ENTROPY VÀ THÔNG TIN

PGS.PTS.NGƯT. Đoàn Phan Tân

Khái niệm Entropy và Thông tin là khái niệm cơ bản của Lý thuyết thông tin.

Lý thuyết thông tin (Infomation theory) là lý thuyết liên quan đến các định luật toán học chi phối việc truyền, tiếp nhận và xử lý thông tin. Chính xác hơn lý thuyết thông tin đề cập tới các vấn đề về đo số lượng thông tin, biểu diễn thông tin (như vấn đề mã hoá), và khả năng của các hệ thống truyền thông có nhiệm vụ truyền, nhận và xử lý thông tin. Việc mã hoá có thể dùng để chuyển các tín hiệu âm thanh và hình ảnh thành tín hiệu điện, điện từ hoặc dùng để bảo mật thông tin.

Lý thuyết thông tin do Claude E. Shanon, một kỹ sư người Mỹ, một chuyên viên về kỹ thuật truyền tin đưa ra vào năm 1948 với bài báo "A mathematical theory of communication", nhằm giải quyết nhu cầu về cơ sở lý thuyết của công nghệ truyền thông. Nhu cầu này nảy sinh do độ phức tạp của quá trình truyền tin trên các kênh truyền thông như các mạng lưới điện thoại, điện báo và truyền thanh. Thuật ngữ thông tin ở đây là để chỉ các thông báo được truyền đi như: tiếng nói và âm nhạc được truyền đi bằng điện thoại hoặc truyền thanh, hình ảnh được truyền đi bằng truyền hình, các dữ liệu số hoá trên các mạng máy tính. Lý thuyết thông tin còn được ứng dụng trong những lĩnh vực khác nhau như điều khiển học, ngôn ngữ học, tâm lý học...

1- Entropy là số đo độ không xác định

Sự không xác định là tính chất chủ yếu của các biến cố ngẫu nhiên. Nhưng rõ ràng là mức độ không xác định của các biến cố ngẫu nghiên khác nhau là khác nhau.

Ví dụ:

- Rất khó đoán trước được người đầu tiên mà ta gập ở ngoài phố là đàn ông hay đàn bà. Nhưng còn khó hơn khi đoán trước người chiến thắng trong một cuộc đua có 10 người tham gia.

- Trong khi đó, gần như tuyệt đối ta có thể khẳng định "màu của con quạ mà ta gập đầu tiên" là màu đen.

Vấn đề đặt ra là, cần phải xây dựng một đại lượng cho phép ta *đánh giá bằng số* độ không xác định của các phép thử , để ta có thể so sánh được chúng vơí nhau (về đọ không xác định).

Trước hết, ta xét các phép thử α có k kết cục đồng khả năng. Rõ ràng đặc trưng bằng số phải tìm của độ không xác định của α phụ thuộc

vào k, tức là một hàm số của k. Rõ ràng hàm f(k) này phải có hai tính chất sau:

- f(1) = 0, vì với k = 1 thì tính không xác định của phép thử α hoàn toàn không có.

- f(k) > f(l) nếu k > l, vì độ không xác định của phép thử α sẽ tăng khi k tăng.

Bây giờ ta xét hai phép thử độc lập α và β. Giả sử α có k kết cục đồng khả năng, β có l kết cục đồng khả năng. Khi đó phép thử hợp $\alpha\beta$, là phép thử thực hiện đồng thời cả hai phép thử α và β, sẽ có kl kết cục đồng khả năng. Rõ ràng độ không xác định của phép thử hợp sẽ lớn hơn độ không xác định của phép các thử thành phần. Một cách tự nhiên ta thừa nhận rằng: *độ không xác định của phép thử $\alpha\beta$ bằng tổng độ không xác định của các phép thử α và β.* Do đó hàm f(k) phải thoả mãn tính chất sau:

$$f(kl) = f(k) + f(l)$$

Ta nhận thấy rằng trong toán học hàm logarit với cơ số lớn hơn 1 là hàm có các tính chất trên. Điều đó gợi ý cho ta lấy số $\log_a k$ làm số đo độ không xác định của phép thử có k kết cục đồng khả năng, trong đó a > 1 để bảo đảm tính đồng biến của hàm số này. Vì vậy:

$$H(\alpha) = \log_a k, \text{ với } a > 1$$

Trong kỹ thuật người ta thường chọn cơ số a = 2, tức là đặt:

$$H(\alpha) = \log_2 k$$

Nếu phép thử α có 2 kết cục đồng khả năng thì k = 2 (ví dụ: phép thử là việc gieo một đồng tiền, các kết cục của nó là việc xuất hiện một trong hai mặt sấp hoặc ngửa), thì $f(2) = \log_2 2 = 1$. Do đó người ta lấy số đo độ không xác định của phép thử α có 2 kết cục đồng khả năng làm đơn vị đo độ không xác định. Đơn vị đó thường gọi là *đơn vị nhị phân*, còn được gọi tắt là *một bít*. (viết tắt của từ *binary digit)*.

Ta xét bảng phân phối xắc suất của phép thử có k kết quả đồng khả năng

Kết cục của phép thử	A_1	A_2	A_3	.....	A_n
Xắc suất	1/k	1/k	1/k	.....	1/k

Vì độ không xác định chung của phép thử là $\log_2 k$, nên có thể thừa nhận rằng : mỗi kết cục riêng biệt, có xắc suất 1/k, có một độ không xác

$$\frac{1}{k}\log_2 k = -\frac{1}{k}\log\frac{1}{k}$$

định bằng:

Do đó, một cách tự nhiên ta thừa nhận rằng trong kết quả của phép thử, cho bởi bảng phân phối xắc suất sau đây:

Kết cục của phép thử	A_1	A_2	A_3
Xắc suất	1/2	1/3	1/6

các kết cục A_1, A_2, A_3 có độ không xác đọnh tương ứng bằng:

$$-\frac{1}{2}\log\frac{1}{2}, \qquad -\frac{1}{3}\log\frac{1}{3}, \qquad -\frac{1}{6}\log\frac{1}{6}$$

Như vậy độ không xác định chung của phép thử này là:

$$-\frac{1}{2}\log\frac{1}{2} \ -\frac{1}{3}\log\frac{1}{3} \ -\frac{1}{6}\log\frac{1}{6}$$

Trong trường hợp tổng quát, với phép thử α có bản phân phối xắc suất:

Kết cục của phép thử	A_1	A_2	A_3	.....	A_n
Xắc suất	$p(A_1)$	$p(A_2)$	$p(A_3)$	.....	$p(A_n)$

thì số đo độ không xác định của nó, ký hiệu là $H(\alpha)$, bằng:

$H(\alpha) = - p(A_1)\log_2 p(A_1) - p(A_2)\log_2 p(A_2) - \ldots\ldots - p(A_n)\log_2 p(A_n)$

con số trên được gọi là entropy của phép thử α.

Tính chất của entropy:

1) $H(\alpha) \geq 0$

Vì $0 \leq p(A_i) \leq 1$, nên $- p(A_i)\log_2 p(A_i) \geq 0$, với mọi i.

2) $H(\alpha) = 0$ khi một trong các xắc suất $p(A_i)$ bằng 1, còn các xắc suất khác bằng 0.

Điều này hoàn toàn phù hợp với ý nghĩa của $H(\alpha)$ là đại lượng đo độ không xác định, vì chỉ khi đó phép thử α mới không chứa độ không xác định nào (Ta nhớ rằng: $p(A_1) + p(A_2) + ...+ p(A_n) = 1$).

Chú ý rằng :

$$\log_2 k = \log_2 10.\log_{10} k = 3{,}32.\lg k$$

nên ta có thể tính loga cơ số 2 thông qua loga cơ số 10.

Ví dụ: Giả sử qua nhiều năm quan sát thời tiết tại một thời điểm người ta thu được kết quả sau:

Thời tiết trong ngày 15 tháng 6 (phép thử α_1)

Các kết cục của phép thử	có mưa	không mưa
Xắc suất	0,4	0,6

Thời tiết trong ngày 15 tháng 11 (phép thử α_2)

Các kết cục của phép thử	có mưa	có tuyết	không mưa
Xắc suất	0,65	0,15	0,2

Entropy tương ứng của hai phép thử này là:

$H(\alpha_1) = - 0,4 \log_2 0,4 - 0,6 \log_2 0,6 = 0,97$

$H(\alpha_2) = - 0,66 \log_2 o,65 - 0,15 \log_2 0,15 - 0,2 \log_2 0,2 = 1,28$

Vậy $H(\alpha_2) > H(\alpha_1)$, do đó tại khu vực đang xét thời tiết ngày 15 tháng 11 khó dự báo hơn thời tiết ngày 15 tháng 6.

2- Entropy và thông tin

Một khái niệm cơ bản của lý thuyết thông tin là số lượng của thông tin trong thông báo, gọi là nội dung thông tin, nó có thể xác định và đo được bằng đại lượng toán học. Thuật ngữ "nội dung" ở đây không liên quan gì đến nội dung của thông báo được truyền đi, mà là xác suất nhận được thông báo đã cho từ một tập hợp các thông báo có thể. Giá trị cao nhất đối với nội dung thông tin được gán cho thông báo có ít khả năng nhất, tức là có độ không xác định lớn nhất. Bởi vì độ không xác định của mọt phép thử càng lớn thì sự xác định kết quả của nó sẽ cho một thông tin càng lớn. Nếu thông báo được mong đợi với 100 - phần trăm chắc chắn thì nội dung của nó bằng 0, và khi đó độ không xác định của nó cũng bằng 0.

Ta biết rằng tăng lượng tin tức về một hiện tượng nào đó cũng là giảm độ chưa biết hoặc độ không xác định của nó. Vì vậy *Entropy $H(\alpha)$ của phép thử α có thể xem là thông tin về α chứa trong bản thân phép thử này*. Đó là thông tin lớn nhất về α mà nó có thể có. Khi α được thực hiện thì $H(\alpha) = 0$. Cho nên có thể nói *Entropy $H(\alpha)$ của phép thử α bằng thông tin nhận được sau khi thực hiện phép thử α, tức là thông tin trung bình chứa trong một kết cục của phép thử.*

Để liên kết nội dung thông tin, ký hiệu là I, với xác suất, Shanon đưa ra công thức đơn giản sau đây:

$$I = \log_2 1/p$$

trong đó p là xác suất của thông báo được truyền đi.

Nếu chú ý rằng p = 1/k, trong đó k là số các kết cục đồng khả năng của phép thử, thì ta thấy công thức trên đồng nhất với công thức:

$$H = \log_2 k$$

Ví dụ: Khi gieo một đồng tiền, thì thông báo "xấp hoặc ngửa" để mô tả kết quả, sẽ không có nội dung thông tin vì đó là một kết cục chắc chắn. Mặt khác mỗi thông báo tách ra "xấp" hoặc " ngửa" sẽ có xác xuất bằng nhau và là p = 1/2 vì có phép thử gieo đồng tiền có k = 2 kết cục đồng khả năng. Ap dụng công thức trên ta thấy nội dung của thông báo "sấp" hoặc "ngửa" có giá trị là:

$$I = \log_2 1/p = \log_2 2 = 1.$$

và Entropy của phép thử là:

$$H = \log_2 k = \log_2 2 = 1.$$

Nội dung của thông tin có thể hiểu đó là số các ký hiệu có thể dùng để biểu diễn thông báo. Trong ví dụ trên, nếu ký hiệu "xấp" là số 1, " ngửa" là số 0, thì chỉ có một cách chọn để biểu diễn thông báo là 1 hoặc 0. Số 0 và 1 là những chữ số của hệ đếm nhị phân, và việc chọn giữa hai ký hiệu đó tương ứng với một đơn vị thông tin nhị phân, hay còn gọi là bit.

Bây giờ giả sử ta gieo ba lần liên tiếp một đồng tiền, thì 8 kết quả đồng khả năng (hay thông báo) có thể biểu diễn như sau: 000, 001, 010,100, 011, 101, 110, 111. Xác suất của mỗi thông báo này là p =1/8, và nội dung thông tin của nó là $\log_2 1/p = \log_2 8 = 3$, đó chính là số bít cần thiết để biểu diễn mỗi thông báo nói trên.

Như vậy, Shanon đã chúng minh được rằng thông tin có thể đo được, tức là với bản tin bất kỳ, ta có thể xác định được nó chứa bao nhiêu đơn vị tin tức.

Thông tin có thể đo được. Đó là phát minh cũng có ý nghĩa về sự hiểu biết của con người đối với thế giới khách quan như ý nghĩa về khả năng đo được của năng lượng. Người ta đã chế tạo ra các máy để sản sinh và chế biến được năng lượng, và giờ đây người ta cũng chế tạo ra các máy để gia công tin tức, đó là máy tính điện tử.

Vì Entropy là đại lượng dùng để chỉ nội dung thông tin trung bình của một thông báo nên nó được ứng dụng để mã hoá các tín hiệu truyền đi.

Ví dụ: Nếu thông báo được truyền đi bao gồm các tổ hợp ngẫu nhiên của 26 chữ cái, một khoảng trống và 5 dấu chấm câu, tổng cộng là

32 ký hiệu, và giả sử rằng xác suất của mỗi ký hiệu là như nhau, thì entropy $H = \log_2 32 = 5$. Điều đó có nghĩa là cần 5 bít để mã hoá mỗi ký hiệu: 00000, 00001, 00010, ..., 11111.

Hiệu quả của việc truyền và lưu trữ thông tin đòi hỏi phải rút gọn số các bít dùng để mã hoá. Những phương pháp của lý thuyết thông tin được sử dụng để mã hoá các thông tin ngữ nghĩa, đưa thông tin vào máy tính để bảo quản và thực hiện các quá trình tìm tin, truyền tin.

* * *