Trường Đại Học Công Nghệ Thông Tin

KHOA MẠNG & TRUYỀN THÔNG

LÝ THUYẾT THÔNG TIN

Bùi Văn Thành

thanhbv@uit.edu.vn

Tháng 7 năm 2013

CHƯƠNG 4

LÝ THUYẾT MÃ
MÃ HÓA NGUỒN

MÃ HÓA NGUỒN TIN

1. **Khái niệm mã và điệu kiện thiết lập mã.**
2. **Điều kiện để mã phân tách được.**
3. **Mã thống kê tối ưu.**

GIỚI THIỆU

- Trong các hệ thống truyền tin, nguồn nhận thường tập hợp các tin mà bên phát dùng để lập nên các bản tin.

- Các tin thường sẽ được ánh xạ (mã hóa) thành một dạng biểu diễn khác thuận tiện hơn để phát đi.
- ***Ví dụ***: Xét một nguồn tin A={a,b,c,d} chúng ta có thể thiết lặp các cặp ánh xạ như sau từ A vào tập các chuỗi {0,1}

$$a \rightarrow 00 \qquad c \rightarrow 10$$

$$b \rightarrow 01 \qquad d \rightarrow 11$$

Vậy để phát đi bản tin cbab, ta phải phát đi tập tin 10010001. Khi nguồn nhận nhận được chuỗi này, thì sẽ xác định được bản tin là cbab.

MÃ HIỆU VÀ NHỮNG THÔNG SỐ CƠ BẢN

- ***Mã hiệu (code):*** Mã hiệu là tập hợp hữu hạn các ký hiệu và phép ánh xạ các tin/bản tin của nguồn tin thành các dãy ký hiệu tương ứng. Tập các ký hiệu và phép ánh xạ này thường sẽ đáp ứng các yêu cầu mà hệ thống truyền tin đặt ra.
- ***Cơ số mã***: Tập các ký hiệu mã dùng để biểu diễn gọi là ***bảng ký hiệu mã***, còn số các ký hiệu thì gọi là **cơ số mã** (m). Mã nhị phân m=2, mã tam phân m=3…
- ***Mã hóa (Encoding)***: Mã hóa là quá trình dùng các ký hiệu mã để biểu diễn các tin của nguồn.
  - Biến đổi nguồi tin thành mã hiệu, biến đổi tập tin này thành tập tin khác có đặc tính thống kê theo yêu cầu.
- Quá trình ngược lại của mã hóa được gọi là ***giải mã (Decoding).***

MÃ HIỆU VÀ NHỮNG THÔNG SỐ CƠ BẢN

- ***Từ mã (code wod), bộ mã***: ***Từ mã*** là chuỗi kí hiệu mã biểu diễn cho tin của nguồn. Tập tất cả các từ mã tương ứng với các tin của nguồn được gọi là ***bộ mã***.
  - Vì vậy có thể nói mã hóa là một phép biến đổi **một – một** giữa ***một tin của nguồn*** và ***một từ mã của bộ mã***.
  - Trong một số trường hợp người ta không mã hóa mỗi tin của nguồn mà mã hóa một bản tin hay khối tin. Lúc này chúng ta có khái niệm ***mã khối***.
  - Các từ mã thường được ký hiệu: ***u,v,w***.
- ***Chiều dài từ mã*** là số ký hiệu trong một từ mã (***l***).

MÃ HIỆU VÀ NHỮNG THÔNG SỐ CƠ BẢN

- ***Chiều dài trung bình của bộ mã*** ($\bar{l}$): $\bar{l} = \sum_{i=1}^{n} p(x_i) l_i$

  $p(x_i)$: xác suất xuất hiện tin x_i của nguồn U được mã hóa.

  n : số từ mã tương ứng số tin của nguồn

  l_i : chiều dài từ mã tương ứng với tin x_i của nguồn.

- ***Phân loại mã***:
  - Một bộ mã được gọi là ***mã đều*** nếu các từ mã của bộ mã có chiều dài bằng nhau.
  - Một bộ mã đều có cơ số mã ***m*** , chiều dài từ mã ***l*** và số lượng từ mã ***n*** bằng với m^l thì được gọi là ***mã đầy***, ngược lại thì gọi là ***mã vơi***.
  - ***Ví dụ***: Cho bảng ký hiệu mã A={0,1}, thì bộ mã X_1 ={0,10,11} là mã không đều, bộ mã X_2 ={00,10,11} là mã đều nhưng vơi, còn bộ mã X_3 ={00,01,10,11} là mã đều và đầy.

ĐIỀU KIỆN PHÂN TÁCH MÃ

Ví dụ:

- Xét bộ mã $\mathbf{X_1}$**= {0,10,11}** mã hóa cho nguồn **A = {a,b,c}**.
- Giả sử bên phát phát đi bản tin **x = abaac**, lúc đó chuỗi từ mã tương ứng được phát đi là **y = 0100011**.
- Vấn đề: bên nhận nhận được y, làm sao tìm x?
- Để làm được quá trình này, bên nhận phải thực hiện một quá trình gọi là ***tách mã***. Với chuỗi ký hiệu y trên, bên nhận chỉ có 1 khả năng tách mã: **0 | 10 | 0 | 0 | 11**.

ĐIỀU KIỆN PHÂN TÁCH MÃ (TT)

- Xét bộ mã $\mathbf{X_2}$**= {0,10,01}** mã hóa cho nguồn A trên.
- Giả sử bên nhận nhận được chuỗi **y = 01010**.
- Với chuỗi ký hiệu y trên, bên nhận có thể có 3 khả năng tách mã: **0 | 10 | 10**; **01 | 0 | 10**; **01 | 01 | 0**. Vì vậy, bên nhận không biết chính xác bên phát đã phát mẫu tin **abb**, hay **cab**, hay **cca**.
- Một mã như vậy không phù hợp cho việc tách mã và được gọi là ***mã không tách được*** (*uniquely undecodable code*).
- Vì vậy, điều kiện để một mã được gọi là ***mã phân tách được*** (*uniquely decodable code*) là không tồn tại dãy từ mã này trùng với dãy từ mã khác của cùng bộ mã.

ĐIỀU KIỆN PHÂN TÁCH MÃ (TT)

- Xét bộ mã $\mathbf{X_3}$= **{010,0101,10100}** mã hóa cho nguồn **A** trên.
- Giả sử bên nhận nhận được chuỗi **y = 01010100101**.
- Với chuỗi ký hiệu y trên, bên nhận chỉ có 1 khả năng tách mã: **0101 | 010 | 0101**. Nhưng việc giải khó khăn hơn so với bộ mã $\mathbf{X_1}$.
- Để nhận biết một bộ mã có phân tích được không, người thường dùng một công cụ được gọi là ***bảng thử mã***.

BẢNG THỬ MÃ

- Bản chất của bảng thử mã là phân tích những ***từ mã dài*** thành những ***từ mã ngắn*** đi đầu.
- Từ mã dài u_1 có thể phân tích thành $v_{11}v_{12}\ldots v_{1k}w_{11}$ trong đó $v_{11}, \ldots v_{1k}$ là các từ mã ngắn, còn w_{11} là phần còn lại của u_1.
- Nếu w_{11} cũng là một từ mã thì bộ mã này là không phân tách được vì chuỗi $v_{11}v_{12}\ldots v_{1k}w_{11}$ có ít nhất hai cách phân tách thành các từ mã, đó là đó là u_1 và $v_{11},v_{12},\ldots,v_{1k},w_{11}$.
- Còn nếu ngược lại, w_{11} không là từ mã thì chúng ta dùng nó để xét tiếp. Trong lần xét tiếp chúng ta xét xem mỗi w_{11}này có là tiếp đầu ngữ của các từ mã hay không, nếu đúng với một từ mã nào đó, giả sử là u_2, thì từ mã này sẽ có dạng $w_{11}v_{21}\ldots v_{2l}w_{22}$ trong đó $v_{21}\ldots v_{2l}$ là các từ mã ngắn (l có thể bằng 0) còn w_{22} là tiếp vị ngữ còn lại.

BẢNG THỬ MÃ (TT)

- Lúc đó tồn tại dãy kí hiệu sau:

  $v_{11}v_{12}..v_{1k}w_{11}v_{21}..v_{2l}w_{22}..w_{(i-1)(i-1)}v_{i1}..v_{im}w_{ii}v_{(i+1)1}..v_{(i+1)n}$

  Và có thể phân tách thành hai dãy từ mã khác nhau.

- ***Cách 1***:

  $v_{11} \mid v_{12} \mid .. \mid v_{1k} \mid w_{11}v_{21}..v_{2l}w_{22} \mid .. \mid w_{(i-1)(i-1)}v_{i1}..v_{im}w_{ii} \mid v_{(i+1)1} \mid .. \mid v_{(i+1)n}$

- ***Cách 2:***

  $v_{11}\ v_{12}\ ..\ v_{1k}\ w_{11} \mid v_{21} \mid .. \mid v_{2l} \mid w_{22}\ ..w_{(i-1)(i-1)} \mid v_{i1} \mid ..v_{im} \mid w_{ii}v_{(i+1)1}\ ..\ v_{(i+1)n}$

CÁCH XÂY DỰNG BẢNG THỬ MÃ

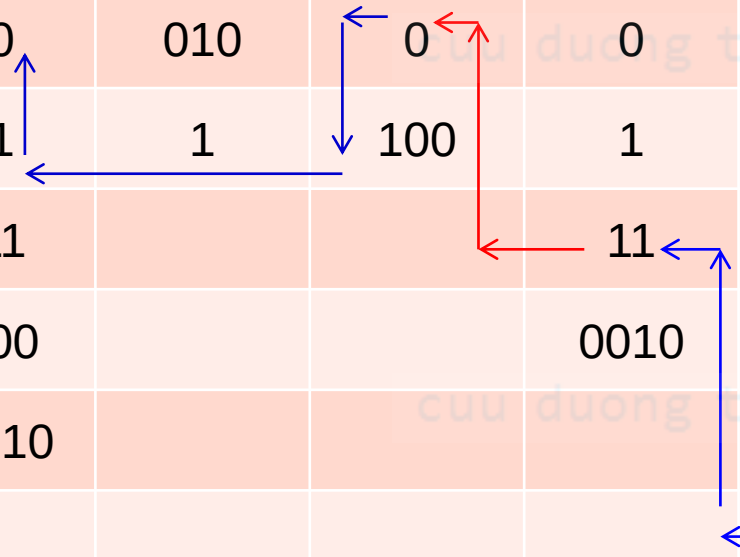
- **B1**. Sắp xếp các từ mã vào cột đầu tiên của bảng (cột 1).
- **B2**. So sánh các từ mã ngắn với các từ mã dài hơn trong cột 1, nếu từ mã ngắn giống phần đầu từ mã dài thì ghi phần còn lại trong từ mã dài sang cột 2.
- **B3**. Đối chiếu các tổ hợp mã trong cột 2 với các từ mã trong cột 1 lấy phần còn lại ghi vào cột tiếp theo (cột 3).
- **B4**. Đối chiếu các tổ hợp mã trong cột 3 với các từ mã trong cột 1… Thực hiện giống như trên cho đến khi không thể điền thêm, hoặc cột mới thêm vào trùng với một cột trước đó, hoặc có một chuỗi trong cột mới trùng với một từ mã.

BẢNG THỬ MÃ (VÍ DỤ)

- Lập bảng thử mã cho bộ mã

A = {00, 01, 011, 1100, 00010}

1	2	3	4	5
00	010	0	0	0
01	1	100	1	1
011			11	11
1100			0010	0010
00010				100
				00



Mã là không phân tách được trên chuỗi 000101100 vì có hai cách phân tách khác nhau

00 | 01 | 011 | 00

00010 | 1100

BẢNG THỬ MÃ (VÍ DỤ)

Lập bảng thử mã cho bộ mã

A = {010, 0001, 0110, 1100, 00011, 00110, 11110, 101011 }

Gợi ý:

Cột 2 ={1}

Cột 3 ={100, 1110, 01011}

Cột 4 ={11}

Cột 5 ={00, 110}

Cột 6 ={01, 0, 011, 110}

Cột 7 ={0, 10, 001, 110, 0011, 0110}

Cột 7 chứa từ mã 0110 nên bảng mã này ***không phải là bảng mã tách được***.

BẢNG THỬ MÃ (VÍ DỤ)

Lập bảng thử mã cho bộ mã

A = {00, 01, 100, 1010, 1011}

B = {10, 100, 01, 011}

Điều kiện cần và đủ để một bộ mã phân tách được là không có phần tử nào trong các cột khác cột 1 trùng với một phần tử trong cột 1.

Định lý Shannon

Cho nguồn tin $X = \{a_1,\ldots,a_k\}$ với các xác suất tương ứng $p_1,\ldots,p_k$. Một bộ mã phân tách được bất kỳ cho nguồn này với cơ số mã m, chiều dài trung bình từ mã sẽ thỏa:

$$\bar{l} \geq \frac{H(X)}{\log m}$$

(trong đó H(X) là entropy của nguồn)

ĐỊNH LÝ MÃ HÓA NGUỒN

$$\frac{H(X)}{\log m} \leq \bar{l} \leq \frac{H(X)}{\log m} + 1$$

Đối với mã nhị phân:

$$H(X) \leq \bar{l} \leq H(X) + 1$$

Bảng mã được gọi là tối ưu tuyệt đối khi:

$$\bar{l} = H(X)$$

Ví dụ

Xét biến ngẫu nhiên $X=\{x_1, x_2, x_3, x_4\}$

Có phân phối: P={1/2, 1/4, 1/8, 1/8}

Có bảng mã $W=\{w_1= 0, w_2=10, w_3=110, w_4=111\}$

Ta tính được độ dài trung bình từ mã:

$\bar{l}$ = 1/2 * 1 + 1/4 *2 + 1/8 * 3 + 1/8 * 3 = 1.75

Tính Entropy của X:

H(X)= H(0.5, 0.25, 0.125, 0.125) = 0.5 +0.5 + 0.375 + 0.375 =1.75

$\bar{l}$ = H(X) nên bảng mã W được gọi là mã thống kê tối ưu.

HIỆU SUẤT LẬP MÃ

- Hiệu suất lập mã *h* được định nghĩa bằng tỉ số của entropy của nguồn với chiều dài trung bình của bộ mã được lập.

$$h = \frac{H(X)}{\bar{l}}$$

- *h* càng tiến tới 1, tính kinh tế của mã càng cao.

Mã hóa Shanno

- Cho nguồn tin $X = \{a_1,\ldots,a_k\}$ với các xác suất tương ứng $p_1,\ldots,p_k$.
- Thuật toán mã hóa theo Shanno như sau:
  - Sắp xếp các xác suất theo thứ tự giảm dần.
  - Định nghĩa $q_1=0$, $q_i = \sum_{j=1}^{i-1} p_j, \forall$ i=1,2,3,..,k. Ở bước này theo giả thiết $p_1 \geq \ldots \geq p_k$.
  - Đổi q_i sang cơ số 2, sẽ được một chuỗi nhị phân.
  - Từ mã được gán cho a_i là l_i ký hiệu lấy từ vị trí sau dấu phẩy của chuỗi nhị phân tương ứng với q_i.
  - Trong đó $l_i = -\log_2 p_i$

Mã hóa Shanno

- Hãy mã hóa nguồn S = {a_1,a_2,a_3,a_4,a_5,a_6} với các xác suất tương ứng 0.3 ; 0.25 ; 0.2 ; 0.12 ; 0.08 ; 0.05.

Tin a_i	Xác suất p_i	$q_i = \sum_{j=1}^{i-1} p_j$	Biểu diễn nhị phân	$l_i = -\log_2 p_i$	Từ mã w_i
a_1	0.3	0	0,00	2	00
a_2	0.25	0.3	0,01001…	2	01
a_3	0.2	0.55	0,10001…	3	100
a_4	0.12	0.75	0,11000…	4	1100
a_5	0.08	0.87	0,11011…	4	1101
a_6	0.05	0.95	0,111100…	5	11110

- H (S) = 2.36; $\bar{l}$ = 2.75; h = 2.36 / 2.75 = 85.82%

Mã hóa Shanno

- Việc sắp xếp các xác suất theo thứ tự giảm dần nhằm mục đích dẫn tới độ dài trung bình của bộ mã là nhỏ.
- Phương pháp Shanno cho kết quả là một mã prefix (một bộ mã không có từ mã nào là phần đầu của từ mã khác).
- Phương pháp Shanno có thể mở rộng cho trường hợp m>2.

Mã hóa Shanno

- Hãy mã hóa nguồn S = {a_1,a_2,a_3,a_4,a_5,a_6,a_7} với các xác suất tương ứng 0.34 ; 0.23 ; 0.19 ; 0.10 ; 0.07 ; 0.06 ; 0.01.

Tin a_i	Xác suất p_i	$q_i = \sum_{j=1}^{i-1} p_j$	Biểu diễn nhị phân	$l_i = -\log_2 p_i$	Từ mã w_i
a_1	0.34	0	0,00	2	00
a_2	0.23	0.34	0,010…	3	010
a_3	0.19	0.57	0,100…	3	100
a_4	0.10	0.76	0,1100…	4	1100
a_5	0.07	0.86	0,1101…	4	1101
a_6	0.06	0.93	0,11101…	5	11101
a_7	0.01	0.99	0,1111110…	7	1111110

- H (S) = 2.37; $\bar{l}$ = 2.99; h = 2.37 / 2.99 = 81%

Mã hóa Fano

- Cho nguồn tin $\mathbf{X = \{a_1,\ldots,a_k\}}$ với các xác suất tương ứng $\mathbf{p_1,\ldots,p_k}$.
- Thuật toán mã hóa theo Fano như sau:
  - Sắp xếp các xác suất theo thứ tự giảm dần.
  - Chia các tin làm hai nhóm có xác suất xấp xỉ bằng nhau.
  - Nhóm đầu lấy trị 0, nhóm sau lấy trị 1.
  - Lặp lại bước 2 cho đến khi tất cả các nhóm chỉ còn lại một tin thì kết thúc.
  - Từ mã ứng với mỗi tin là chuỗi bao gồm các kí hiệu theo thứ tự lần lượt được gán cho các nhóm có chứa xác suất tương ứng của tin.

Mã hóa Fano

- Hãy mã hóa nguồn S = {a_1,a_2,a_3,a_4,a_5,a_6} với các xác suất tương ứng 0.3 ; 0.25 ; 0.2 ; 0.12 ; 0.08 ; 0.05.

Tin	Xác suất	Phân nhóm lần				Từ mã
		1	2	3	4	
a_1	0.3	0	0			00
a_2	0.25	0	1			01
a_3	0.2	1	0			10
a_4	0.12	1	1	0		110
a_5	0.08	1	1	1	0	1110
a_6	0.05	1	1	1	1	1111

- H (S) = 2.36; $\bar{l}$ = 2.38; h = 2.36 / 2.38 = 99.17%

Mã hóa Fano

- Hãy mã hóa nguồn $S = \{a_1,a_2,a_3,a_4,a_5,a_6,a_7,a_8\}$ với các xác suất tương ứng 0.23 ; 0.2 ; 0.14 ; 0.12 ; 0.1 ; 0.09 ; 0.06 ; 0.06.

a_i	p_i	1	2	3	4	wi
a_1	0.23	0	0			00
a_2	0.2	0	1			01
a_3	0.14	1	0	0		100
a_4	0.12	1	0	1		101
a_5	0.1	1	1	0	0	1100
a_6	0.09	1	1	0	1	1101
a_7	0.06	1	1	1	0	1110
a_8	0.06	1	1	1	1	1111

a_i	p_i	1	2	3	4	wi
a_1	0.23	0	0			00
a_2	0.2	0	1	0		010
a_3	0.14	0	1	1		011
a_4	0.12	1	0	0		100
a_5	0.1	1	0	1		101
a_6	0.09	1	1	0		110
a_7	0.06	1	1	1	0	1110
a_8	0.06	1	1	1	1	1111

- $\bar{l}_1 = 2.88$

$\bar{l}_2 = 2.89$

Nhận xét

- **Chú ý**: Trong nhiều trường hợp có nhiều hơn một cách chia thành các nhóm có tổng xác suất gần bằng nhau, ứng với mỗi cách chia có thể sẽ cho ra các bộ mã có chiều dài trung bình khác nhau.
- **Nhận xét**: Phương pháp Fano thường cho kết quả tốt hơn phương pháp Shanno.

Phương pháp mã hóa tối ưu Huffman

- ***Bổ đề***: Cho nguồn tin **S = {a_1,…,a_k}** với các xác suất tương ứng **p_1,…,p_k** . Gọi **l_1,…,l_k** là chiều dài các từ mã tương ứng với bộ mã tối ưu cho **S**. Nếu **$p_i > p_j$** thì **$l_i \leq l_j$**.
- ***Định lý 1***: Trong một bộ mã tối ưu (m=2) cho một nguồn tin, thì hai từ mã tương ứng với hai tin có xác suất nhỏ nhất phải có chiều dài bằng nhau và có thể làm cho chúng chỉ khác nhau duy nhất ở bit cuối.
- ***Định lý 2***: Xét nguồn mới **S' = {a'$_1$,…,a'$_{k-1}$}** với các xác suất tương ứng **p'$_1$,…,p'$_{k-1}$** . Trong đó **p'$_i$ = p_i** với **$1 \leq i \leq k-2$** còn **p'$_{k-1}$ = p_{k-1} + p_k**. Nếu **{w'$_1$,…,w'$_{k-1}$}** làm một mã tối ưu cho **S'** thì mã nhận được theo qui tắc sau là mã tối ưu cho **S**.

$$w_i = w'_i \qquad 1 \leq i \leq k-2$$
$$w_{k-1} = w'_{(k-1)} + \text{"0"}$$
$$wk = w'_{(k-1)}{}^{1} + \text{"1"}$$

Phương pháp mã hóa tối ưu Huffman

- **Bộ mã thỏa mãn:**
  - Đầu vào: tập nguồn đã được thống kê.
  - Đầu ra: Các từ mã tương ứng có chiều dài tối ưu, có độ lệch chuẩn (σ) cực tiểu.

$$\sigma^2 = \sum_{i=1}^{N} p_i (l - \bar{l})^2$$

Giải thuật mã hóa tối ưu Huffman

- B1: Sắp xếp các xác suất theo thứ tự giảm dần $p_1 \geq \ldots \geq p_k$.
- B2: Gán 0 tới bit cuối của w_{k-1} và 1 tới bit cuối của w_k hoặc ngược lại. Quy ước theo chiều thứ nhất.
- B3: Kết hợp p_k và p_{k-1} để tạo thành một tập xác suất mới $p_1, \ldots, p_{k-2}, p_{k-1} + p_k$.
- B4: Lặp lại các bước trên cho tập tin mới này.

Ví dụ

- Hãy mã hóa nguồn S = {a_1,a_2,a_3,a_4,a_5,a_6} với các xác suất tương ứng 0.3 ; 0.25 ; 0.2 ; 0.12 ; 0.08 ; 0.05.

a_i	p_i	Lần 1	Lần 2	Lần 3	Lần 4	wi
a_1	0.3	0.3	0.3	0.45	0.55	00
a_2	0.25	0.25	0.25	0.3	0.45	01
a_3	0.2	0.2	0.25	0.25		11
a_4	0.12	0.13	0.2			101
a_5	0.08	0.12				1000
a_6	0.05					1001

- H (S) = 2.36; $\bar{l}$ = 2.38; h = 2.36 / 2.38 = 99.17%

Nhận xét

- Trong trường hợp nếu xác suất $\mathbf{p_{k-1} + p_k}$ bằng với một xác suất $\mathbf{p_i}$ nào đó thì chúng ta có thể đặt $\mathbf{p_{k-1} + p_k}$ nằm dưới hoặc nằm trên xác suất $\mathbf{p_i}$ thì kết quả chiều dài trung bình vẫn không thay đổi cho dù các từ mã kết quả có thể khác nhau.
- So sánh với phương pháp Fano ta thấy trong trường hợp trên thì cả hai phương pháp cho hiệu suất như nhau.
- Tuy nhiên, trong trường hợp tổng quát, phương pháp **Fano** không phải là phương pháp mã hóa tối ưu.