j(x) d\mu(x) \equiv L(A_1, A_2, \ldots, A_k) \tag{4}$$

Từ đó, bài toán đặt ra là: Hãy xây dựng một phân hoạch $A_1^*, A_2^*, \ldots, A_k^*$ sao cho cực tiểu hóa giá trị L:

$L^* = L(A_1^*, A_2^*, \ldots, A_k^*)$, tức là $L^* \leq L$.

Ta có bổ đề sau đây:

<u>Bổ đề 1</u>: Giả sử $A_1^*, A_2^*, \ldots, A_k^*$ là một phân hoạch trên tập $\Omega$ thỏa mãn điều kiện: $[x \in A_i^*] \Rightarrow [q_i(x) \leq q_j(x), j = 1, 2, \ldots, k]$. Khi đó, $L^* = L(A_1^*, A_2^*, \ldots, A_k^*) \leq L = L(A_1, A_2, \ldots, A_k)$ đối với mọi phân hoạch $A_1, A_2, \ldots, A_k$ tùy ý trên $\Omega$.

Chứng minh

Thật vậy, từ giả thiết trên ta có:

$$L = L(A_1, A_2, \ldots, A_k) = \sum_{j=1}^{k} \int_{A_j} q_j(x) d\mu(x) = \sum_{j=1}^{k} \sum_{i=1}^{k} \int_{A_j \cap A_i^*} q_j(x) d\mu(x)$$

$$= \sum_{i=1}^{k} \sum_{j=1}^{k} \int_{A_i^* \cap A_j} q_j(x) d\mu(x) \geq \sum_{i=1}^{k} \sum_{j=1}^{k} \int_{A_i^* \cap A_j} q_i(x) d\mu(x) = \sum_{i=1}^{k} \int_{A_i^*} q_i(x) d\mu(x)$$

$= L(A_1^*, A_2^*, \ldots, A_k^*) = L^*$. Đây là điều phải chứng minh.

Chú ý: Để đơn giản trong thực hành ta giả thiết:

$$z_{ij} = \begin{cases} 0 & \text{nếu } i=j \\ 1 & \text{nếu } i \neq j \end{cases} \tag{5}$$

Và đặt:

$$c(x) = \sum_{i=1}^{k} \pi_i f_i(x), \tag{6}$$

Từ (3), (5) và (6) ta suy ra:

$$q_j(x) = c(x) - \pi_j f_j(x) \tag{7}$$

Từ đó, $q_t(x) \leq q_j(x)$ $\forall$ j = 1, 2,…, k nếu và chỉ nếu:

$$\pi_t f_t(x) \geq \pi_j f_j(x) \ \forall \ j = 1, 2, \ldots, k \tag{8}$$

Như vậy nếu tồn tại một $t \neq j$ mà

$$\pi_t f_t(x) > \pi_j f_j(x) \ \forall \ j = 1, 2, \ldots, k \tag{8'}$$

thì quyết định của ta về việc $x \in A_t$ là tối ưu.

Trường hợp tồn tại $t_1 \neq t_2$, $t_1 \neq j$, $t_2 \neq j$, j = 1, 2,…, k mà

$\pi_{t_1} f_{t_1}(x) \geq \pi_j f_j(x)$ và đồng thời $\pi_{t_2} f_{t_2}(x) \geq \pi_j f_j(x)$ $\forall j \neq t_1, t_2$.

Khi đó ta sử dụng quy tắc: $x \in A_{t_1}$ nếu $t_1 > t_2$ và $x \in A_{t_2}$ nếu $t_2 > t_1$.

Bây giờ, giả sử cho một tập hữu hạn $\Omega$ đã được phân hoạch tối ưu (theo nghĩa nêu trên). Để đơn giản phân hoạch đó được ký hiệu là $A_1, A_2, \ldots, A_k$, $k \geq 2$ và cho trước. Giả sử $f_1, f_2, \ldots, f_k$ là các hàm mật độ xác suất lần lượt trên $A_1, A_2, \ldots, A_k$.

Ta ký hiệu tập hợp G = {$f_1$, $f_2$, …, $f_k$} và h là một hàm mật độ xác suất nào đó của đại lượng ngẫu nhiên Y. Vấn đề đặt ra là hãy trả lời câu hỏi: có tồn tại một i, i = 1, 2,…, k mà $y \in A_i$ hay không ?

Sau đây là câu trả lời cho câu hỏi trên:

<u>Bổ đề 2:</u> Cho $f_1, f_2, \ldots, f_k$ là k hàm mật độ xác suất lần lượt trên $A_1, A_2, \ldots, A_k$. Trong đó {$A_1, A_2, \ldots, A_k$} là phân hoạch như trong Bổ đề 1. Giả sử X là một đại lượng ngẫu nhiên trên $\Omega$ với h là một hàm mật độ xác suất của X trên không gian $\Omega$. Khi đó:

1/ Nếu tích phân $\int_\Omega h(x) \log \frac{f_i(x)}{f_j(x)} d\mu(x) > 0$ với mọi $j \neq i$. Khi đó, $h = f_i$ $\mu$- hầu khắp nơi trên $\Omega$, đặc biệt là trên $A_i$.

2/ Nếu tồn tại một $j \neq i$ mà $\int_\Omega h(x) \log \frac{f_i(x)}{f_j(x)} d\mu(x) < 0$. Khi đó, $h \neq f_i$.

3/ Nếu $\int_\Omega h(x) \log \frac{f_i(x)}{f_j(x)} d\mu(x) = 0$. Khi đó, không có câu trả lời.

4/ Trường hợp $\int_\Omega h(x) \log \frac{f_i(x)}{f_j(x)} d\mu(x) < 0$ với mọi $j \neq i$. Khi đó, $h \notin G = \{f_1, f_2, \ldots, f_k\}$.

Chứng minh

Để chứng minh Bổ đề 2, ta sử dụng Bổ đề 3 với nội dung như sau:

<u>Bổ đề 3</u>: Giả sử f và g là hai làm số thực, không âm và khả tích đối với độ đo $\mu$ nào đó trên miền $\Omega$ và sao cho thỏa mãn điều kiện:

Tích phân $\int_\Omega (f - g) d\mu(x) \geq 0$ (9)

Khi đó tích phân $\int_\Omega f \log \frac{f}{g} d\mu(x) \geq 0$ (10)

và nó bằng 0 khi và chỉ khi f = g $\mu$- hầu khắp nơi trên $\Omega$.

Chứng minh

Ta chứng minh Bổ đề 3 cho trường hợp f và g là những hàm rời rạc.

Mệnh đề 1: Cho hai chuỗi số thực không âm và hội tụ: $\sum a_i$, $\sum b_i$ với $a_i$, $b_i \geq 0$ sao cho $(\sum a_i - \sum b_i) \geq 0$. Khi đó $\sum a_i \log \frac{a_i}{b_i} \geq 0$. (11)

Bất đẳng thức (11) chỉ bằng 0 khi và chỉ khi $a_i = b_i$ với mọi i = 1, 2, 3,…

Chứng minh

Bất đẳng thức (11) tương đương với bất đẳng thức (12) sau đây:

$$\sum_i a_i \log \frac{b_i}{a_i} \leq 0 \qquad (12)$$

Ta sẽ chứng minh (12) như sau:

Trước hết, ta xét hàm số f(x) = lnx (logarit nêpe) ($\log_a x = \log_e x = \ln x$). Bây giờ ta khai triển hàm f(x) trong lân cận V(1) = (1-$\varepsilon$, 1+$\varepsilon$).

Ta có: lnx = ln(x-1+1)=ln[(x-1)+1] = (x-1)-(x-1)$^2$(2$\xi^2$)$^{-1}$, trong đó $\xi \in$(1, x).

Vì vậy, $\sum_i a_i \log \frac{b_i}{a_i} = \sum_i a_i[(\frac{b_i}{a_i} - 1) - (\frac{b_i}{a_i} - 1)^2 (2\xi_i^2)]$ = $(\sum b_i - \sum a_i)$ - $\sum_i a_i (\frac{b_i}{a_i} - 1)^2 (2\xi_i^2)] \leq 0$ vì $\sum b_i \leq \sum a_i$. Đây là điều phải chứng minh.

Bây giờ ta chứng minh Bổ đề 2. Ta ký hiệu tập hợp G = {$f_1$, $f_2$,…, $f_k$} và h(x) là hàm mật độ xác suất của đại lượng ngẫu nhiên X trên $\Omega$. Ta giả thiết h$\in$G.

1/ Giả sử tích phân $\int_\Omega h(x) \log \frac{f_i(x)}{f_j(x)} d\mu(x) > 0$, ta cần chứng minh rằng h = $f_i$ $\mu$-hầu khắp nơi trên $\Omega$, đặc biệt là $A_i$. Thật vậy, giả sử trái lại rằng h $\neq$ $f_i$, tức là có tồn tại một l để h = $f_l$. Từ đó và từ giả thiết:

$\int_\Omega h(x) \ln \frac{f_i(x)}{f_j(x)} d\mu(x) > 0$ với mọi j$\neq$ i, j = 1, 2, …, k. Ta suy ra: $\int_\Omega f_l(\mathrm{x}) \ln \frac{f_i(x)}{f_j(x)} d\mu(x) > 0$, vì bất đẳng thức đó đúng cho mọi j nên nó cũng đúng cho j = l, tức là $\int_\Omega f_l(x) \ln \frac{f_i(x)}{f_l(x)} d\mu(x) > 0$, hay $\int_\Omega f_l(x) \ln \frac{f_l(x)}{f_i((x)} d\mu(x) < 0$. Điều này trái với kết quả của Bổ đề 3, vậy h = $f_i$ và Bổ đề được chứng minh.

2/ Hiển nhiên (suy ra từ 1).

3/ Trường hợp $\int_\Omega h(x) \log \frac{f_i(x)}{f_j(x)} d\mu(x) = 0$ thì theo Bổ đề 3 $f_i$= $f_j$ $\mu$- hầu khắp nơi trên $\Omega$, nên chúng ta không có cơ sở để kết luận h = $f_i$ hay h = $f_j$.

4/ Trường hợp $\int_{\Omega} h(x)\log\frac{f_i(x)}{f_j(x)}d\mu(x) < 0$ với mọi j $\neq$ i.

Nếu có tồn tại l mà h = $f_l$, thế thì $\int_{\Omega} f_l(x)\ln\frac{f_i(x)}{f_l(x)}d\mu(x) < 0$. Điều này trái với Bổ đề 3, vậy h $\neq$ $f_j$ với mọi j = 1, 2, …, k. Tức là h$\notin$G. Đó là điều cần chứng minh.

## 4. ỨNG DỤNG

***Ví dụ 1:*** **Xác định khóa mã DES (Data Encryption Standard)**

Giả sử ta nhận được một bản mã X = ($x_1$, $x_2$,…, $x_n$) được mã bởi DES. Vấn đề đặt ra là hãy giải bản mã đó, tức là hãy xác định đúng K = ($k_1$, $k_2$, …, $k_{56}$) với $k_i \in \{0, 1\}$ với i = 1, 2,…, 56. Trong trường hợp này, khóa đúng K sẽ thuộc vào không gian khóa $\Omega$={$K^{(1)}$, $K^{(2)}$, …, $K^{(2^{56})}$}, tức là lực lượng của $\Omega$ là $|\Omega| = 2^{56}$. Đây là một lực lượng quá lớn mà hệ thống máy tính của ta hiện nay không đủ khả năng tìm được khóa đúng trong $\Omega$ bằng phương pháp "vét cạn" trong thời gian đa thức. Đã có một số phương pháp tấn công được công bố [.]. Tuy nhiên các phương pháp đó chỉ mang tính lý thuyết và còn mang tính thành công theo nghĩa xác suất. Trong ví dụ này, chúng ta đòi hỏi:

i) Đã có một hệ thống tính toán hiệu năng cao.
ii) Có một hoặc nhiều bản mã bởi mã hóa DES.

Bước đầu tiên của thuật toán là chia không gian $\Omega$ thành 2 tập con rời nhau: $A_1$, $A_2$. Trong đó một trong hai tập hợp đó chứa khóa đúng K với xác suất 1, chẳng hạn đó là tập $A_1$. Vấn đề là phân hoạch như thế nào ?

Ta biết rằng mối quan hệ giữa bản mã, khóa và bản rõ có sự tương ứng 1-1. Đó là mỗi khối mã 64 bít tương ứng với một khối bản rõ 64 bít (8 bytes) và tương ứng 1-1 với khóa K. Số tất cả các khối rõ 64 bít (8 bytes) có thể là $26^8 \approx 1,6.2^{11}$. Như vậy khóa đúng chỉ có thể thuộc vào tập hợp $A_1$ với lực lượng $|A_1| \approx 1,6.2^{11} \leq 2^{12}$.

Đây là một con số mà với công nghệ hiệu năng cao (hoặc công nghệ cluster) hoàn toàn có thể vét cạn để xác định khóa đúng nếu chúng ta có một tiêu chuẩn bản rõ tốt.

Như vậy, không gian khóa của DES là $\Omega$ có thể phân hoạch thành 2 tập con (2 lớp): $\Omega = A_1 \cup A_2$ với $A_1 \cap A_2 = \varnothing$. Trong đó $|A_1| \approx 2^{12}$ và $|A_2| = 2^{56}/2^{12} \approx 2^{44}$.

Tiêu chuẩn bản rõ được trình bày ở Bổ đề 2 với k=2 (lớp).

Sau đây là một ví dụ đơn giản bằng số:

Cho bản mã y = AJJIT BRJHF XHXRJ JUBAR SMITR UHXVU GGQBI HE. Biết rằng đây là bản mã dùng mật mã thay thế đơn với ngôn ngữ tiếng Việt. Hãy giải bản mã đó (tức tìm khóa mã thay thế đơn).

Đây là bài toán khá đơn giản nên ta làm như sau:

1/ Phân lớp: Trong tiếng Việt (viết theo lối Telex) có 6 ký tự cao tần nhất là "T, H, A, N, O, I" ta gán cho lớp $A_1$. Số ký tự thấp tần nhất là "P, Q, K, J, X, Z" được gán cho lớp $A_2$. Các ký tự còn lại được gán cho lớp $A_3$.

2/ Tính tần số các ký tự của bản mã y, ta nhận được kết quả như sau:

| A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2 | 3 | 0 | 0 | 1 | 1 | 2 | 4 | 3 | 5 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 4 | 1 | 2 | 3 |

| V | W | X | Y | Z |
|---|---|---|---|---|
| 1 | 0 | 3 | 0 | 0 |

Nhận xét: Có 7 ký tự của bản mã y có tần số cao nhất là {B, H, I, J, R, U, X}.

Ta chú ý lớp thứ nhất $A_1$.

3/ Thực hiện tìm khóa K: Về lý thuyết ta có tất cả 7! = 1.2.3.4.5.6.7 = 840 phép thay thế để tìm khóa đúng. Nhưng thực tế với 12 vòng thay thế chúng ta đã xác định được khóa giải mã K là:

| A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| c | j | . | . | m | f | e | a | n | o | . | . | . | . | . | . | t | h | u | g | i |

| V | W | X | Y | Z |
|---|---|---|---|---|
| v | . | x | . | . |

Các dấu "." ở dưới các ký tự mã là chưa xác định được vì bản mã y quá ngắn.

3/ Bản thông báo được giải mã ra là:

x = "coongj hoaf xax hooij chur nghiax Vieetj nam"./.

***<u>Ví dụ 2</u>:***

Để làm ví dụ, trước hết ta có nhận xét sau đây: Xét hàm số f(x) =lnx với ln là logarit nêpe và x>0. Rõ ràng f(x) là hàm đồng biến nếu x>1. Trong trường hợp 3 của bổ đề 2, nếu giá trị x ≈ 1 thì ln1≈ 0 và do đó khi lấy phần nguyên ta đều nhận được giá trị phần nguyên của nó là 0. Để giảm thiểu trường hợp này, ta cần nhân giá trị của f(x) bởi một hằng số k đủ lớn trước khi lấy phần nguyên, chẳng hạn ta lấy k = 10.

Cho 2 dãy bít: $X_1$= 10011 10100 00110; $X_2$= 11000 11010 01001. Bây giờ giả sử ta nhận được dãy đặc trưng Y = 10100 01110 00001. Hãy trả lời câu hỏi Y = $X_1$ hay Y = $X_2$ ? Áp dụng bổ đề trên, ta tính

$f_1 = A_1 = \begin{bmatrix} 4 & 3 \\ 4 & 3 \end{bmatrix}$, $f_2 = A_2 = \begin{bmatrix} 4 & 4 \\ 4 & 2 \end{bmatrix}$

$[10\ln f_1 / f_2] = A_3 = \begin{bmatrix} 0 & -3 \\ 0 & 4 \end{bmatrix}$ và $h = B = \begin{bmatrix} 6 & 3 \\ 3 & 2 \end{bmatrix}$

Do đó, $S = tr(A_3.B^T) = -1 < 0$.

Vậy, Y = $X_2$. Kết quả này phù hợp với thực tế là khoảng cách hamming $d(B,X_2) = 5 < d(B,X_1) = 9$. Chú ý rằng các ma trận $A_1$, $A_2$, và B là các tần số bộ đôi móc xích tương ứng với các vectơ $X_1$, $X_2$ và Y ở trên. Còn tr(A) là vết (trace) của ma trận A và $X^T$ là ma trận chuyển vị của ma trận X.

## 5. KẾT LUẬN

Bài báo trình bày kết quả của việc giải bài toán phân lớp có giám sát (supervised classification) và với số k lớp đã cho trước. Trường hợp này đơn giản hơn bài toán phân lớp không có giám sát (non-supervised classification) với số lớp k chưa biết. Nhưng nó

phục vụ yêu cầu của bài toán định vị đối tượng đã đặt ra. Ngoài ra, bài báo mới chỉ giải quyết được đối tượng y có thuộc $A_i$ hay không, chứ chưa tìm được $x \in A_i$ mà y = x theo một nghĩa nào đó. Để giải quyết tiếp vấn đề này, chúng tôi sẽ đưa ra hai tiếp cận:

- Một là xây dựng khoảng cách d(x,y), $x \in A_i$, từ đó x = y khi và chỉ khi $d(x,y) = \min_{x \in A_i}(d(x',y))$.
- Hai là, nhờ sự hỗ trợ và phối hợp với cơ quan chuyên môn.

## TÀI LIỆU THAM KHẢO


[1]. R. Leonard, G. Doddington, "*Automatic Languge Identification*", Technical Report RADC-TR-74-200 (Rome Air Development Center) August 1975.

[2]. R. Leonard, "*Language Recognition Test and Evaluation*", Technical Report RADC-TR-80-83, March 1980.

[3]. A.S. House, E.P. Neuberg, "*Toward Automatic Identification of the Languages of an utterance. J. Acoust. Soc. Am. 62(3)*", 708-717 (1977).

[4]. J. Gauvain, A. Messaoudi, H. Schwenk, "*Language recognition using phone lattices*", In International Speech Communication Association (INTERSPEECH), pp. 25-28 (2004).

[5]. W. Shen, W. Campbell, T. Gleason, D. Reynolds, E. Singer, "*Experiments with lattice-based PPRLM Language Identification*". In Speaker and Language Recognition Workshop, pp. 1-6 (2006).

[6]. H. Li, B. Ma, C. H. Lee, "*Avector space modelling approach to spoken Language Identification*", IEEE. Trans. Audio Speech Lang. Process. 15(1), pp.271-284 (2007).

[7]. G. R. Botha, E. Barmard, "*Fators that affect the accuracy of text-based language Identification*", Comput, Speech Lang. 26(5), 307-20 (2012).

[8]. S. M. Siniscalchi, J. Reed, T. Svendsen, C. H. Lee, "*Universal attribute characterization of spoken Languages for automatic spoken Language recognition*", Comput. Speech Lang – 27(1), pp. 209-227 (2013).

[9]. V. R. Reddy, S. Maity, K. S. Rao, "*Identification of Indian Languages using multi-level spectral and prpsodic Features*", Int. J. Speech Technol (Springer) 16(4), pp.489-511 (2013).

[10]. G. L. Stuber, "*Propagation Modeling*", Principles of Mobile Communication; (2012).

[11]. Z. Sanaci et.al, "*Heterogengeneity in Mobile Cloud Computing: Taxonomy and Open Challenges*", IEEE Commun. Survey & Tutorial, vol. 16, no.1, pp.369-392, 2014.

[12]. Arch W.Nafflor – George R.Sell 1985: The "*Linear Operation Theory*", in National and Technical Science, 1985.

[13]. Hồ Văn Canh, Nguyễn Viết Thế, "*Phần 1 Nhập môn: Phân tích thông tin có bảo mật*'', Nhà xuất bản Hà Nội T&T – 2010.

[14]. K.S. Rao and D. Nandi, "*Languge Identification using Excitation*", Springer Briefs in Speech Technology, DOI 10. 1007/978_3_31q_17725_0_2, 2015.

[15]. Phạm Anh Phương, Quách Hải Thọ, " *Một phương pháp quản lý dữ liệu tham gia phân lớp trong mô hình học bán giám sát*", Kỷ yếu Hội nghị FAIR tại Đà Nẵng, 8/2017, DOI: 10. 15625/vap, 2017, 00059.

**ABSTRACT**

A METHOD DETECTING AN OBJECT BASED ON SUPERVISED CLASSIFICATION

*The aim of the paper is to study and propose a method detecting an object based on supervised classification. To resolve the problem, the paper proposes two lemmas (lemma 1 and lemma 2) developed from the lemma of Information and Statistical Theory, and makes some suggestions applied in pratise.*

**Keywords:** Classify; Language Identification; Characteristics.

*Nhận bài ngày 01 tháng 7 năm 2018*
*Hoàn thiện ngày 24 tháng 8 năm 2018*
*Chấp nhận đăng ngày 20 tháng 9 năm 2018*

*Địa chỉ:* [1] Bộ Công an;
[2] Học viện Công nghệ Bưu chính Viễn thông.
[*] Email: hovancanh@gmail.com.