MỞ ĐẦU

Cuộc cách mạng thông tin kỹ thuật số đã đem lại những thay đổi sâu sắc trong xã hội và trong cuộc sống của chúng ta. Mạng Internet toàn cầu đã biến thành một xã hội ảo nơi diễn ra quá trình trao đổi thông tin trong mọi lĩnh vực chính trị, quân sự, quốc phòng, kinh tế, thương mại… Và chính trong môi trường mở và tiện nghi như thế xuất hiện những vấn nạn, tiêu cực đang rất cần đến các giải pháp hữu hiệu cho vấn đề an toàn thông tin như nạn xuyên tạc thông tin, ăn cắp thông tin v.v... Vấn đề đặt ra là thông tin phải bảo mật vậy thông tin phải được mã hóa, mã hóa được đánh giá là tốt nhất trong bảo mật thông tin, do đó an toàn bảo mật thông tin được đặt lên hàng đầu không chỉ riêng ở Việt Nam mà cả trên thế giới. Khi thông tin mã hóa bằng hệ mã như vậy làm thế nào để xác định hệ mã đó tốt để thông tin được bảo mật an toàn? Thám mã thông tin chưa biết có thể giúp đánh giá được hệ mã là tốt hay xấu. Vậy, vấn đề của việc thám mã là ta đi tìm khóa của hệ mã đó, làm thế nào để biết khóa đó là đúng? chỉ khi khóa đúng thì thông tin đó mới đúng. Được sự gợi ý của thầy em đã tìm hiểu đề tài này. Trong bài luận văn em tập trung nghiên cứu vấn đề nhận dạng ngôn ngữ (Recognition of language) tự nhiên dựa vào phân hoạch không gian (hay nhận dạng theo thống kê toán học), trong đó một lớp ngôn ngữ tiêu biểu được nghiên cứu đó là Tiếng Anh. Em đã xây dựng một hệ mã và ứng dụng nhận dạng ngôn ngữ để tìm khóa hệ mã này.

Luận văn của em gồm có 3 chương :

Chương 1. Khái quát về nhận dạng.

Chương 2. Ứng dụng lý thuyết thống kê toán học giải bài toán nhận dạng ngôn ngữ tự nhiên và ứng dụng vào việc dò tìm khóa mã trong phân tích mật mã.

Chương 3. Xây dựng thuật toán giấu tin và sử dụng kỹ thuật nhận dạng bản rõ để dò tìm khóa.

Cuối cùng em có trình bày kết quả đạt được cũng như đánh giá hiệu quả của thuật toán. Do đây là một đề tài khó đối với em vì nó liên quan đến cơ sở toán học như xích Markov, lý thuyết xác suất thống kê, trong luận văn không tránh khỏi những sai sót. Em mong được các thầy, cô chỉ bảo để luận văn của em được đạt chất lượng tốt hơn. Em xin chân thành cảm ơn.

- **Phương pháp nghiên cứu:**
    - Nghiên cứu tài liệu (Tài liệu kỹ thuật thống kê toán học các quá trình Markov, tài liệu tổng quan về giấu thông tin trong ảnh).
    - Các quy luật ngôn ngữ như là một quá trình ngẫu nhiên dừng, không hậu quả.
- **Nội dung nghiên cứu:**
    - Tính tần số bộ đôi móc xích của ngôn ngữ Tiếng Anh
    - Nghiên cứ về thuật toán giấu thông tin trong ảnh giúp cho việc thực hiện tìm khóa.
    - Nghiên cứu cơ sở của lý thuyết sác xuất – thống kê toán học
    - Nghiên cứu, xây dựng tiêu chuẩn nhận dạng và lập trình thể hiện thuật toán trên ngôn ngữ Matlab.

CHƯƠNG 1

TỔNG QUAN VỀ NHẬN DẠNG

1.1. Tổng quan về nhận dạng

Nhận dạng (pattern recognition) là một ngành thuộc lĩnh vực học máy (machine learning). Nhận dạng nhằm mục đích phân loại dữ liệu (là các mẫu) dựa trên: hoặc là kiến thức tiên nghiệm (a priori) hoặc dựa vào thông tin thống kê được trích rút từ các mẫu có sẵn. Các mẫu cần phân loại thường được biểu diễn thành các nhóm của các dữ liệu đo đạc hay quan sát được, mỗi nhóm là một điểm ở trong một không gian đa chiều phù hợp. Đó là không gian của các đặc tính để dựa vào đó ta có thể phân loại. Quá trình nhận dạng dựa vào những mẫu học biết trước gọi là nhận dạng có ***thầy*** hay ***học có thầy*** (supervised learning); trong trường hợp ngược lại là ***học không có thầy*** (unsupervised learning).

Trong lý thuyết nhận dạng nói chung có ba cách tiếp cận khác nhau:

- Nhận dạng dựa vào phân hoạch không gian.
- Nhận dạng cấu trúc.
- Nhận dạng dựa vào kỹ thuật mạng nơ ron.

Hai cách tiếp cận đầu là các kỹ thuật kinh điển. Cách tiếp cận thứ ba hoàn toàn khác. Nó dựa vào cơ chế đoán nhân, lưu trữ và phân biệt đối tượng mô phỏng theo hoạt động của hệ thần kinh con người. Các cách tiếp cận trên sẽ trình bày trong các phần dưới đây.

Các ứng dụng phổ biến là nhận dạng tiếng nói tự động, phân loại văn bản thành nhiều loại khác nhau (ví dụ: những thư điện tử nào là spam/non-spam), nhận dạng tự động các mã bưu điện viết tay trên các bao thư, hay hệ thống nhận dạng danh tính dựa vào mặt người. Ba ví dụ cuối tạo thành lãnh vực con phân tích ảnh của **nhận dạng** với đầu vào là các ảnh số.

1.1.1. Không gian biểu diễn đối tượng, không gian diễn dịch

Không gian biểu diễn đối tượng [1]

Các đối tượng khi quan sát hay thu thập được, thường được biểu diễn bởi tập các đặc trưng hay đặc tính. Như trong trường hợp xử lý ảnh, ảnh sau khi được tăng cường để nâng cao chất lượng, phân vùng và trích chọn đặc tính được biểu diễn bởi các đặc trưng như biên, miền đồng nhất,v.v. Người ta thường phân các đặc trưng này theo các loại như: đặc trưng tôpô, đặc trưng hình học và đặc trưng chức năng. Việc biểu diễn ảnh theo đặc trưng nào phụ thuộc vào ứng dụng tiếp theo. Ở đây ta đưa ra một cách hình thức việc biểu diễn các đối tượng. Giả sử đối tượng X (ảnh, chữ viết, dấu vân tay,v.v.); được biểu diễn bởi n thành phần (n đặc trưng): $X=\{x_1,x_2,...,x_n\}$; mỗi x_i biểu diễn một đặc tính. Không gian biểu diễn đối tượng thường gọi tắt là không gian đối tượng X và được ký hiệu là:

$$X =\{X_1,X_2,...,X_n\}$$

trong đó mỗi X_i biểu diễn một đối tượng. Không gian này có thể là vô hạn. Để tiện xem xét chúng ta chỉ xét tập X là hữu hạn.

Không gian diễn dịch

Không gian diễn dịch là tập các tên gọi của đối tượng. Kết thúc quá trình nhận dạng ta xác định được tên gọi cho các đối tượng trong tập không gian đối tượng hay nói là đã nhận dạng được đối tượng. Một cách hình thức gọi Ω là tập tên đối tượng:

$\Omega=\{w_1,w_2,...,w_k\}$ với w_i, i =1,2,...,k là tên các đối tượng:

Quá trình nhận dạng đối tượng là một ánh xạ f: $X \rightarrow \Omega$ với f là tập các quy luật để định một phần tử trong X ứng với một phần tử Ω. Nếu tập các quy luật và tập tên các đối tượng là biết trước như trong nhận dạng chữ viết (có 26 lớp từ A đến Z), người ta gọi là nhận dạng có thầy. Trường hợp thứ hai là nhận dạng không có thày. Đương nhiên trong trường hợp này việc nhận dạng có khó khăn hơn.

1.1.2. Mô hình và bản chất của quá trình nhận dạng

1.1.2.1. Mô hình

Việc chọn lựa một quá trình nhận dạng có liên quan mật thiết đến kiểu mô tả mà người ta sử dụng để đặc tả đối tượng. Trong nhận dạng, người ta phân chia làm hai họ lớn: [1]

- Họ mô tả theo tham số;

- Họ mô tả theo cấu trúc.

Cách mô tả được lựa chọn sẽ xác định *mô hình* của đối tượng. Như vậy, chúng ta sẽ có hai loại mô hình: *mô hình theo tham số và mô hình cấu trúc*.

- Mô hình tham số sử dụng một vectơ để đặc tả đối tượng, mỗi phần tử của vectơ mô tả một đặc tính của đối tượng. Thí dụ như trong các đặc trưng chức năng, người ta sử dụng các hàm cơ sở trực giao để biểu diễn. Và như vậy ảnh sẽ được biểu diễn bởi một chuỗi các hàm trực giao. Giả sử C là đường bao của ảnh và C(i,j) là điểm thứ i trên đường bao, i = 1, 2, ..., N (đường bao gồm N điểm)

Giả sử tiếp:

$$x_0 = \frac{1}{N}\sum_{i=1}^{N} x_i$$

$$y_0 = \frac{1}{N}\sum_{i=1}^{N} y_i$$

là tọa độ tâm điểm. Như vậy, momen trung tâm bậc p, q của đường bao là

$$\mu_{pq} = \frac{1}{N}\sum_{i=1}^{N}(x_i - x_0)^p (y_i - y_0)^q \qquad (1.1)$$

Vectơ tham số trong trường hợp này chính là các momen μ_{ij} với i=1,2,...,p và j=1,2,...,q. Còn trong các đặc trưng hình học người ta hay sử dụng chu tuyến, đường bao, diện tích và tỉ lệ $T = 4\Pi S/p^2$, với S là diện tích, p là chu tuyến.

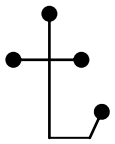
Việc lựa chọn phương pháp biểu diễn sẽ làm đơn giản cách xây dựng. Tuy nhiên, việc lựa chọn đặc trưng nào là hoàn toàn phụ thuộc vào ứng dụng. Thí dụ, trong nhận dạng chữ, các tham số là các dấu hiệu:

- Số điểm chạc ba, chạc tư,

- Số điểm chu trình,

- Số điểm ngoặt,

- Số điểm kết thúc,

Chẳng hạn với chữ t có 4 điểm kết thúc, 1 điểm chạc tư, ... 

• **Mô hình cấu trúc**: Cách tiếp cận của mô hình này dựa vào việc mô tả đối tượng nhờ một số khái niệm biểu thị các đối tượng cơ sở trong ngôn ngữ tự nhiên. Để mô tả đối tượng, người ta dùng một số dạng nguyên thủy như đoạn thẳng, cung,.v.v... Chẳng hạn, một hình chữ nhật được định nghĩa gồm 4 đoạn thẳng vuông góc với nhau từng đôi một. Trong mô hình này người ta sử dụng một bộ kí hiệu kết thúc V_t, một bộ kí hiệu không kết thúc gọi là V_n. Ngoài ra, có dùng một tập các luật sản xuất để mô tả cách xây dựng các đối tượng phù hợp dựa trên các đối tượng đơn giản hơn các đối tượng nguyên thủy (tập V_t). Trong cách tiếp cận này, ta chấp nhận một khẳng định là: Cấu trúc một dạng là kết quả của việc áp dụng luật sản xuất theo những nguyên tắc xác định từ một dạng gốc bắt đầu. Một cách hình thức, ta có thể coi mô hình này tương đương một văn phạm $G = (V_t, V_n, P, S)$ với:

- V_t là bộ kí hiệu kết thúc,

- V_n là bộ kí hiệu không kết thúc,

- P là luật sản xuất,

- S là dạng (kí hiệu bắt đầu)

1.1.2.2. Bản chất của quá trình nhận dạng

Quá trình nhận dạng gồm 3 giai đoạn chính [1]:

- Lựa chọn mô hình biểu diễn đối tượng,

- Lựa chọn luật ra quyết định (phương pháp nhận dạng) và suy diễn quá trình học.

- Học nhận dạng.

Khi mô hình biểu diễn đã được xác định, có thể là định lượng (mô hình tham số) hay định tính (mô hình cấu trúc), quá trình nhận dạng chuyển sang giai đoạn học. Học là giai đoạn rất quan trọng. Thao tác học nhằm cải thiện, điều chỉnh việc phân hoạch tập đối tượng thành các lớp.

Việc nhận dạng là tìm ra quy luật và các thuật toán để có thể gán đối tượng vào một lớp hay nói một cách khác gán cho đối tượng một tên.

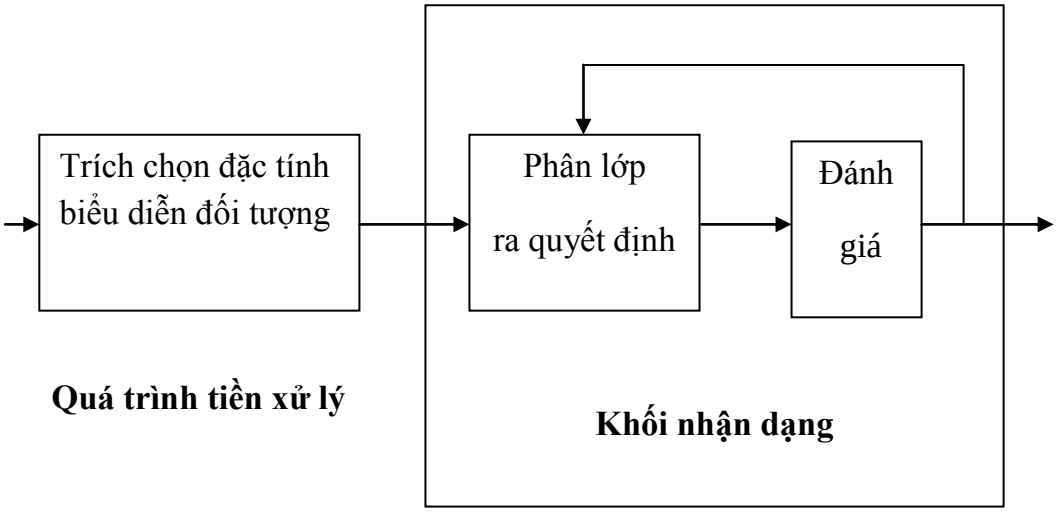
Học có thầy (supervised learning)

Kỹ thuật phân loại nhờ kiến thức biết trước gọi là học có thầy. Đặc điểm cơ bản của kỹ thuật này là người ta có một thư viện các mẫu chuẩn. Mẫu cần nhận dạng sẽ được đem đối sánh với mẫu chuẩn để xem nó thuộc loại nào. Thí dụ như trong một ảnh viễn thám, người ta muốn phân biệt một cánh đồng lúa, một cánh rừng hay một vùng đất hoang mà đã có các miêu tả về các đối tượng đó. Vấn đề chủ yếu là thiết kế một hệ thống để có thể đối sánh đối tượng trong ảnh với mẫu chuẩn và quyết định gán cho chúng vào một lớp. Việc đối sánh nhờ vào các thủ tục ra quyết định dựa trên một công cụ gọi là *hàm phân lớp* hay *hàm ra quyết định*. Hàm này sẽ được đề cập trong phần sau.

Học không có thầy (unsupervised learning)

Kỹ thuật học này tự định ra các lớp khác nhau và xác định các tham số đặc trưng cho từng lớp. Học không có thày đương nhiên là khó khăn hơn. Một mặt, do số lớp không được biết trước, mặt khác những đặc trưng của các lớp cũng không biết trước. Kỹ thuật này nhằm tiến hành mọi cách gộp nhóm có thể và chọn lựa cách tốt nhất. Bắt đầu từ tập dữ liệu, nhiều thủ tục xử lý khác nhau nhằm phân lớp và nâng cấp dần để được một phương án phân loại.

Nhìn chung, dù là mô hình nào và kỹ thuật nhận dạng ra sao, một hệ thống nhận dạng có thể tóm tắt theo sơ đồ sau:



Hình 1.1. Sơ đồ tổng quát một hệ nhận dạng.

1.2. Nhận dạng dựa trên phân hoạch không gian.

Trong kỹ thuật này, các đối tượng nhận dạng là các đối tượng định lượng, mỗi đối tượng được biểu diễn bởi một vectơ nhiều chiều. Trước tiên, ta xem xét một số khái niệm như: phân hoạch không gian, hàm phân biệt sau đó sẽ đi vào một số kỹ thuật cụ thể.

1.2.1. Phân hoạch không gian

Giả sử không gian đối tượng X được định nghĩa: $X=\{X_i,i=1,2,...,m\}$, X_i là một vectơ. Người ta nói P là một phân hoạch của không gian X thành các lớp C_i, $C_i \subset X$ nếu: $C_i \cap C_j = \Phi$ với $i \neq j$ và $\cup C_i = X$

Nói chung, đây là trường hợp lý tưởng: tập X tách được hoàn toàn. Trong thực tế, thường gặp không gian biểu diễn tách được từng phần. Như vậy phân loại là dựa vào việc xây dựng một ánh xạ $f: X \rightarrow P$. Công cụ xây dựng ánh xạ này là các hàm phân biệt (Descriminant functions).

1.2.2. Hàm phân lớp hay hàm ra quyết định

Để phân đối tượng vào các lớp, ta phải xác định số lớp và ranh giới giữa các lớp đó. Hàm phân lớp hay hàm phân biệt là một công cụ rất quan trọng. Gọi {g} là lớp các hàm phân lớp. Lớp hàm này được định nghĩa như sau:

Nếu $\forall$ i $\neq$ k, $g_k(X)>g_i(X)$ thì ta quyết định X$\in$lớp k.

Như vậy để phân biệt k lớp, ta cần k-1 hàm phân biệt. Hàm phân biệt g của một lớp nào đó thường dùng là hàm tuyến tính, có nghĩa là:

$g(X)= W_0+W_1X_1+W_2X_2+...+W_kX_k$

trong đó:

- W_i là các trọng số gán cho các thành phần X_i.

- W_0 là trọng số để viết cho gọn.

Trong trường hợp g là tuyến tính, người ta nói việc phân lớp là tuyến tính hay siêu phẳng (hyperplan).

Các hàm phân biệt thường được xây dựng dựa trên khái niệm khoảng cách hay dựa vào xác suất có điều kiện.

Lẽ tự nhiên, khoảng cách là một công cụ rất tốt để xác định xem đối tượng có "*gần nhau*" hay không. Nếu khoảng cách nhỏ hơn một ngưỡng τ nào đấy ta coi đối tượng là giống nhau và gộp chúng vào một lớp. Ngược lại, nếu khoảng cách lớn hơn ngưỡng, có nghĩa là chúng khác nhau và ta tách thành hai lớp.

Trong một số trường hợp, người ta dựa vào xác suất có điều kiện để phân lớp cho đối tượng. Lý thuyết xác suất có điều kiện được Bayes nghiên cứu khá kỹ và chúng ta có thể áp dụng lý thuyết này để phân biệt đối tượng.

Gọi: $P(X/C_i)$ là xác suất để có X biết rằng có xuất hiện lớp C_i

$P(C_i/X)$ là xác suất có điều kiện để X thuộc lớp C_i

với X là đối tượng nhận dạng, C_i là các lớp đối tượng (lớp thứ i)

Quá trình học cho phép ta xác định $P(X/C_i)$ và nhờ công thức Bayes về xác suất có điều kiện áp dụng trong điều kiện nhiều biến, chúng ta sẽ tính được $P(C_i/X)$theo công thức: $P(C_i/X) = \frac{P(X/C_i)P(C_i)}{\sum_{i=1}^{n} P(C/X_i)P(C_i)} = \frac{P(X/C_i)P(C_i)}{P(X)}$ (1.2)

Nếu $P(C_i/X)>P(C_k/X)$ với $\forall i \neq k$ thì $X \in C_i$. Tùy theo các phương pháp nhận dạng khác nhau, hàm phân biệt sẽ có các dạng khác nhau.

1.2.3. Nhận dạng thống kê

Nếu các đối tượng nhận dạng tuân theo luật phân bố Gauss, mà hàm mật độ xác suất cho bởi:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-m)^2}{2\sigma^2}\right\} \qquad \forall x$$

người ta có dùng phương pháp ra quyết định dựa vào lý thuyết Bayes. Lý thuyết Bayes thuộc loại lý thuyết thống kê nên phương pháp nhận dạng dựa trên lý thuyết Bayes có tên là phương pháp thống kê.

Quy tắc Bayes

- Cho không gian đối tượng $X = \{X_l, l = 1,2,...,L\}$, với $X_1 = \{x_1,x_2,...,x_p\}$

- Cho không gian diễn dịch $\Omega = \{C_1,C_2,...,C_r\}$,r là số lớp

Quy tắc Bayes phát biểu như sau:

ε: $X \rightarrow \Omega$ sao cho $X \in C_k$ nếu $P(C_k/X) > P(C_1/X)$ $\forall l \neq k$, l=1,2,...,r.

Trường hợp lý tưởng là nhận dạng luôn đúng, có nghĩa là không có sai số. Thực tế, luôn tồn tại sai số ε trong quá trình nhận dạng. Vấn đề ở đây là xây dựng quy tắc nhận dạng với sai số ε là nhỏ nhất.

Phương pháp ra quyết định với ε tối thiểu

Ta xác định $X \in C_k$ nhờ xác suất $P(C_k/X)$. Vậy nếu có sai số, sai số sẽ được tính bởi $1-P(C_k/X)$. Để đánh giá sai số trung bình, người ta xây dựng một ma trận L(r, r) giả thiết là có n lớp.

Ma trận L được định nghĩa như sau

$$L_{k,j} = \begin{cases} l_{k,j} > 0 & \text{nếu } k \neq j \\ l_{k,j} \leq 0 & k = j \end{cases} \quad (1.3)$$

Như vậy, sai số trung bình của sự phân lớp sẽ là:

$$r_k(X) = \sum_{j=1}^{r} l_{k,j} P(C_j / X) \quad (1.4)$$

Để sai số nhỏ nhất ta cần có r_k là min. Từ công thức (1.2) và (1.4) ta có:

$$r_k(X) = \sum_{j=1}^{r} l_{k,j} P(X / C_j) P(C_j) \quad (1.5)$$

Vậy, quy tắc ra quyết định dựa trên lý thuyết Bayes có tính đến sai số được phát biểu như sau:

$$X \in C_k \text{ nếu } p_k < p_p \text{ với } p \neq k,\ p=1,2,...,r. \quad (1.6)$$

với p_k là $r_k(X)$.

Trường hợp đặc biệt với 2 lớp C_1 và C_2, ta dễ dàng có:

$$X \in C_1 \text{ nếu } P'(X/C_1) > \frac{l_{12} - l_{22}}{l_{11} - l_{21}} \frac{P(C_2)}{P(C_1)} P(X/C_2) \quad (1.7)$$

Giả sử thêm rằng xác suất phân bố là đều $P(C_1) = P(C_2)$, sai số là như nhau ta có:

$$X \in C_1 \text{ nếu } P(X/C_1) > P(X/C_2) \quad (1.8)$$

1.2.4. Một số thuật toán nhận dạng tiêu biểu trong tự học

Thực tế có nhiều thuật toán nhận dạng học không có thầy. Ở đây, chúng ta xem xét ba thuật toán hay được sử dụng: Thuật toán nhận dạng dựa vào khoảng cách lớn nhất, thuật toán K-trung bình (K mean) và thuật toán ISODATA. Chúng ta lần lượt xem xét các thuật toán này vì chúng có bước tiếp nối, cải tiến từ thuật toán này qua thuật toán khác.

1.2.4.1. Thuật toán dựa vào khoảng cách lớn nhất

a) Nguyên tắc

Cho một tập gồm m đối tượng, ta xác định khoảng cách giữa các đối tượng và khoảng cách lớn nhất ứng với phần tử xa nhất tạo nên lớp mới. Sự phân lớp được hình thành dần dần dựa vào việc xác định khoảng cách giữa các đối tượng và các lớp.

b) Thuật toán [1]

Bước 1

- Chọn hạt nhân ban đầu: giả sử $X_1 \in C_1$ gọi là lớp g_1. Gọi Z_1 là phần tử trung tâm của g_1.

- Tính tất cả các khoảng cách $D_{j1} = D(X_j,Z_1)$ với j =1,2,...,m.

- Tìm $D_{k1} = \max_j D_{j1}$. X_k là phần tử xa nhất của nhóm g_1. Như vậy X_k là phần tử trung tâm của lớp mới g_2, kí hiệu Z_2.

- Tính $d_1 = D_{12} = D(Z_1,Z_2)$.

Bước 2

- Tính các khoảng cách D_{j1}, D_{j2}.

- $D_{j1} = D(X_j,Z_1)$, $D_{j2} = D(X_j,Z_2)$.. Đặt $D_k^{(2)} = \max_j D_j$

Nguyên tắc chọn

- Nếu $D_k^{(2)} < \theta\ d_1$ kết thúc thuật toán. Phân lớp xong.

- Nếu không, sẽ tạo nên nhóm thứ ba. Gọi X_k là phần tử trung tâm của g_3, kí hiệu Z_3.

- Tính $d_3 = (D_{12} + D_{13} + D_{23})/3$

với θ là ngưỡng cho trước và $D_{13} = (Z_1,Z_3)$, $D_{23} = D(Z_2,Z_3)$.

Quá trình cứ lặp lại như vậy cho đến khi phân xong. Kết quả là ta thu được các lớp với các đại diện là $Z_1,Z_2,...,Z_m$.

1.2.4.2. Thuật toán K trung bình (giả sử có K lớp)

a) Nguyên tắc

Khác với thuật toán trên, ta xét K phần tử đầu tiên trong không gian đối tượng, hay nói một cách khác ta cố định K lớp. Hàm để đánh giá là hàm khoảng cách Euclide:

$$J_k = \Sigma_{x \in gk} D(X,Z_k) = \sum_{j=1}^{k} D_2 (X_j,Z_k) \quad (1.9)$$

J_k là hàm chỉ tiêu với lớp C_k. Việc phân vùng cho k hạt nhân đầu tiên được tiến hành theo nguyên tắc khoảng cách cực tiểu. Ở đây, ta dùng phương pháp đạo hàm để tính cực tiểu.

Xét $\frac{\partial J_{jk}}{\partial Z_k} = 0$ với Z_k là biến. Ta dễ dàng có (1.9) min khi:

$$\sum_{i=1}^{N}(X_i - Z_k) = 0 \Rightarrow Z_k = \frac{1}{N_c}\sum_{j=1}^{N_c} Z_j \quad (1.10)$$

Công thức (1.10) là giá trị trung bình của lớp C_k và điều này lý giải tên của phương pháp.

b)Thuật toán [1]

- Chọn N_c phần tử (giả thiết có N_c lớp) của tập T. Gọi các phần tử trung tâm của các lớp đó là: $X_1,X_2,...X_{Nc}$.

- Thực hiện phân lớp

$X \in C_k$ nếu $D(X,Z_k) = \text{Min } D(X,Z_j)^{(1)}$, j =1,...,$N_c$. là lần lặp thứ nhất.

Tính tất cả Z_k theo công thức (1.10).

Tiếp tục như vậy cho đến bước q.

$X \in G_k(q\text{-}1)$ nếu $D(X,Z_k^{(q-1)}) = \min_1 D(X,Z_1^{(q-1)})$.

Nếu $Z_k^{(q-1)} = Z_k^{(q)}$ thuật toán kết thúc, nếu không ta tiếp tục thực hiện phân lớp.

1.2.4.3. Thuật toán ISODATA

ISODATA là viết tắt của từ Iteractive Self Organizing Data Analysis. Nó là thuật toán khá mềm dẻo, không cần cố định các lớp trước. Các bước của thuật toán mô tả như sau: [1]

- Lựa chọn một phân hoạch ban đầu dựa trên các tâm bất kỳ. Thực nghiệm đã chứng minh kết quả nhận dạng không phụ thuộc vào phân lớp ban đầu.

- Phân vùng bằng cách sắp các điểm vào tâm gần nhất dựa vào khoảng cách Euclide.

- Tách đôi lớp ban đầu nếu khoảng cách lớn hơn ngưỡng t_1.

Xác định phân hoạch mới trên cơ sở các tâm vừa xác định lại và tiếp tục xác định tâm mới.

- Tính tất cả các khoảng cách đến tâm mới.

- Nhóm các vùng với tâm theo ngưỡng t_2.

Lặp các thao tác trên cho đến khi thỏa tiêu chuẩn phân hoạch.

1.3. Nhận dạng theo cấu trúc

1.3.1. Biểu diễn định tính

Ngoài cách biểu diễn theo định lượng như đã mô tả ở trên, tồn tại nhiều kiểu đối tượng mang tính định tính. Trong cách biểu diễn này, người ta quan tâm đến các dạng và mối quan hệ giữa chúng. Giả thiết rằng mỗi đối tượng được biểu diễn bởi một dãy ký tự. Các đặc tính biểu diễn bởi cùng một số ký tự. Phương pháp nhận dạng ở đây là nhận dạng lôgic, dựa vào hàm phân biệt là hàm Bool. Cách nhận dạng là nhận dạng các từ có cùng độ dài.

Giả sử hàm phân biệt cho mọi ký hiệu là $g_a(x)$, $g_b(x)$,..., tương ứng với các ký hiệu a,b,... Để dễ dàng hình dung, ta giả sử có từ "abc" được biểu diễn bởi một dãy ký tự $X = \{x_1,x_2,x_3,x_4\}$. Tính các hàm tương ứng với 4 ký tự và có:

$$g_a(x_1) + g_b(x_2) + g_c(x_3) + g_c(x_4)$$

Các phép cộng ở đây chỉ phép toán OR. Trên cơ sở tính giá trị cực đại của hàm phân biệt, ta quyết định X có thuộc lớp các từ "abc" hay không.

1.3.2. Phương pháp ra quyết định dựa vào cấu trúc

1.3.2.1. Một số khái niệm

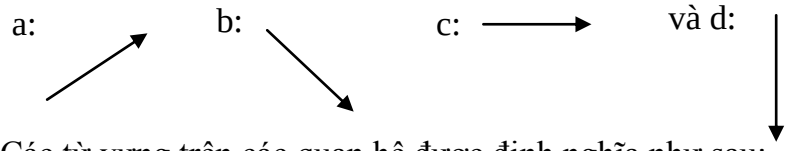
Thủ tục phân loại và nhận dạng ở đây gồm 2 giai đoạn: Giai đoạn đầu là giai đoạn xác định các quy tắc xây dựng, tương đương với việc nghiên cứu một văn phạm trong một ngôn ngữ chính thống. Giai đoạn tiếp theo khi đã có văn phạm là xem xét tập các dạng có được sinh ra từ các dạng đó không? Nếu nó thuộc tập đó coi như ta đã phân loại xong. Tuy nhiên, văn phạm là một vấn đề lớn. Trong nhận dạng cấu trúc, ta mới chỉ sử dụng được một phần rất nhỏ mà thôi.

Như trên đã nói, mô hình cấu trúc tương đương một văn phạm G:

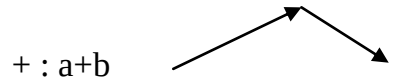
G = {V_n, V_t, P, S}. Có rất nhiều kiểu văn phạm từ chính tắc, phi ngữ cảnh. Ở đây, xin giới thiệu một ngôn ngữ có thể được áp dụng trong nhận dạng cấu trúc: Đó là ngôn ngữ PLD (Picture Language Description).

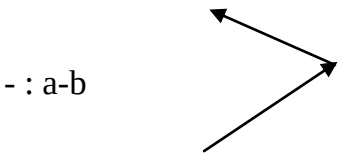
Ví dụ: Ngôn ngữ PLD

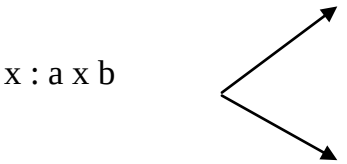
Trong ngôn ngữ này, các từ vựng là các vạch có hướng. Có 4 từ vựng cơ bản:

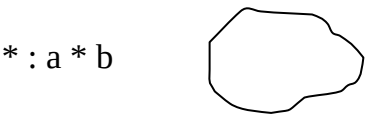


Các từ vựng trên các quan hệ được định nghĩa như sau:









Văn phạm sinh ra các mô tả trong ngôn ngữ được định nghĩa bởi:

$G_A = \{V_n, V_T, P, S\}$

Với $V_n = \{A, B, C, D, E\}$ và $V_T = \{a, b, c, d\}$. S là kí hiệu bắt đầu và P là tập luật sản xuất. Ngôn ngữ này thường dùng nhận dạng các mạch điện.

1.3.2.2. Phương pháp nhận dạng

Các đối tượng cần nhận dạng theo phương pháp này được biểu diễn bởi một câu trong ngôn ngữ L(G). Khi đó thao tác phân lớp chính là xem xét một đối tượng có thuộc văn phạm L(G) không? Nói cách khác nó được sinh ra bởi các luật của văn phạm G không? Như vậy sự phân lớp là theo cách tiếp cận cấu trúc đòi hỏi phải xác định:

- Tập V_t chung cho mọi đối tượng.

- Các quy tắc sinh V để sản sinh ra một câu và chúng khác nhau đối với mỗi lớp.

- Quá trình học với các câu biểu diễn các đối tượng mẫu l nhằm xác định văn phạm G.

- Quá trình ra quyết định: Xác định một đối tượng X được biểu diễn một câu l_x. Nếu l_x nhận biết bởi ngôn ngữ $L(G_x)$ thì ta nói rằng X ɔ C_k.

Nói cách khác, việc ra quyết định phân lớp là dựa vào phân tích cú pháp G_k biểu diễn lớp C_k của văn phạm. Cũng như trong phân tích cú pháp ngôn ngữ, có phân tích trên xuống, dưới lên, việc nhận dạng theo cấu trúc cũng có thể thực hiện theo cách tượng tự.

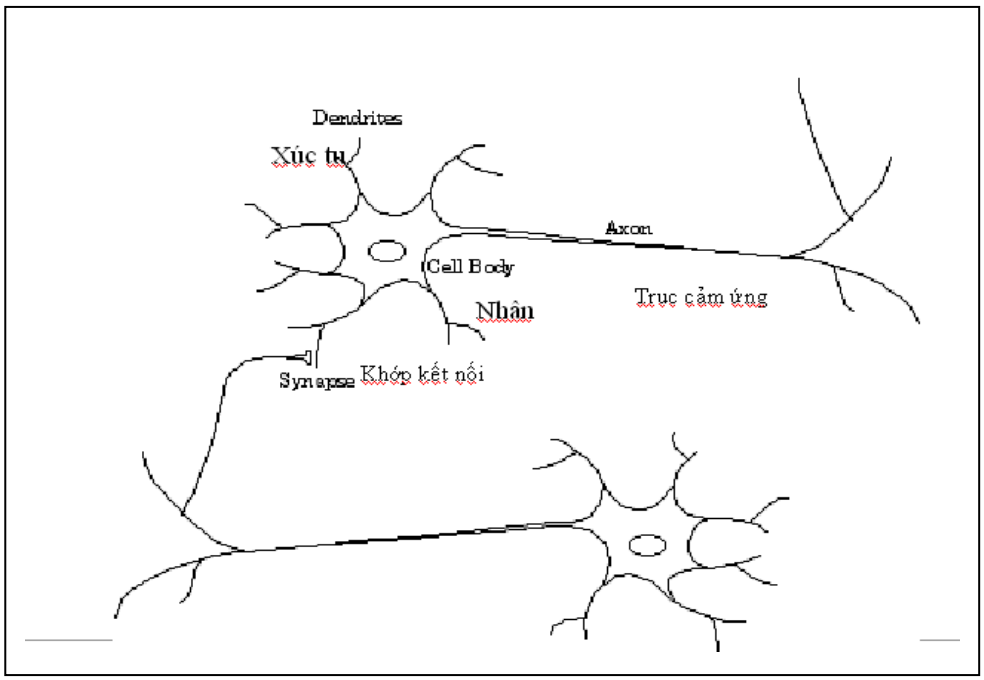
Việc nhận dạng theo cấu trúc là một ý tưởng và dẫu sao cũng cần được nghiên cứu thêm.

1.4. Mạng nơron nhân tạo và nhận dạng theo mạng nơron

Trước tiên, cần xem xét một số khái niệm về bộ não cũng như cơ chế hoạt động của mạng nơron sinh học. [3]

1.4.1. Bộ não và Nơron sinh học

Các nhà nghiên cứu sinh học về bộ não cho ta thấy rằng các nơron (tế bào thần kinh) là đơn vị cơ sở đảm nhiệm những chức năng xử lý nhất định trong hệ thần kinh, bao gồm não, tủy sống, các dây thần kinh. Mỗi nơron có phần thân với nhân bên trong (gọi là soma), một đầu thần kinh ra (gọi là sợi trục axon) và một hệ thống dạng cây các dây thần kinh vào (gọi là dendrite). Các dây thần kinh vào tạo thành một lưới dày đặc xung quanh thân tế bào, chiếm diện tích khoảng 0,25 mm^2, còn dây thần kinh ra tạo thành trục dài có thể từ 1 cm đến hàng mét. Đường kính của nhân tế bào thường chỉ là 10^{-4}m. Trục dây thần kinh ra cũng có thể phân nhánh theo dạng cây để nối các dây thần kinh vào hoặc trực tiếp với nhân tế bào các nơron khác thông qua các khớp nối (gọi là Synapse). Thông thường, mỗi nơron có thể gồm vài trục tới hàng trăm ngàn khớp nối để nối các nơron khác. Người ta ước lượng rằng lưới các dây thần kinh ra cùng với các khớp nối bao phủ diện tích khoảng 90% bề mặt nơron (hình 1.2)



Hình 1.2. Cấu tạo nơron sinh học

Các tín hiệu truyền trong các dây thần kinh vào và dây thần kinh ra của các nơron là tín hiệu điện và được thực hiện thông qua các quá trình phản ứng và giải phóng các chất hữu cơ. Các chất này được phát ra từ các khớp nối dẫn tới các dây thần kinh vào sẽ làm tăng hay giảm điện thế của nhân tế bào. Khi điện thế này đạt tới một

ngưỡng nào đó, sẽ tạo ra một xung điện dẫn tới trục dây thần kinh ra. Xung này được truyền theo trục, tới các nhánh rẽ khi chạm tới các khớp nối với các nơron khác sẽ giải phóng các chất truyền điện. Người ta chia làm hai loại khớp nối: khớp nối kích thích (Excitatory) hoặc khớp nối ức chế (Inhibitory).

Phát hiện quan trọng nhất trong ngành nghiên cứu về bộ não là các liên kết khớp thần kinh khá mềm dẻo, có thể biến động và chỉnh đổi theo thời gian tùy thuộc vào các dạng kích thích. Hơn nữa, các nơron có thể sản sinh các liên kết mới các nơron khác và đôi khi, lưới các nơron có thể di chú từ vùng này sang vùng khác trong bộ não. Các nhà khoa học đây chính là cơ sở quan trọng để giải thích cơ chế của bộ não con người.

Phần lớn các quá trình xử lý thông tin đều xảy ra trên vỏ não. Toàn bộ vỏ não được bao phủ bởi mạng các tổ chức cơ sở có dạng hình thùng tròn với đường kính khoảng 0,5 mm, độ cao khoảng 4mm. Mỗi đơn vị cơ sở này chứa khoảng 2000 nơron. Người ta chỉ ra rằng mỗi vùng não có những chức năng. Điều rất đáng ngạc nhiên là các nơron rất đơn giản trong cơ chế làm việc, nhưng mạng các nơron liên kết với nhau lại có khả năng tính toán, suy nghĩ, ghi nhớ và điều khiển. Có thể điểm qua những chức năng cơ bản của bộ não như sau:

- Bộ nhớ được tổ chức theo các bó thông tin và truy cập theo nội dung (có thể truy xuất thông tin dựa theo giá trị các thuộc tính của đối tượng);

- Bộ não có khả năng tổng quát hóa, có thể truy xuất các tri thức hay các mối liên kết chung của các đối tượng tương ứng với một khái niệm chung nào đó;

- Bộ não có khả năng dung thứ lỗi theo nghĩa có thể điều chỉnh hoặc tiếp tục thực hiện ngay khi có những sai lệch do thông tin bị thiếu hoặc không chính xác. Ngoài ra, bộ não còn có thể phát hiện và phục hồi các thông tin bị mất dựa trên sự tương tự giữa các đối tượng;

- Bộ não có khả năng xuống cấp và thay thế dần dần. Khi có những trục trặc tại các vùng não (do bệnh, chấn thương) hoặc bắt gặp những thông tin hoàn toàn mới lạ, bộ não vẫn tiếp tục làm việc;

- Bộ não có khả năng học.

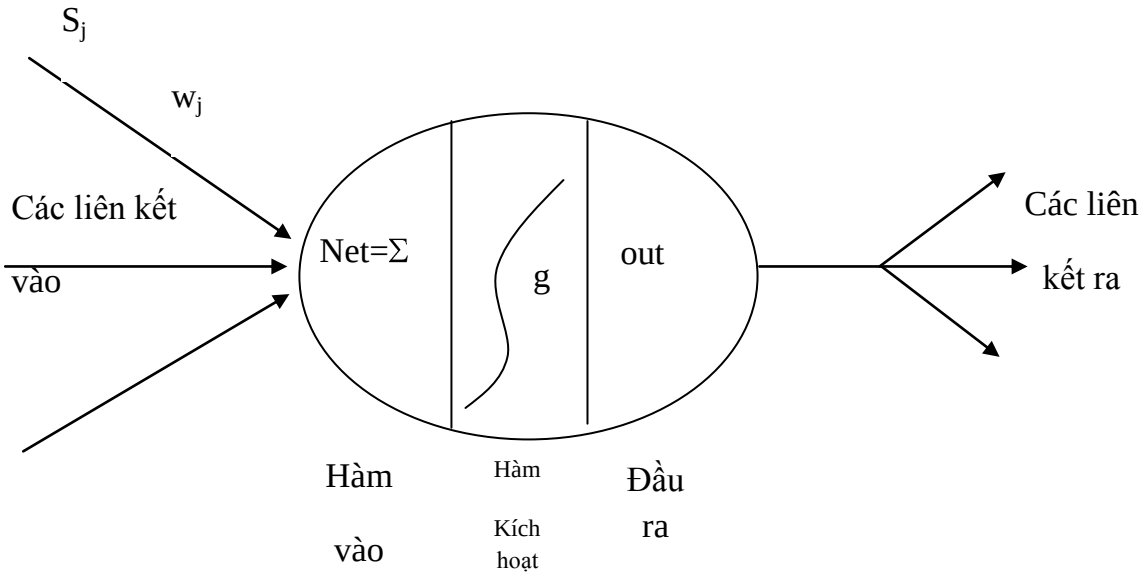
Cách tiếp cận mạng nơron nhân tạo có ý nghĩa thực tiễn lớn cho phép tạo ra các thiết bị có thể kết hợp khả năng song song cao của bộ não với tốc độ tính toán cao của máy tính. Tuy vậy, cần phải có một khoảng thời gian dài nữa để các mạng nơron nhân tạo có thể mô phỏng được các hành vi sáng tạo của bộ não con người. Chẳng hạn, bộ não có thể thực hiện một nhiệm vụ khá phức tạp như nhận ra khuôn mặt người quen sau không quá một giây, trong khi đó một máy tính tuần tự phải thực hiện hàng tỉ phép tính (khoảng 10 giây) để thực hiện cùng thao tác đó, nhưng với chất lượng kém hơn nhiều, đặc biệt trong trường hợp thông tin không chính xác, không đầy đủ.

1.4.2. Mô hình mạng nơron

Mạng nơron nhân tạo (Artificial Neural Network) bao gồm các nút (đơn vị xử lý, nơron) được nối với nhau bởi các liên kết nơron. Mỗi liên kết kèm theo một trọng số nào đó, đặc trưng cho hoạt tính kích hoạt/ức chế giữa các nơron. Có thể xem các trọng số là phương tiện để lưu giữ thông tin dài hạn trong mạng nơron và nhiệm vụ của quá trình huấn luyện (học) mạng là cập nhật các trọng số khi có thêm các thông tin về các mẫu mô phỏng hoàn toàn phù hợp môi trường đang xem xét.

Trong mạng, một số nơron được nối với môi trường bên ngoài như các đầu ra, đầu vào.

1.4.2.1. Mô hình nơron nhân tạo



Hình 1.3. Mô hình nơron nhân tạo

Mỗi nơron được nối với các nơron khác và nhận được các tín hiệu s_j từ chúng với các trọng số w_j. Tổng các thông tin vào có trọng số là:

$$\text{Net} = \sum w_j s_j .$$

Người ta gọi đây là thành phần tuyến tính của nơron. Hàm kích hoạt g (còn gọi là hàm chuyển) đóng vai trò biến đổi từ Net sang tín hiệu đầu ra out.

out =g(Net).

Đây là thành phần phi tuyến của nơron. Có ba dạng hàm kích hoạt thường được dùng trong thực tế:

Hàm dạng bước

$$\text{step}(x) = \begin{cases} 1 & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases} \quad \text{hoặc} \quad \text{step}(x) = \begin{cases} 1 & \text{if } x \geq \theta \\ 0 & \text{if } x < \theta \end{cases}$$

Hàm dấu

$$\text{sign}(x) = \begin{cases} 1 & \text{if } x \geq 0 \\ -1 & \text{if } x < 0 \end{cases} \qquad \text{hoặc} \qquad \text{sign}(x) = \begin{cases} 1 & \text{if } x \geq \theta \\ -1 & \text{if } x < \theta \end{cases}$$

Hàm sigmoid được tính $\text{Sigmoid}(x) = \frac{1}{1+e^{-\alpha(x+\theta)}}$

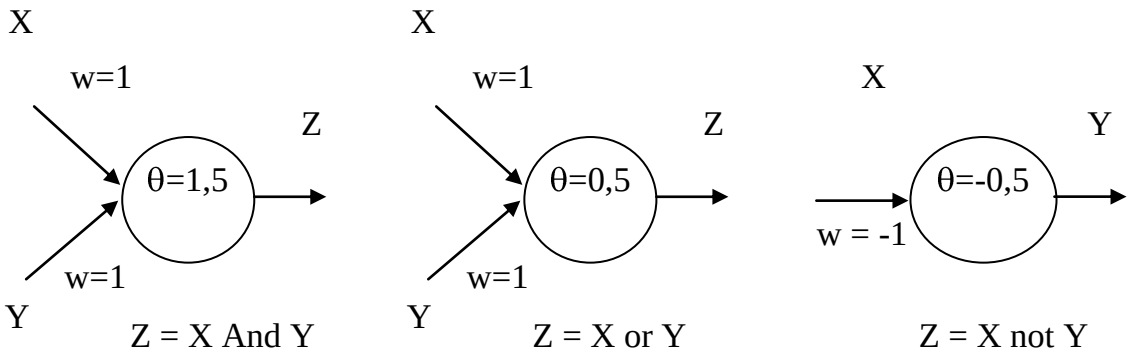
Ở đây ngưỡng θ đóng vai trò làm tăng tính thích nghi và khả năng tính toán của mạng nơron. Sử dụng ký pháp vectơ, S = (s_1,...,s_n) vectơ tín hiệu vào, w=(w_1,...,w_n) vectơ trọng số, ta có

out = g(Net), Net =SW.

Trường hợp xét ngưỡng θ, ta dùng biểu diễn vectơ mới $S' = (s_1,...s_n,\theta)$, $W' = (w_1,...,w_n,-1)$.

Khả năng biểu diễn của nơron

Bộ vi xử lý máy tính dựa trên tích hợp các mạch logic cơ sở. Có thể thấy rằng các nơron hoàn toàn mô phỏng khả năng tính toán của các mạch cơ sở AND, OR, NOT.



1.4.2.2. Mạng nơron

Mạng nơron là hệ thống bao gồm nhiều phần tử xử lý đơn giản (nơron) hoạt động song song. Tính năng của hệ thống này tùy thuộc vào cấu trúc của hệ thống, các trọng số liên kết nơron và quá trình tính toán tại các nơron đơn lẻ. Mạng nơron có thể học từ dữ liệu mẫu và tổng quát hóa dựa trên các dữ liệu mẫu học. Trong mạng nơron, các nơron đón nhận tín hiệu vào gọi là nơron vào và các nơron đưa thông tin ra gọi là nơron ra.

1.5. Kết luận

Có rất nhiều vấn đề nhận dạng khác mà chúng ta chưa đề cập đến như nhận dạng tín hiệu, nhận dạng tiếng nói, v.v. Các vấn đề này nằm trong lý thuyết nhận dạng. Mục đích của chương này nhằm cung cấp một cách nhìn tổng quan về nhận dạng. Các hướng nghiên cứu khác nhau hiện nay trên thế giới về lĩnh vực nhận dạng nói chung.

CHƯƠNG 2

ỨNG DỤNG LÝ THUYẾT THỐNG KÊ TOÁN HỌC ĐỀ GIẢI BÀI TOÁN NHẬN DẠNG NGÔN NGỮ TỰ NHIÊN VÀ ỨNG DỤNG VÀO VIỆC DÒ TÌM KHÓA MÃ TRONG PHÂN TÍCH MẬT MÃ

Kỹ thuật nhận dạng bằng thống kê toán học có nhiều ý nghĩa trong nghiên cứu và thực tiễn. Nó không những được ứng dụng trong nhận dạng ngôn ngữ mà còn đối với hình ảnh, âm thanh, tiếng nói v.v... Trong phạm vi nghiên cứu này, tác giả trình bày một ứng dụng quan trọng. Đó là ứng dụng kỹ thuật thống kê Toán học để nhận dạng các ngôn ngữ tự nhiên (lớp ngôn ngữ la tinh) ứng dụng nhận dạng ngôn ngữ vào việc tìm khóa với một thuật toán mã hóa, thông tin đã được mã hóa và đã được giấu vào trong ảnh. Đây là những hướng ứng dụng mới và có ý nghĩa trong thực tiễn, đặc biệt đối với an ninh quốc phòng. Sau đây là nội dung của nghiên cứu.

2.1. Dạng tổng quát của bài toán

Giả sử ta có một tập hữu hạn $X = \{x^1, x^2, \ldots, x^m\}$ các đối tượng, mỗi đối tượng x^i được đặc trưng bởi n tham số nào đó (như vậy ta hoàn toàn có thể coi X là một tập con, hữu hạn trong không gian Euclid n chiều R^n). Vấn đề đặt ra là: Hãy chia tập X thành K tập con $G_1, G_2, \ldots, G_K$ (với $K \geq 2$); sao cho:

i. $G_i \neq \phi$; với $\forall$ i = 1, 2, . ., k

ii. $G_i \cap G_j = \phi$; với $\forall i, j$; $i \neq j$ và $1 \leq i,j \leq K$ (2.1)

iii. $\bigcup_{i=1}^{k} G_i = X$

Sao cho tổn thất là bé nhất và tốc độc chấp nhận được trong thực tế.

Bài toán này có ý nghĩa thực tiễn quan trọng trong nhiều lĩnh vực Khoa học Kỹ thuật, Tin học, Kinh tế Xã hội và đặc biệt là trong An ninh Quốc phòng, như: phân biệt giọng nói của một đối tượng hình sự nào đó với giọng nói của người khác; hoặc phân

biệt các ngôn ngữ tự nhiên thuộc một lớp các ngôn ngữ nào đó trong An ninh thông tin khi kiểm soát tự động thư tín điện tử Internet…

Ở đây có hai trường hợp xảy ra:

i. Trường hợp số K là đã biết.

ii. Trường hợp số K là chưa biết.

Cách giải quyết bài toán nhận dạng các ngôn ngữ tự nhiên:

1. Xây dựng cơ sở dữ liệu về đặc trưng của các ngôn ngữ.

2. Xây dựng ma trận chuyển trạng thái cho ngôn ngữ đã cho trong cơ sở dữ liệu; tính ước lượng ma trận chuyển trạng thái tương ứng cho mỗi ngôn ngữ.

3. Giải quyết bài toán nhận dạng các ngôn ngữ tự nhiên trong trường hợp số lớp K là đã biết và số lớp K là chưa biết.

2.2. Một số khái niệm và thuật toán

Giả sử $X = \{x \mid x = (x_1, x_2, \dots, x_n)$; x_i là một số nguyên không âm $\forall$ i=1,2, …n$\}$ là một tập hợp tùy ý hữu hạn các véc tơ n thành phần. Với n là một số nguyên dương cho trước, cố định; $x \in X$ được gọi là một đối tượng X. Ta có các khái niệm sau:

2.2.1. Khoảng cách giữa hai đối tượng, hai tập hợp

Với $x,y \in X$, khi đó khoảng cách giữa hai đối tượng x và y được định nghĩa là:

$$d(x,y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + ... + (x_n - y_n)^2} = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

Cho $G_1, G_2 \subseteq X$, khi đó khoảng cách giữa hai tập hợp G_1 và G_2 được định nghĩa là:

$$d(G_1, G_2) = \frac{1}{n_1.n_2} \sum_{x \in G_1} \sum_{x \in G_2} d(x,y)$$

Với $n_1 = \|G_i\|$ là lực lượng của tập G_i với i = 1,2.

2.2.2. Giải bài toán trường hợp cho trước số k

Tư tưởng của phương pháp giải là tìm cách ghép các đối tượng có quan hệ "gần gũi" nhau nhất vào chung một lớp. Như vậy để giải quyết bài toán chúng ta cần xây dựng độ đo của sự gần gủi. Vậy thế nào là độ đo sự gần gủi? [2]

Định nghĩa 1: Một độ đo sự gần gũi giữa 2 đối tượng tùy ý x, y thuộc không gian X đối tượng là một ánh xạ d: X→R (với R là đường thẳng thực) sao cho:

i) $d(x, y) \geq 0 \ \forall\ x, y$ và $d(x, y) = 0 \Leftrightarrow x = y$

ii) $d(x, y) = d(y, x) \ \forall\ x, y \in X$

iii) $d(x, y) \leq d(x, z) + d(z, y) \ \forall\ x, y, z \in X$

Đối với việc giải bài toán phân lớp, chúng ta còn cần đến khái niệm quan hệ gần gủi giữa hai tập hợp. Ta có định nghĩa như sau:

Định nghĩa 2: Giả sử G_1, G_2 là hai tập hợp con tùy ý. Chúng đa dùng khái niệm khoảng cách giữa hai tập hợp để đo sự gần gủi giữa hai tập hợp. Độ đo sử gần gũi của G_1, G_2 được định nghĩa như sau:

$$d(G_1, G_2) = \frac{1}{n_1.n_2} \sum_{x \in G_1} \sum_{x \in G_2} d(x, y)$$

Thuật toán:

Trên cơ sở 2 định nghĩa vừa nêu, tác giả đưa ra thuật toán giải bài toán cho trường hợp số k≥2 cho trước như sau:

Giả sử tập hợp $X=\{x_1, x_2, .., x_n\}$ với $x_i \in R^m, \quad i = 1,2,..,n, \quad n \geq k$

Step1: Đặt $G_1=\{x_1\}$, $G_2=\{x_2\}$, .., $G_n=\{x_n\}$. Với cách phân hoạch tập X như này, rõ ràng thỏa mãn điều kiện (2.1)

Step2: Nếu n = k thì thuật toán dừng và G_1, G_2, .., G_k là kết quả của bài toán.

Step3: Đặt $S(G_{i_0}, G_{j_0}) = \min_{t \neq t'} S(G_t, G_{t'})$

Step4: Đặt $G_1 = G_{i_0} \cup G_{j_0}$. Như vậy ở bước này lần thứ nhất G_1, G_2, .., G_n chỉ còn G_1, G_2, .., G_{n-1} và có thể tồn tại $S(G_{i_0}, G_{j_0}) = \min_{t \neq t'} S(G_t, G_{t'})$ và đồng thời $(G_{i_1}, G_{j_1}) = \min_{t \neq t'} S(G_t, G_{t'})$ lúc đó ta nhóm tất cả tập hợp cùng độ "gần gủi" này thành 1 tập con, và như vậy, một cách tổng quát ta giả sử tại bước thứ l, tập X được phân thành $k^{(l)}$ tập con: $G_1, G_2, .., G_{k^{(l)}}$ (không mất tính tổng quát, để đơn giản kí hiệu ta vẫn kí hiệu như vậy)

Step5: Nếu $k^{(l)} = k$, tức là $G_1, G_2, .., G_{k^{(l)}} = G_1, G_2, .., G_k$ thì thuật toán kết thúc và $G_1, G_2, .., G_{k^{(l)}}$ là đáp số bài toán. Ngược lại thì trở lại **Step3**

<u>Tính đúng đắn của thuật toán:</u>

Ký hiệu:

$S_{ij} = d(x^i, x^j)$

$$S(G_k, G_k) = d(G_k, G_k) = \frac{1}{n_k^2} \sum_{x^i \in G_k} \sum_{x^i \in G_k} d(x^i, x^j) = \frac{1}{n_k^2} \sum_{x^i \in G_k} \sum_{x^j \in G_k} S_{ij}$$

$$S(G_{t_1}, G_{t_2}) = d(G_{t_1}, G_{t_2}) = \frac{1}{n_{t_1} n_{t_2}} \sum_{x^i \in G_{t_1}} \sum_{x^j \in G_{t_2}} d(x^i, x^j) = \frac{1}{n_{t_1} n_{t_2}} \sum_{x^i \in G_{t_1}} \sum_{x^j \in G_{t_2}} S_{ij}$$

Với n_k là lực lượng của G_k, n_{t_1} là lực lượng của G_{t_1}, n_{t_2} là lực lượng của G_{t_2}

Ta có định lý sau:

Định lý: Điều kiện cần và đủ để phép phân hoạch tập X thành K tập con thỏa điều kiện (2.1):

- ✓ $G_i \neq \phi$; với $\forall$ i = 1, 2, . ., k
- ✓ $G_i \cap G_j = \phi$; với $\forall$i, j ; i≠j và $1 \leq$ i,j $\leq$ K
- ✓ $\bigcup_{i=1}^{k} G_i = X$

đúng đắn là: $\max_k \{S(G_k, G_k)\} \le \min_{t_1,t_2} \{S(G_{t_1}, G_{t_2})\}$; với $t_1 \# t_2$

Ta có:

$$S(G,G) = \frac{2}{n_G(n_G-1)} \sum_{i=1}^{n_G} \sum_{j=i+1}^{n_G} S_{ij}$$

S(G,G) được gọi là đại lượng đặc trưng cho sự "gần gũi" giữa các đối tượng x^i trong tập G.

Ví dụ 2.3: Cho $X=\{X_1, X_2, X_3\}$ với X_i có các giá trị sau đây

i	$\mathbf{X_1}$	$\mathbf{X_2}$	$\mathbf{X_3}$
1	58	30	19
2	46	34	16
3	45	39	31
4	61	50	44

Hãy phân tập X thành 2 (k=2) lớp G_1, G_2 sao cho

1) $G_i \neq \phi$ i=1,2

2) $G_1 \cap G_2 = \phi$ 3) $G_1 \cup G_2 = X$

Giải:

$$d_{ij} = d(x_i, x_j) = \begin{cases} \sum_{k=1}^{4} \left| \ln \frac{X_{ki}}{X_{kj}} \right| & \text{Nếu } X_{ki} \text{ và } X_{kj} > 0 \ \forall \ k=1,2,.. \\ \infty & \text{Nếu } X_{ki} \text{ hoặc } X_{kj} = 0 \text{ với một k nào đó} \end{cases}$$

Chọn độ đo sự gần gủi:Rõ ràng rằng việc xác định độ đo khoảng cách như vậy thỏa mãn điều kiện của

Định nghĩa 1, tức là $d(x, y) \ge 0$; $d(x, y) = 0 \Leftrightarrow x = y$; $d(x, y) = d(y, x)$

và $d(x, y) \le d(x, z) + d(z, y) \ \forall \ x, y, z \in X$

Step1: Đặt $G_1=\{X_1\}$, $G_2=\{X_2\}$, $G_3=\{X_3\}$

Step3: Ta có

$$S(X_1, X_2) = \sum_{k=1}^{4} \left| \ln \frac{X_{k1}}{X_{k2}} \right| == \ln \frac{58}{30} + \ln \frac{46}{34} + \ln \frac{45}{39} + \ln \frac{61}{50} \approx 1,303$$

$$S(X_1, X_3) = \sum_{k=1}^{4} \left| \ln \frac{X_{k1}}{X_{k3}} \right| = \ln \frac{58}{19} + \ln \frac{46}{16} + \ln \frac{45}{31} + \ln \frac{61}{44} \approx 2,871$$

$$S(X_2, X_3) = \sum_{k=1}^{4} \left| \ln \frac{X_{k2}}{X_{k3}} \right| = \ln \frac{30}{19} + \ln \frac{34}{16} + \ln \frac{39}{31} + \ln \frac{50}{44} \approx 1,568$$

Do đó $\min_{1 \le i < j \le 3} S(X_i, X_j) = 1,303 = S(G_1, G_2)$

Step4: Ghép G_1, G_2 thành 1 lớp. với việc ghép này ta có $k^{(1)} = 2 = k$ nên thuật toán kết thúc và 2 lớp cần tìm là $G_1=\{X_1, X_2\}$, $G_2=\{ X_3\}$

2.2.3. Giải bài toán trường hợp số k chưa cho biết trước

Đây là trường hợp tổng quát và hay gặp trong thực tế. Trong trường hợp này, chúng ta xây dựng thuật toán xác định số k. Sau khi tìm được số k, bài toán trở về trường hợp giải bài toán số k biết trước

Giả sử $X = \{X_1, X_2,..,X_n\}$ với $X_i \in R^m$ i=1,2,..n ; m≥1 là tập tùy ý các đối tượng, $S_{ij} = d(X_i , X_j)$ là khoảng cách giữa hai đối tượng X_i , X_j. S_{ij} có thể định nghĩa một cách tùy ý thỏa mãn ba tính chất tương đương với (2.1):

$S_{ij} \ge 0$ và $S_{ij} = 0 <=> i = j$

$S_{ij} = S_{ji} \forall$ i, j

$S_{ij} \le S_{ik} + S_{kj} \forall$ i, j, k

Ta ký hiệu: $S = (S_{ij})_{m \times m}$ i,j = 1, 2, .., m

Do tính chất của d_{ij} nên ma trận S cấp m x m là ma trận đối xứng có $d_{ii} = 0,\ i = \overline{1,m}$

Để xác định hằng số K, ta đặt :

Step1: Đặt $F(k) = \sum_{i \in P_k} \sum_{j \in Q_i} d_{ij}$

Trong đó P_k là ma trận con cấp k x m của ma trận S (Nghĩa là P_k là ma trận có k dòng lấy trong m dòng của ma trận S và có m cột) với k ≤ m-1. Một cách tổng quát, đối với ma trận $S_{m\,x\,m}$. Lúc đó ∀k<m, sẽ có $C_m^k = \frac{m!}{(m-k)!k!}$ ma trận cấp k. Còn Q_i là tập các chỉ số cột của ma trận P_k.

Bài toán đặt ra như sau: Hãy xác định số k với 2 ≤ k ≤ m-1, bé nhất có thể được sao cho $F(k) = \min_{P_k, Q_k} F(k)$.

Bổ đề: Để tìm min($F_i(k)$) với 2 < k < m, ta dựa vào bổ đề sau:

Các ma trận P_k^i làm cho $F_i(k)$ đạt min là các ma trận chứa P_k^{i-1} làm cho $F_i(k-1)$ đạt min.

Nội dung xác định số k như sau:

- Ứng với mỗi k cụ thể:

  - K=2 ta lập tất cả ma trận con P_k của S.
  - 2 < K < m ta lập các ma trận P_k của S và thỏa mãn bổ đề trên.

- Tiếp theo, đối với mỗi cột của ma trận con P_k, ta tìm phần tử bé nhất; sau đó lấy tổng tất cả các phần tử bé nhất trong m cột đó của ma trận P_k.

- Ta chọn k = u : thỏa mãn $F_v(u)$ đạt min với $2 \le u \le m$; $v = 1,2,\ldots, C_m^l$

Ví dụ 2.4: Giả sử $S = \begin{bmatrix} 0 & 1 & 2 & 3 & 4 \\ 1 & 0 & 1 & 5 & 6 \\ 2 & 1 & 0 & 3 & 4 \\ 3 & 5 & 3 & 0 & 2 \\ 4 & 6 & 4 & 2 & 0 \end{bmatrix}$

khi đó một trong các ma trận P_2 là $P_2 = \begin{bmatrix} 0 & 1 & 2 & 3 & 4 \\ 1 & 0 & 1 & 5 & 6 \end{bmatrix}$, v.v.

Với k = 2 ta sẽ có C_5^2 ma trận P_2 như vậy, ta có $C_5^2 = \frac{5!}{3!2!} = 10$ ma trận P_2. Đó là:

Ma trận P_2^i	**$F_i(2)$**

$P_2^{(1)} = \begin{bmatrix} 0 & 1 & 2 & 3 & 4 \\ 1 & 0 & 1 & 5 & 6 \end{bmatrix}$	$F_1(2) = 0 + 0 + 1 + 3 + 4$	8
$P_2^{(2)} = \begin{bmatrix} 0 & 1 & 2 & 3 & 4 \\ 2 & 1 & 0 & 3 & 4 \end{bmatrix}$	$F_2(2) = 0 + 1 + 0 + 3 + 4$	8
$P_2^{(3)} = \begin{bmatrix} 0 & 1 & 2 & 3 & 4 \\ 3 & 5 & 3 & 0 & 2 \end{bmatrix}$	$F_3(2) = 0 + 1 + 2 + 0 + 2$	5
$P_2^{(4)} = \begin{bmatrix} 0 & 1 & 2 & 3 & 4 \\ 4 & 6 & 4 & 2 & 0 \end{bmatrix}$	$F_4(2) = 0 + 1 + 2 + 2 + 0$	5
$P_2^{(5)} = \begin{bmatrix} 1 & 0 & 1 & 5 & 6 \\ 2 & 1 & 0 & 3 & 4 \end{bmatrix}$	$F_5(2) = 1 + 0 + 0 + 3 + 4$	8
$P_2^{(6)} = \begin{bmatrix} 1 & 0 & 1 & 5 & 6 \\ 3 & 5 & 3 & 0 & 2 \end{bmatrix}$	$F_6(2) = 1 + 0 + 1 + 0 + 2$	4
$P_2^{(7)} = \begin{bmatrix} 1 & 0 & 1 & 5 & 6 \\ 4 & 6 & 4 & 2 & 0 \end{bmatrix}$	$F_7(2) = 1 + 0 + 1 + 2 + 0$	4
$P_2^{(8)} = \begin{bmatrix} 2 & 1 & 0 & 3 & 4 \\ 3 & 5 & 3 & 0 & 2 \end{bmatrix}$	$F_8(2) = 2 + 1 + 0 + 0 + 2$	5
$P_2^{(9)} = \begin{bmatrix} 2 & 1 & 0 & 3 & 4 \\ 4 & 6 & 4 & 2 & 0 \end{bmatrix}$	$F_9(2) = 2 + 1 + 0 + 2 + 0$	5
$P_2^{(10)} = \begin{bmatrix} 3 & 5 & 3 & 0 & 2 \\ 4 & 6 & 4 & 2 & 0 \end{bmatrix}$	$F_{10}(2) = 3 + 5 + 3 + 0 + 0$	11

$\longrightarrow$ $\min(F_i(2)) = 4$ và các ma trận con tương ứng là: P_2^6 và P_2^7

Với k=3, bình thường số ma trận chon cấp 3 x m của S là chỉnh hợp chập 3 của 5 tức cũng cần tính 10 ma trận nhưng khi áp dụng bổ đề trên ta chỉ cần xét các ma trận sau:

Ma trận P_3^i	**$F_i(3)$**	
$P_3^{(1)} = \begin{bmatrix} 1 & 0 & 1 & 5 & 6 \\ 3 & 5 & 3 & 0 & 2 \\ 0 & 1 & 2 & 3 & 4 \end{bmatrix}$	$F_1(3) = 0 + 0 + 1 + 0 + 2$	3
$P_3^{(2)} = \begin{bmatrix} 1 & 0 & 1 & 5 & 6 \\ 3 & 5 & 3 & 0 & 2 \\ 2 & 1 & 0 & 3 & 4 \end{bmatrix}$	$F_2(3) = 1 + 0 + 0 + 0 + 2$	3
$P_3^{(3)} = \begin{bmatrix} 1 & 0 & 1 & 5 & 6 \\ 3 & 5 & 3 & 0 & 2 \\ 4 & 6 & 4 & 2 & 0 \end{bmatrix}$	$F_3(3) = 1 + 0 + 1 + 0 + 0$	2

$P_3^{(4)} = \begin{bmatrix} 1 & 0 & 1 & 5 & 6 \\ 4 & 6 & 4 & 2 & 0 \\ 0 & 1 & 2 & 3 & 4 \end{bmatrix}$	$F_4(3) = 0 + 0 + 1 + 2 + 0$	3
$P_3^{(5)} = \begin{bmatrix} 1 & 0 & 1 & 5 & 6 \\ 4 & 6 & 4 & 2 & 0 \\ 2 & 1 & 0 & 3 & 4 \end{bmatrix}$	$F_5(3) = 1 + 0 + 0 + 2 + 0$	3
$P_3^{(6)} = \begin{bmatrix} 1 & 0 & 1 & 5 & 6 \\ 4 & 6 & 4 & 2 & 0 \\ 3 & 5 & 3 & 0 & 2 \end{bmatrix}$	$F_6(3) = 1 + 0 + 1 + 0 + 0$	2

$\longrightarrow$ $\min(F_i(3)) = 2$ và các ma trận con tương ứng là: P_3^6 và P_3^3

Trường hợp k=4 ta chỉ cần xét các ma trận sau:

Ma trận P_4^i	**$F_i(4)$**	
$P_4^{(1)} = \begin{bmatrix} 1 & 0 & 1 & 5 & 6 \\ 4 & 6 & 4 & 2 & 0 \\ 3 & 5 & 3 & 0 & 2 \\ 0 & 1 & 2 & 3 & 4 \end{bmatrix}$	$F_1(4) = 0 + 0 + 1 + 0 + 0$	1
$P_4^{(2)} = \begin{bmatrix} 1 & 0 & 1 & 5 & 6 \\ 4 & 6 & 4 & 2 & 0 \\ 3 & 5 & 3 & 0 & 2 \\ 2 & 1 & 0 & 3 & 4 \end{bmatrix}$	$F_1(4) = 1 + 0 + 0 + 0 + 0$	2

$\longrightarrow$ $\min(F_i(4)) =$ 1và các ma trận con tương ứng là: P_4^1 và P_4^2

Vậy $\min(F_v(u)) = F_1(4) = F_2(4) = 1$; Suy ra: K =4. Đến đây bài toán quay về số K đã biết.

2.3. Mô hình xích Markov và phép kiểm định thống kê cho bài toán nhận dạng ngôn ngữ

Chúng ta biết rằng nhận dạng ngôn ngữ là một trong những yêu cầu cực kỳ quan trọng và cần thiết của quá trình phân tích mật mã nói chung. Để nhận dạng được một ngôn ngữ nào đó, trước hết chúng ta cần toán học hóa ngôn ngữ đó như một xích Markov hữu hạn trạng thái. Trên cơ sơ đó, chúng ta sẽ xây dựng một số tiêu chuẩn cụ thể để "nhận dạng " một ngôn ngữ. Vấn đề giải quyết trong nghiên cứu này là: Sử

dụng các phép kiểm định giả thiết xác suất thống kê trên mô hình ngôn ngữ với giả định Markov.

2.3.1 Mô hình xích Markov

Mô hình xích Markov (gọi tắt là Markov) hay xích ngôn ngữ với giả định Markov là một dạng mô hình xác suất thống kê nhận dạng mẫu được áp dụng phổ biến trong xử lý ngôn ngữ. Mô hình xích Markov của ngôn ngữ là mô hình hữu hạn trạng thái có tính dừng (ergodic).

Mô hình Markov của ngôn ngữ được định nghĩa bằng tập 5 tham số

$$(m, A, \{Y_t\}, P, r).$$

Trong đó

$m \in Z^+$: là số các trạng thái mô hình Markov có thể nhận

$A=\{a_1, a_2,..,a_m\}$: là không gian các trạng thái.

$\{Y_t\}$ $t \in T$: là quá trình ngẫu nhiên dừng. $T \subset Z=\{0, \pm1, \pm\}$

P: là ma trận các xác suất chuyển trạng thái

r: là cấp của xích Markov.

Ví dụ, mô hình xích Markov cho tiếng Anh có thể có tham số m=26 và A là tập các ký tự trong Alphabet của ngông ngữ (các ký tự ASCII từ A đến Z). Nếu phân biệt chữ in hoa với chữ in thường hoặc cần xử lý thêm dấu gián cách từ, dấu câu và số đếm, tham số m sẽ tăng lên và không gian trạng thái A đồng thời sẽ mở rộng thêm.

Khi đề xuất mô hình xác suất thống kê, Markov giả định rằng trạng thái hiện tại của mô hình chỉ phụ thuộc vào một số ít các trạng thái mà mô hình đã trải qua trước đó. Số trạng thái phụ thuộc như vậy được gọi là bậc của mô hình và là tham số quyết định độ phức tạp của mô hình.

Biến cố ngẫu nhiên Y_t biểu diễn trạng thái thuộc không gian A mà mô hình nhận tại thời điểm t, Tập $\{Y_t\}$ biểu diễn đoạn mẫu quan sát. Lực lượng của $\{Y_t\}$ cần được lựa chọn thỏa mãn các điều kiện thống kê để qui luật xác suất thể hiện rõ, đồng thời thỏa mãn điều kiện tối thiểu thời gian tính toán trong nhận dạng tự động đáp ứng được

thời gian thực. Tham số m hay kích thước của không gian trạng thái A quyết định nhiều đến độ dài mẫu cần lựa chọn $\{Y_t\}$.

Ma trận xác suất chuyển trạng thái P là tham số cần nhiều bộ nhớ của mô hình Markov. Xác suất chuyển trạng thái thể hiện mối quan hệ giữa các trạng thái phụ thuộc trên mô hình Markov. Bậc của mô hình càng tăng, không gian bộ nhớ cần sử dụng càng lớn và tính phức tạp của ma trận xác suất chuyển trạng thái càng cao. Với r=1, trạng thái hiện tại của mô hình chỉ phụ thuộc vào một trạng thái trýớc ðó, ma trận xác suất chuyển trạng thái chính là xác suất bộ ðôi có ðiều kiện của hai trạng thái xuất hiện liên tiếp nhau của mô hình, không gian bộ nhớ cần để lưu trữ sẽ bằng m^2. Với r=2, trạng thái hiện tại phụ thuộc vào hai trạng thái trước đó, ma trận xác suất chuyển biểu diễn trong không gian ba chiều bởi kích thước bộ nhớ chiếm dụng bằng m^3. Một cách tổng quát, nếu r=k thì không gian bộ nhớ cần để lưu trữ là m^{k+1}. Trong nghiên cứu này ta có m=26 và chọn r=1. Ma trận xác suất chuyển được tính toán bằng ước lượng hợp lý nhất trên tập mẫu có kích thước cỡ trên 100.000 biểu hình cho ngôn ngữ Tiếng Anh.

Ma trận xác suất chuyển trạng thái có thể đơn giản ước lượng từ các mẫu cơ bản. Nói chung các xác suất chuyển P_{ij} $(1\le i,j\le m)$ thường là chưa biết. Nếu mẫu thống kê là đủ lớn thì ước lượng của P_{ij} là $\hat{P}_{ij} = \frac{n_{ij}}{n_i}$;$1 \le i, j \le m$

Trong đó n_{ij} là số lần (tần số) xuất hiện trạng thái j khi cho trước trạng thái i

còn $n_i = \sum_{j=1}^{26} n_{ij}$; $i, j = \overline{1,26}$.

Trong trường hợp độ dài mẫu bé thì ước lượng P_{ij} được cho bởi công thức sau:

$\hat{P}_{ij} = \frac{n_{ij} + c}{n_i + c.m}$;$1 \le i, j \le m$ với c là hằng số thường được chọn c=0,5 hoặc c=1/m [2, 4]

2.3.2 Phép kiểm định thống kê cho bài toán nhận dạng ngôn ngữ đã biết

Định nghĩa : Một bản rõ $X = X_1 X_2 \ldots X_n$; $n \geq 1$ được gọi là có nghĩa (hoặc hợp lý) nếu phân bố thực nghiệm của x phù hợp với phân bố của một xích Markov hữu hạn trạng thái có cấp $r \geq 0$.

Chuyển bài toán nhận dạng ngôn ngữ đã biết về bài toán kiểm tra giả thiết thống kê : "Cho trước một mẫu X. Hãy kiểm tra giả thuyết : " *X được sinh ra từ một xích Markov B đã biết* " với đối thiết : "*X được sinh ra từ một xích Markov khác B nhưng có cùng cấp*". Hay gọi là phân biệt hai ngôn ngữ.

Nếu ta ký hiệu :

- P_M là ma trận các xác suất chuyển đổi với xích Markov sinh ra mẫu x.

- P_B là ma trận các xác suất chuyển đã biết.

Bài toán xem xét mẫu văn bản với giả thiết H_0 : *mẫu P_M phù hợp với ngôn ngữ P_B* và đối thiết H_1 : *mẫu P_M không phù hợp với ngôn ngữ P_B* được áp dụng mô hình ngôn ngữ của tiếng Anh. Để giải quyết bài toán cần xem xét tỷ số hợp lý của mẫu đối với mô hình đối sánh. Mô hình cho kết quả tỷ số hợp lý cao hơn sẽ xác định là ngôn ngữ được dùng để viết ra mẫu. Tuy nhiên các phép kiểm định hoàn toàn có khả năng rơi vào trạng thái tỷ số hợp lý trên các mô hình ngôn ngữ không loại trừ giả thiết H_0 của nhau. Lý do các ngôn ngữ có cùng gốc La tinh (như là : tiếng Anh, Pháp và Đức) có chung 26 chữ cái. Cần có thêm so sánh tuyệt đối cho các trường hợp trong miền không phân định rõ ràng này.

Thống kê Sinkov và các phép kiểm định cho bài toán nhận dạng ngông ngữ

Nếu cho một dãy mẫu thử $X=x_1x_2...x_n$ đơn biểu, thống kê Sinkov tính điểm cho X bằng biểu thức: $S_1 = \sum_{i=1}^{n} \ln(p_{xi})$

Trong đó ứng với mỗi i, p_{xi} là xác suất tiên nghiệm của ký tự x_i. Ví dụ: nếu xác suất tiên nghiệm của ký tự D, E, S lần lượt là 0,044; 0,130 và 0,063 thì giá trị S_1 cho chuỗi X=DES sẽ là ≈-3,12 - 2,04 – 2,67 = -7,92. Xem xét giá trị S_1, Sinkov khẳng định

tính hợp lý của chuỗi DES bằng trực giác. Tuy nhiên, Sinkov không giải thích cơ sở lý thuyết đưa đến thống kê kể trên và không đưa ra cá đặc tính phân bố xác suất của nó. Ông cũng không chỉ rõ thủ tục ra quyết định công nhận hay bác bỏ giả thiết trên cơ sở giá trị S_1 tính toán được. Để chấp nhận hay bác bỏ một giả thiết, Sinkov đơn thuần thử mọi khả năng và chấp nhận khả năng cho giá trị S_1 cao nhất.

Sinkov cũng chỉ ra rằng các giá trị S_r (với r=2, 3,..) áp dụng cho biểu hình bộ đôi, bộ ba có hiệu quả trong việc đo mức độ ăn khớp của hai hay ba cột trong thám mã chuyển vị bảng. Nhưng ông không đề cập đến vấn đề nên sử dụng xác suất không điều kiện hay xác suất có điều kiện của bộ biểu hình. [2, 4]

2.4. Ứng dụng vào việc dò tìm khóa mã trong phân tích mật mã.

2.4.1. Giới thiệu về phân tích mật mã.

Chúng ta đã biết rằng nhận dạng ngôn ngữ là một trong những yêu cầu cực kỳ quan trọng và cần thiết của quá trình phân tích mật mã nói chung. Để nhận dạng được ngôn ngữ nào đó, trước hết chúng ta phải toán học quá ngôn ngữ đó như là một xích Markov hữu hạn trạng thái. Trên cơ sở chúng ta sẽ xây dựng một số tiêu chuẩn cụ thể để nhận dạng một ngôn ngữ.

Việc nhận dạng ngôn ngữ trong quá trình phân tích mật mã thường gặp phải ba khó khăn sau:

1. Mã thám là một ngành đối lập với mật mã. Nhiều khi nhà lập mã tìm cách đánh lừa các nhà phân tích mật mã.

2. Bản rõ được nhận dạng nhiều trường hợp là rất ngắn và chưa hoàn thiện

3. Bản rõ khi nhận được thường bị sai lệch nhiều chỗ do nhiễu.

Bản rõ có thể là 1 dãy số, một chuỗi bít nhị nguyên, một dãy mã Hexa.... chứ không nhất thiết là một văn bản đọc được có nghĩa . Khi nói đến bản rõ thì có thể không đọc được có nghĩa, khi nói bản rõ có nghĩa thì tức là bản rõ đó là văn bản đọc được có cấu trúc cú pháp.

2.4.2. Giới thiệu kỹ thuật giấu thông tin trong ảnh dùng cho việc thám mã.

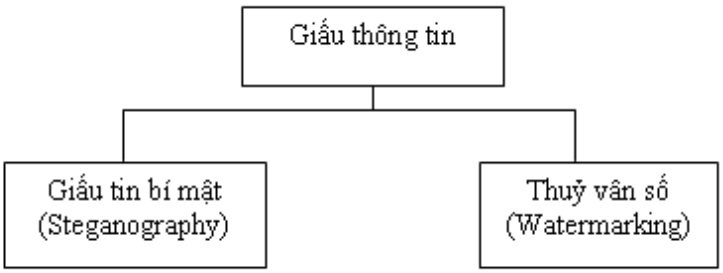
2.4.1.1. Định nghĩa giấu tin.

Giấu tin là một kỹ thuật giấu hoặc nhúng một lượng thông tin số nào đó vào trong một đối tượng dữ liệu số khác (giấu tin nhiều khi không phải là hành động giấu cụ thể mà chỉ mang ý nghĩa quy ước).

2.4.1.2. Mục đích của giấu tin: có 2 mục đích của giấu thông tin.

- Bảo mật cho những dữ liệu được giấu

Có thể thấy 2 mục đích này hoàn toàn trái ngược nhau và dần phát triển thành 2 lĩnh vực với những yêu cầu và tính chất khác nhau.



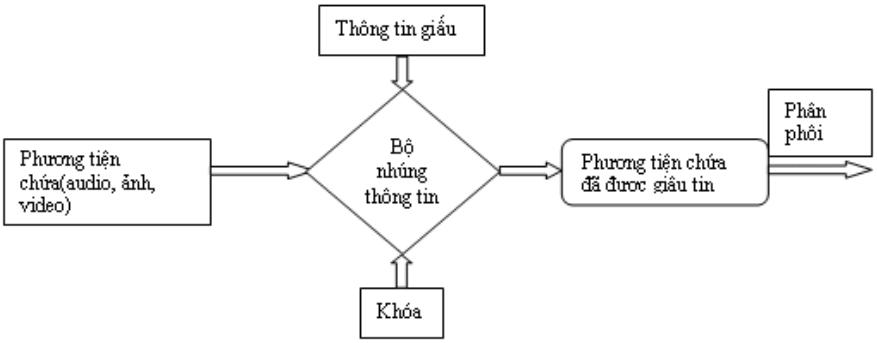
Hình 2.1: Hai lĩnh vực chính của kỹ thuật giấu thông tin.

- Kỹ thuật giấu thông tin bí mật (Steganography): với mục đích đảm bảo an toàn và bảo mật thông tin tập trung vào các kỹ thuật giấu tin để có thể giấu được nhiều thông tin nhất. Thông tin mật được giấu kỹ trong một đối tượng khác sao cho người khác không phát hiện được.

- Kỹ thuật giấu thông tin theo kiểu đánh giấu (watermarking) để bảo vệ bản quyền của đối tượng chứa thông tin thì lại tập trung đảm bảo một số các yêu cầu như đảm bảo tính bền vững… đây là ứng dụng cơ bản nhất của kỹ thuật thuỷ vân số.

2.4.1.3. Mô hình kỹ thuật giấu thông tin cơ bản

Giấu thông tin vào phương tiện chứa và tách lấy thông tin là 2 quá trình trái ngược nhau và có thể mô tả qua sơ đồ khối của hệ thống như sau:



Hình 2.2:Lược đồ chung cho quá trình giấu tin.

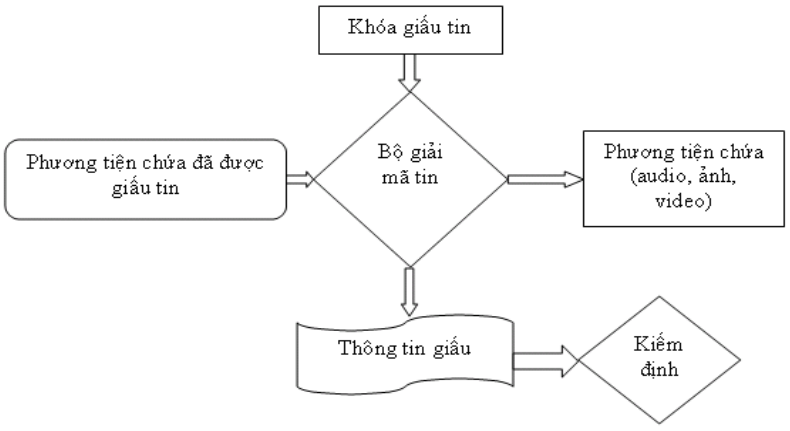
- Thông tin cần giấu tuỳ theo mục đích của người sử dụng, nó có thể là thông điệp (với các tin bí mật) hay các logo, hình ảnh bản quyền.

- Phương tiện chứa: các file ảnh, text, audio… là môi trường để nhúng tin

Bộ nhúng thông tin: là những chương trình thực hiện việc giấu tin

Đầu ra: là các phương tiện chứa đã có tin giấu trong đó

- Tách thông tin từ các phương tiện chứa diễn ra theo quy trình ngược lại với đầu ra là các thông tin đã được giấu vào phương tiện chứa.



Hình 2. 3: Lược đồ chung cho quá trình giải mã

Hình vẽ trên chỉ ra các công việc giải mã thông tin đã giấu. Sau khi nhận được đối tượng phương tiện chứa có giấu thông tin, quá trình giải mã được thực hiện thông qua một bộ giải mã tương ứng với bộ nhúng thông tin cùng với khoá của quá trình nhúng. Kết quả thu được gồm phương tiện chứa gốc và thông tin đã giấu.

Trong đề tài này tác giả đã xây dựng một chương trình giấu thông tin vào trong ảnh sử dụng kỹ thuật NSAS. Nhúng dữ liệu bằng cách dịch chuyển các biểu đồ tần số. Xây dựng một biểu đồ ảnh của ảnh gốc để có 1 điểm cực đại và 1 đểm cực tiểu, sau đó các dữ liệu nhúng bằng cách di chuyển biểu đồ tần số. Kỹ thuật này tìm được vị trí Max và chỉ khi nhập đúng vị trí Max này thì thông tin mới được tách ra.

CHƯƠNG 3

XÂY DỰNG THUẬT TOÁN GIẤU TIN VÀ SỬ DỤNG KỸ THUẬT NHẬN DẠNG BẢN RÕ DÒ TÌM KHÓA

Trong chương 1 và chương 2 tác giả đã trình bày các kỹ thuật cần thiết cho việc nhận dạng ngôn ngữ tự nhiên và ứng dụng vào dò tìm khóa trong phân tích mật mã. Trong chương này tác giả sẽ trình bày thuật toán trong một ngôn ngữ cụ thể. Đó là : Phân loại ngôn ngữ Tiếng Anh và bản rõ ngẫu nhiên. Sau đó tác giả đi xây dựng thuật toán giấu thông tin trong ảnh kết hợp với kỹ thuật nhận dạng để tìm khóa. Cả 2 thuật toán trên tác giả viết trên ngôn ngữ Matlab 7.7.

3.1. Bài toán

Cho một mẫu văn bản x nào đó thuộc ngôn ngữ Anh hoặc dãy ngẫu nhiên (chữ cái hoặc chuyển sang mã cơ số 16) nhưng chưa biết x thuộc loại nào trong 2 loại nêu trên. Hãy xác định xem x thuộc ngôn ngữ cụ thể nào.

Ta ký hiệu A_0 là lớp đại diện cho ngôn ngữ không đọc được, A_1 đại diện cho lớp ngôn ngữ tiếng Anh. Vậy bài toán là hãy xác định xem x thuộc lớp đại diện nào trong 2 lớp vừa nêu? (ở đây số k đã được biết trước là k=2).

Nội dung thuật toán gồm 2 phần: **Phần off-line** (xây dựng cơ sở dữ liệu để máy học) và **phần on-line** (phần nhận biết trực tiếp). Sau đây là nội dung cụ thể của 2 phần đó

3.2. Thuật toán nhận dạng

3.3.1. Phần off-line.

Xây dựng ước lượng ma trận các xác suất chuyển trạng thái P của mô hình Markov ứng với các ngôn ngữ tự nhiên tiếng Anh. Trong ước lượng này, tác giả chọn độ dài mẫu xấp xỉ 10.000 ký tự bỏ qua dấu gián cách từ dấu chấm (.), dấu phẩy (,), ..., không phân biệt chữ in hoa và chữ in thường. Nó là tổng hợp các loại văn bản thuộc các chuyên ngành khác nhau: Chính trị, kinh tế, văn học, tin học, địa lý, quân sự, thể thao, ngoại giao, lịch sử, y tế, giáo dục, pháp luật. Mỗi loại được chọn xấp xỉ 1000 ký tự.

Tính tần số bộ đôi móc xích của dãy X, tức là ta tính số lần xuất hiện các cặp chữ cái la tinh của dãy đó. Giả sử, tần số các chữ cái đó lần lượt được ký hiệu là $X=(x_{ij})$

$$x_{t,t+1}(i,j)=\begin{cases}1 & \text{Nếu cặp (i,j) xuất hiện tại thời điểm (vị trí) t và t+1}\\ 0\end{cases}$$

Trong trường hợp khác. với i,j = a,b,...,z.

với $t=\overline{1,N-1}$

Rõ ràng ta có $n_{ij}=\sum_{t=1}^{n-1}x_{t,t+1}(i,j)$ là số lần xuất hiện cặp ký tự (i,j) đứng kề nhau.

Ta có xác suất xuất hiện cặp (i,j) là P_{ij} . Do các ngôn ngữ có tính dừng nên xác suất chuyển P_{ij} không phụ thuộc vào t từ là:

$$x_{t,t+1}(i,j)=\begin{cases}1 & \text{với xác suất } P_{ij}\\ 0 & \text{với xác suất } q_{ij}=1-P_{ij}\end{cases}$$

Bây giờ ta tính kỳ vọng toán học của đại lượng ngẫu nhiên n_{ij}, ta có kỳ vọng

$$E[n_{ij}]=E\left[\sum_{t=1}^{n-1}X_{t,t+1}(i,j)\right]=\sum_{t=1}^{n-1}(E.X_{t,t+1}(i,j))=(n-1)P_{ij}\quad\forall i,j;1\le i,j\le 26$$

Vì vậy, rõ ràng là $\frac{n_{ij}}{n-1}$ là ước lượng không chệch đối với tham số P_{ij} $\forall$ P_{ij}. Bây giờ, giả sử cho i cố định, bằng lý luận hoàn toàn tương tự, ta có $\frac{n_{ij}}{n-1}$ là ước lượng không chênh của P_{ij} với i cố định. Trong đó $n_i=\sum_{j=1}^{26}n_{ij}$; i = 1,2,...,26.

Tuy nhiên có tồn tại một ước lượng có chênh tốt hơn ước lượng không chệch theo nghĩa sai số trung bình bình phương nhỏ nhất. Ước lượng đó là

$\hat{P}_{ij} = \frac{n_{ij} + c}{n_i + m.c}$ trong đó c là hằng số thường được chọn c=0,5 hoặc c=1/m còn m=26 (Số chữ cái của bảng ngôn ngữ). Nếu lấy c=1/m ta có

$$\hat{P}_{ij} = \frac{n_{ij} + 0{,}0385}{n_i + 1}; i, j = 1,2,...,26$$

Đó là ước lượng có chênh nhưng với sai số trung bình bình phương bé nhất của xác suất chuyển P_{ij} của ma trận chuyển P trong mô hình Markov của ngôn ngữ tự nhiên Anh. Kết quả tính $\hat{P}_{ij}$ được cho ở bảng ứng sau:

BẢNG 3.1. ƯỚC LƯỢNG BỘ ĐÔI MÓC XÍCH TIẾNG ANH (A_1)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
A	4	21	29	53	2	11	31	4	33	4	6	65	23	171	2	14		87	77	127	7	25	8	1	9	1
B	13				56				8	2		23			11			15	4	2	13				15	
C	33		7	1	71			36	18		10	9	1		51	3		10		29	11				3	
D	41	16	10	5	66	18	3	9	59		1	4	15	6	16	4		21	18	53	21	5	15		3	
E	88	20	58	126	57	41	20	16	51	1	4	56	57	154	35	7	6	194	153	66	9	26	31	13	5	
F	19	3	5	1	19	21	1	3	30	2		11	1		52			26	8	49	7	3	3		2	
G	21	4	3	2	35	1	3	15	18			6	1	4	21	1	1	23	9	21	9		5		1	
H		1	4		271	5	1	6	58				3	2	44	1		3	12	18	6		5		3	
I	40	7	53	24	25	9	11	3			2	38	26	212	58	2	1	46	82	120	1	3		4		3
J	3				5				1						46						3					
K	2				11				13					2	2				6	2	1		2		1	
L	46	2	5	12	65	7	5	2	44	1	1	54	2	2	26	1	1	2	16	23	9		1		34	
M	52	14	1		67			3	1				7	1	17	9	1	2	12	3	8		1		2	
N	42	10	48	126	63	19	107	12	32	1	6	6	9	7	54	7	1	7	47	132	6	5	14		13	
O	7	12	14	17	5	97	3	5	14			19	2	136	13	3		95	23	42	58	16	28		4	1
P	19	1			37			4	8			15	1		28	9		33	14	7	6					
Q																					17					
R	86	8	16	24	177	4	8	8	78	1	10	6	26	16	60	4		24	38	56	6	11	4		29	
S	66	9	17	9	76	13	1	47	78	3		8	11	12	58	7	6	9	48	121	35	1	28		4	
T	59	23	7	1	79	6	2	332	133	1		14	10	6	79	7		51	50	56	21	3	30		25	
U	12	5	9	6	9	1	6		9		1	20	5	31	2	5		48	40	31		3			1	
V	7				77				29						5										3	
W	38	1	1		39			33	36			4	1	8	15				4	2			1			
X	1		2			1			3						1	5				4			1			
Y	15	5	4	2	7	12	2	6	10			3	7	5	20	8		4	16	31			5			
Z	1				4																					

Từ số liệu ở bảng A_1

$$A_1 = (a_{ij}^{(1)}) \quad i, j = 1,2,...,26$$

ta chuyển sang các bảng B_1 theo công thức sau:

$$B_1 = (b_{ij}^{(1)}) \quad i, j = 1,2,...,26$$

Trong đó:

$$b_{ij}^{(1)} = \begin{cases} \left[7 \lg \dfrac{14,79}{a_{ij}^{(1)}} \right] & \text{if } a_{ij}^{(1)} > 0 \\ 11 & \text{if } a_{ij}^{(1)} = 0 \end{cases} \qquad i, j = 1,2,...,26$$

Trong đó lg(.): là lôgarit cơ số 10

[x] = số nguyên lớn nhất nhưng bé hơn hoặc bằng x.

Hệ số k = 7 này là là kết quả thực nghiệm giúp cho việc nhận dạng giữa các lớp được tốt hơn.

Gọi $A = (a_{ij})_{26x26}$ với $a_{ij} = 1/26 \ \forall \ i,j = 1,2,\ldots,26$. Ma trận của dãy ngẫu nhiên.

Như vậy, mỗi phần tử $A_0 = (a_{ij}^{(0)})$ của ma trận bộ đôi của dãy ngẫu nhiên là Hệ số $14,79 \approx \dfrac{10.000}{26 * 26}$, lấy 2 chữ số thập phân sau dấu phẩy.

Ví dụ 3.1: [-1,5] = -2 [-1,95] = -2 [-1] = -1
[1,5] = 1 [3] = 3 [0,3] = 0

ta có bảng như sau:

BẢNG 3.2: ƯỚC LƯỢNG ĐỐI SÁNH CỦA TIẾNG ANH VỚI MẪU NGẪU NHIÊN (B1)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
A	4	-2	-3	-4	6	0	-3	4	-3	4	2	-5	-2	-8	6	0	11	-6	-6	-7	2	-2	1	8	1	8
B	0	11	11	11	-5	11	11	11	1	6	11	-2	11	11	0	11	11	-1	4	6	0	11	11	11	-1	11
C	-3	11	2	8	-5	11	11	-3	-1	11	1	1	8	11	-4	4	11	1	11	-3	0	11	11	11	4	11
D	-4	-1	1	3	-5	-1	4	1	-5	11	8	4	-1	2	-1	4	11	-2	-1	-4	-2	3	-1	11	4	11
E	-6	-1	-5	-7	-5	-4	-1	-1	-4	8	4	-5	-5	-8	-3	2	2	-8	-8	-5	1	-2	-3	0	3	11
F	-1	4	3	8	-1	-2	8	4	-3	6	11	0	8	11	-4	11	11	-2	1	-4	2	4	4	11	6	11
G	-2	4	4	6	-3	8	4	-1	-1	11	11	2	8	4	-2	8	8	-2	1	-2	1	11	3	11	8	11
H	11	8	4	11	-9	3	8	2	-5	11	11	11	4	6	-4	8	11	4	0	-1	2	11	3	11	4	11
I	-4	2	-4	-2	-2	1	0	4	11	11	6	-3	-2	-9	-5	6	8	-4	-6	-7	8	4	11	4	11	4
J	4	11	11	11	3	11	11	11	8	11	11	11	11	11	-4	11	11	11	11	11	4	11	11	11	11	11
K	6	11	11	11	0	11	11	11	0	11	11	11	11	6	6	11	11	11	2	6	8	11	6	11	8	11
L	-4	6	3	0	-5	2	3	6	-4	8	8	-4	6	6	-2	8	8	6	-1	-2	1	11	8	11	-3	11
M	-4	0	8	11	-5	11	11	4	8	11	11	11	2	8	-1	1	8	6	0	4	1	11	8	11	6	11
N	-4	1	-4	-7	-5	-1	-7	0	-3	8	2	2	1	2	-4	2	8	2	-4	-7	2	3	0	11	0	11
O	2	0	0	-1	3	-6	4	3	0	11	11	-1	6	-7	0	4	11	-6	-2	-4	-5	-1	-2	11	4	8
P	-1	8	11	11	-3	11	11	4	1	11	11	-1	8	11	-2	1	11	-3	0	2	2	11	11	11	11	11
Q	11	11	11	11	11	11	11	11	11	11	11	11	11	11	11	11	11	11	11	11	-1	11	11	11	11	11
R	-6	1	-1	-2	-8	4	1	1	-6	8	1	2	-2	-1	-5	4	11	-2	-3	-5	2	0	4	11	-3	11
S	-5	1	-1	1	-5	0	8	-4	-6	4	11	1	0	0	-5	2	2	1	-4	-7	-3	8	-2	11	4	11
T	-5	-2	2	8	-6	2	6	-10	-7	8	11	0	1	2	-6	2	11	-4	-4	-5	-2	4	-3	11	-2	11
U	0	3	1	2	1	8	2	11	1	11	8	-1	3	-3	6	3	11	-4	-4	-3	11	4	11	11	8	11
V	2	11	11	11	-6	11	11	11	-3	11	11	11	11	11	3	11	11	11	11	11	11	11	11	11	4	11
W	-3	8	8	11	3	11	11	-3	-3	11	11	4	8	1	-1	11	11	11	3	6	11	11	8	11	11	11
X	8	11	6	11	11	8	11	11	4	11	11	11	11	11	8	3	11	11	11	4	11	11	8	11	11	11
Y	-1	3	4	6	2	0	6	2	1	11	11	4	2	3	-1	1	11	4	-1	-3	11	11	3	11	11	11
Z	8	11	11	11	4	11	11	11	11	11	11	11	11	11	11	11	11	11	11	11	11	11	11	11	11	11

3.3.2. Phần on-line

Giả sử X là mẫu nào đó $X = x_1x_2..x_N$ với $x_i \in \{a,b,..,z\}$, i=1,2,..,N. Vấn đề đặt ra X thuộc ngôn ngữ Tiếng Anh hay là một dãy ngẫu nhiên nào đó?

Ta tiến hành như sau:

Step1: Tính tần số bộ đôi móc xích của dãy X, tức là ta tính số lần xuất hiện các cặp chữ cái la tinh của dãy đó. Giả sử, tần số các chữ cái đó lần lượt được ký hiệu là $F= (f_{ij})$ với i,j = a,b,...,z.

$$f_{t,t+1}(i,j) = \begin{cases} 1 & \text{Nếu cặp (i,j) xuất hiện tại thời điểm (vị trí) t và t+1} \\ 0 & \text{Trong trường hợp khác với } t = \overline{1, N-1} \end{cases}$$

Step2: i=0

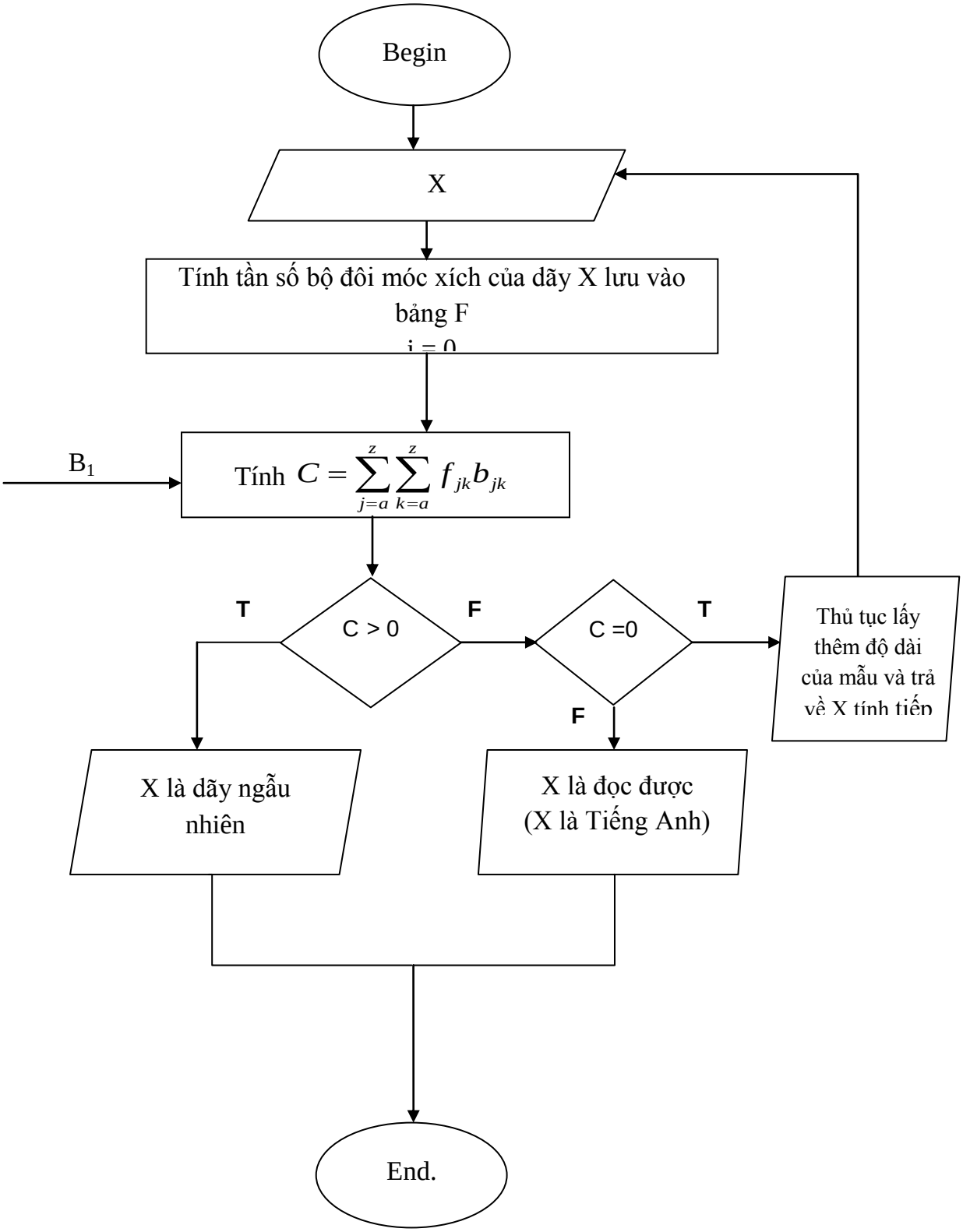
Step3: Tính Tr(F.B) bằng công thức

$$C = \sum_{j=a}^{z}\sum_{k=a}^{z} f_{jk}b_{jk}$$

Step4: Nếu C > 0 thì thuật toán dừng và kết luận X thuộc lớp dãy ngẫu nhiên

Step5: Nếu C = 0 thì thuật toán thông báo yêu cầu nhập thêm độ dài của mẫu cần kiểm tra và quay lại **Step1**.

Step6: Nếu C = 0 sai thì thuật toán kết thúc và thông bào X thuộc bản rõ Tiếng Anh.



Hình 3.1. Sơ đồ khối của thuật toán

3.3. Thuật toán NSAS

Đầu vào: 1 ảnh cấp xám, chuỗi thông điệp.

Đầu ra: Ảnh giấu tin

Bước 1: Quét tất cả các ảnh và xây dựng biểu đồ tần suất $H_1(x)$, $x \in |0,255|$. Trong histogram, có điểm cực đại a, điểm cực tiểu b. Không mất tính khái quát, giả sử a < b.

Bước 2: Thiết lập k = 0. Giá trị k được sử dụng để cho biết số bít dữ liệu nhúng.

Bước 3: Quét tất cả các ảnh 1 lần nữa. Nếu quét được giá trị điểm ảnh a, trích 1 bit dữ liệu từ S, thiết lập k = k +1 và tiếp tục bước 4 để nhúng dữ liệu S, nếu không, thực hiện bước 5.

Bước 4: Nếu bit dữ liệu là 1, thì thiết lập giá trị điểm ảnh quét được là a+1, nếu không có thay đổi gì cho những điểm ảnh này, quay lại bước 3 tiếp tục quá trình nhúng.

Bước 5: Nếu tất cả các giá trị điểm ảnh quét được nằm trong khoảng (a, b), thì cộng các giá trị điểm ảnh đó thêm 1. Ghi lại vị trí các điểm ảnh có giá trị điểm ảnh = b.

Thuật toán tách tin:

Đầu vào: Ảnh giấu tin.

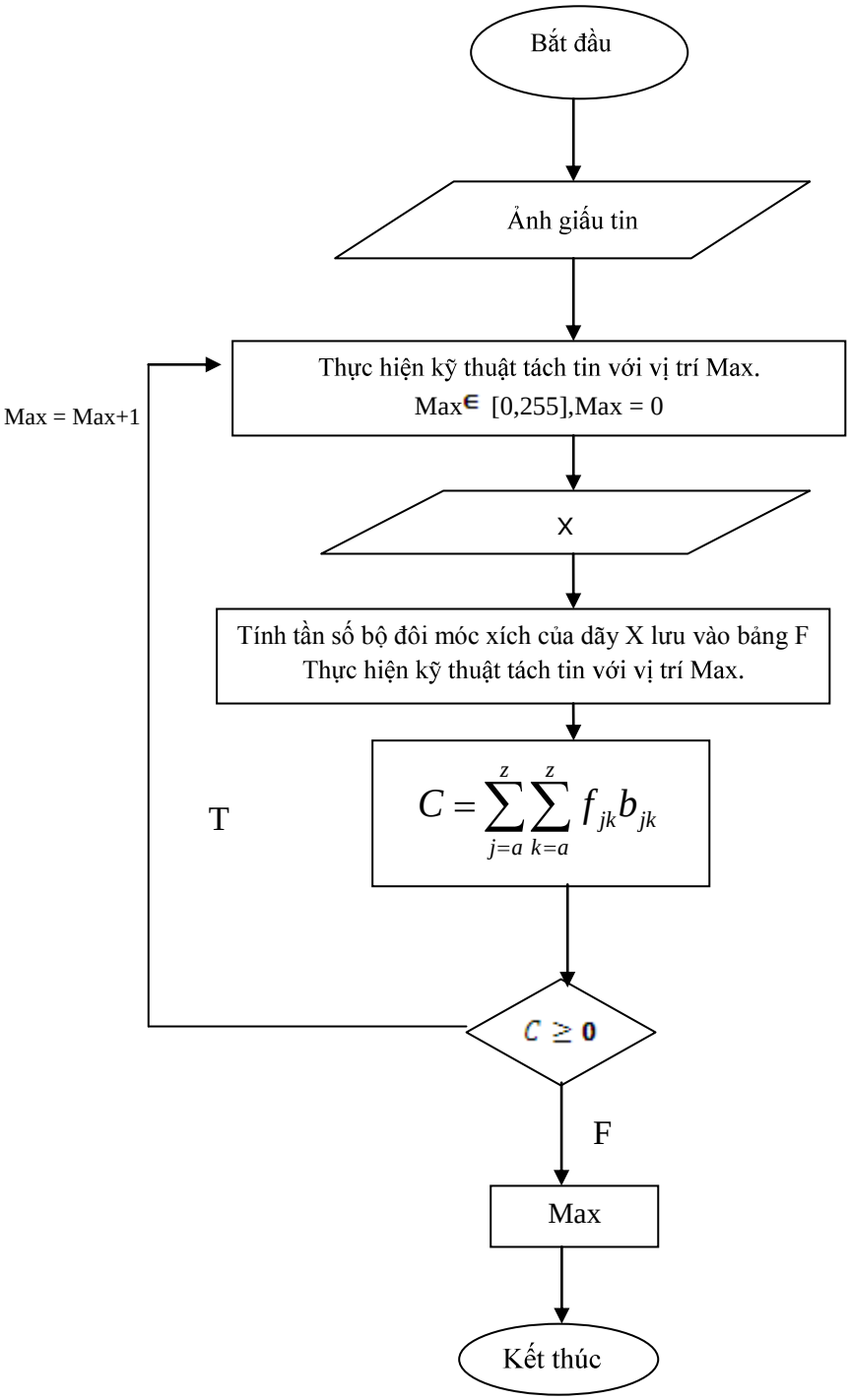
Đầu ra: Ảnh khôi phục và chuỗi thông điệp.

Bước 1: Thiết lập k = 0.

Bước 2: Quét tất cả các ảnh theo thứ tự như trong quá trình nhúng. Nếu quét được giá trị là a, thì đặt k = k+1 và tách bit 0 khỏi a. Nếu quét được giá trị là a+1, thì k = a+1 và tách bit 1 ra khỏi a. Nếu giá trị quét nằm trong khoảng (a, b) thì các giá trị điểm ảnh quét được trừ đi 1. Nếu vị trí các điểm ảnh được ghi trong bản đồ L, thì thiết lập giá trị các điểm ảnh quét được là b.

Bước 3: Lặp lại bước 2 cho đến khi k = |S|

3.3. Ứng dụng nhận dạng ngôn ngữ tiếng anh tìm khóa của thuật toán NSAS.



Hình 3.2. Sơ đồ khối của thuật toán

3.4. Một số ví dụ

Ví dụ 3.1: Ta kiểm tra mẫu văn bản:

Cho X =Thoong xoth phaart iuoof ghtfc ytrung phoith ghuiyr jokjp.

Vậy với thuật toán thì nó nhận ra như thế nào. Quá trình thực hiện nhý sau:

Step 1: Tính tần số bộ đôi móc xích, được bảng sau (Ký hiệu là bảng F)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
A	1																	1								
B																										
C																									1	
D																										
E																										
F			1				1																			
G								2								1										
H	1						1								2	1				1	1					
I																				1	1				1	
J															1	1										
K										1																
L																										
M																										
N							2																			
O						1			1		1			1	2					1						
P								2																		
Q																										
R										1										1	1					
S																										
T						1		3	1									1								
U									1					1	1											
V																										
W																										
X															1											
Y																		1		1						
Z																										

Sử dụng bảng B_1: tính $C = \sum_{j=a}^{z} \sum_{k=a}^{z} f_{jk} b_{jk}$

Dòng	Các phép tính tương ứng trên dòng của bảng B_1 và F
A	1.4+ 1.(-6) +
C	1.(4) +
F	1.3 + 1.8 +
G	2.(-8) +1.8+
H	1.11 + 1.8 + 2.(-4) +1.(-4) +1.(-2) +
I	1.(-7) + 1.8 + 1.11 +
J	1.11+1.11 +
K	1.11+
N	1.(-7) +
O	1.3 + 1.0 + 1.11 + 1.(-7) + 0.2 + 1.(-4) +
P	1.11 +
R	1.8 + 1.(-5) + 1.2 +
T	1.(-6) + 1.(-10) + 1.(-7) + 1.(-4) +
U	1.1 + 1.(-3) +1.(6) +
X	1.11 +
Y	1.4 + 1.(-3) = 57

Vậy C=26 > 0. Suy ra X là văn bản không đọc được.

***Ví dụ 3.3*:** Cho văn bản X= In cryptanalysis, how can a computer program recognize when it has discovered all or part of the secret message:

Tính tần số bộ đôi móc xích, được bảng sau (Ký hiệu là bảng F)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
A			1				1					2	1	2				1	1							
B																										
C	1														3			2								
D	1								1																	
E			2	1										1				2	2	1			1			
F																				1						
G					1									1				1								
H	1				2										1											
I														1					1	1						1
J																										
K																										
L												1			1										1	
M					1											1										
N	21								2					1												
O						1	21						1					1				1	1			
P	1																	1		1	1					
Q																										
R	1				3											3					1				1	
S	1		1	1	1			1	1										1							
T	1				1			2					1		1											
U																				1						
V					1																					
W			1					1																		
X																										
Y																1			1							
Z					1																					

Sử dụng bảng B_1: tính $C = \sum_{j=a}^{z} \sum_{k=a}^{z} f_{jk} b_{jk}$

Dòng	Các phép tính tương ứng trên dòng của bảng B_1 và F
A	1.(-3) + 1.(-3) + 2.(-5) + 1.(-2) + 2.(-8)+1.(-6)+1.(-6)+
C	1.(-3) + 3.(-4) + 2.1 +
D	1.(-4) + 1.(-5) +
E	2.(-5) + 1.(-7) + 1.(-8) + 2.(-8) + 2.(-8) + 1.(-5) + 1.(-3)+
F	1.(-4) +
G	1.(-3) +1.(4)+1(-2) +
H	1.(11) + 2.(-9) + 1(-4) +
I	1.(-9) + 1.(-6) + 1.(-7) +1.(-6) +
L	1.(-4) +1.(-2) + 1. (-3) +
M	1.(-5) + 1.1 +
N	2.(-4) + 2.(-3) + 1.(2) +
O	1.(-6) + 2.4 + 1.6 + 1.(-6) + 1.(-4) + 1.(-5) +
P	1.(-1) + 1.11 +1.2+1.2+
R	1.(-6) + 3.(-8) + 3.4 +1.2+1.(-3)+
S	1.(-5) + 1.(-1) + 1.(-2) + 1.(-8)+1.(-1) +1.(-6)+1.(-3) +
T	1.(-5)+1.(-6)+ 2.(-10) + 1.0 + 1.(-6) +
U	1.(-3) +
W	1.8 + 1.(-3) +
Y	1.(-1) + 1.1 +
Z	1.4 = -271

Vậy C = -357 < 0. Suy ra X là văn bản Tiếng Anh.

3.5. Kết quả đạt được

Tác giả đã xây dựng chương trình viết bằng Malab 7.7 để thử nghiệm điểm kiểm định theo quy trình nhận dạng và tìm khóa như đã đề xuất.

Các mẫu thử được lựa chọn là tài liệu tiếng Anh thuộc các lĩnh vực: Chính trị, kinh tế, văn học, tin học, địa lý, quân sự, thể thao, ngoại giao, lịch sử, y tế, giáo dục, pháp luật; với độ dài khác nhau.

Độ chính xác trong các trường hợp lấy độ dài kiểm tra khác nhau

Độ dài (Ký tự)	Tiếng Anh			Dãy ngẫu nhiên		
	T.số	Đúng	Tỷ lệ	T.Số	Đúng	Tỷ lệ
50	200	198	99,5%	20	18	90%
60	200	199	99%	20	20	100%
70	200	200	100%	20	20	100%
80	200	200	100%	20	20	100%

Độ chính xác trong các trường hợp tìm khóa

Độ dài (Ký tự)	Ứng dụng nhận dạng tìm khóa	
	T.số	Tỷ lệ
50	150	100%
60	150	100%
70	150	100%
80	150	100%

3.6. Đánh giá thuật toán

Thuật toán dựa trên xích Markov cấp 1 hữu hạn trạng thái. Xây dựng ước lượng ma trận các xác suất chuyển trạng thái của mô hình Markov ứng với ngôn ngữ tự nhiên tiếng Anh. Tính tần số bộ đôi móc xích của dãy ký tự thuật toán nhận dạng được văn bản tiếng anh . Thuật toán này có khả năng mở rộng ra áp dụng cho việc nhận dạng cho mọi ngôn ngữ. Để nhận dạng các ngôn ngữ khác chúng ta cũng cấn xây dựng bảng tần số bộ đôi móc xích cho ngôn ngữ đó giống như bảng A1 và xây dựng các bảng đối sánh giữa các ngôn ngữ giống như B1.

Trong phần ứng dụng nhận dạng ngôn ngữ cho việc tìm khóa, do thời gian nghiên cứu hạn chế tác giả chỉ thực hiện trên thuật toán cụ thể với 1 khóa duy nhất . Việc ứng dụng nhận dạng cho việc tìm khóa sẽ làm giảm bớt thời gian cho người thám mã giúp người thám mã có thể thực hiện công việc tốt hơn.

KẾT LUẬN

Trong luận văn của mình, em đã tập trung nghiên cứu các kỹ thuật nhận dạng ngôn ngữ tự nhiên, tìm hiểu phương pháp giấu thông tin trong ảnh và ứng dụng nhận dạng để tìm khóa với thuật toán giấu tin cụ thể luận văn đã đạt được các kết quả sau:

- Trình bày những vấn đề cơ bản nhất của kỹ thuật nhận dạng nói chung, các hướng nghiên cứu hiện nay trên thế giới.
- Trình bày vấn đề cơ bản về giấu tin trong ảnh và xây dựng thuật toán giấu tin trong ảnh.
- Xây dựng được thuật toán nhận dạng nhanh bản rõ ngôn ngữ tự nhiên Anh.
- Xây dựng được thuật toán ứng dụng nhận dạng ngôn ngữ để tìm khóa.
- Các thuật toán được thể hiện trên máy PC bằng ngôn ngữ Matlap, kết quả thử nghiệm tốt, nhanh.

*** Những vấn đề có thể nghiên cứu tiếp tục:**

Từ thuật toán này ta có thể mở rộng ra tính toán nhận dạng bản rõ Tiếng Pháp, Tiếng Đức… và các ngôn ngữ la tinh và phi la tinh khác. Ta thấy vấn đề khó khăn nhất của việc xây dựng thuật toán chính là phần offline hay là xây dựng được ma trận tần số bộ đôi móc xích (B1) của ngôn ngữ cần nhận dạng (cái này đòi hỏi sự hiểu biết về ngôn ngữ để chọn lựa các mẫu tính toán và thơi gian công sức lớn). Vì độ chính xác càng cao thì độ chính xác của bảng đối sánh giữa các ngôn ngữ càng cao (B1). Khi đó, thuật toán tính toán cho ra một kết quả tốt hơn chỉ với dãy mẫu ngắn; giúp cho hệ thống chạy nhanh khi với số lượng mẫu khổng lồ.

TÀI LIỆU THAM KHẢO

Tiếng Việt

[1]. Lương Mạnh Bá, Nguyễn Thanh Thuỷ, *Nhập môn xử lý ảnh số*, Nhà xuất bản khoa học và kỹ thuật, 1999. tr.154-170.

[2]. Hồ Văn Canh, Phạm Quốc Doanh, *Thuật toán nhận dạng các ngôn ngữ tự nhiên*, 2002. tr. 3-20.

Tiếng Anh

[5]. Anderson, Roland. 1989. Recognizing complet and partial plaintext. *Cryptologia.* 13(2):161-166.

[6]. Anderson, T.W.and Leo A.Goodman.1957. Statistical inference about Markov chains, *Annals of Mathematical Statistics,*28: 89-110

[7]. Andrew R. Web . 2002. John Wiley & Sons, Ltd. *Statistical Pattern Recognition, Second Edition.*

[8]. R.GaneSan, AlanT.Sherman(1993), "Statiscal Techniques for language Recognition An introduction and Guide for Cryptanalysts. 121-126

[9]. Seber, George Arthur Fredederick. 2008. John Wiley & Sons, Inc. "MISCELLANEOUS INEQUALITIES", *A Matrix Handbook for Statisticians:*296-298

[10]. Richard O Duda, Peter E Hart, David G Stork. Wiley-interscience. **"**Bayesian decision theory", *Pattern Classification, Second Edition*: 39-78

LỜI CẢM ƠN

Em xin chân thành cảm ơn các thầy cô trong bộ môn tin cũng như các thầy cô trong trường đã trang bị cho em những kiến thức cơ bản cần thiết để em có thể hoàn thành báo cáo.

Xin chân thành cảm ơn các anh, các chị và các bạn sinh viên K12 trường Đại học Dân Lập Hải Phòng đã luôn động viên, giúp đỡ và nhiệt tình chia sẻ với tôi những kinh nghiệm học tập, công tác trong suốt khoá học.

Đặc biệt em xin bày tỏ lòng biết ơn sâu sắc đến **TS.Hồ Văn Canh** đã tận tình giúp đỡ em hình thành, nghiên cứu và hoàn chỉnh luận văn.

Mặc dù đã có nhiều cố gắng, song do sự hạn hẹp về thời gian, điều kiện nghiên cứu và trình độ, luận văn không tránh khỏi những khiếm khuyết. Em chân thành mong nhận được sự đóng góp ý kiến của các thầy, cô giáo.

Một lần nữa em xin cảm ơn!

Hải Phòng, tháng 11 năm 2012

Người thực hiện luận văn

Nguyễn Doãn Tùng

MỤC LỤC

DANH MỤC BẢNG

DANH MỤC HÌNH