Thuật toán Cây quyết định (Decision Tree)

**Thuật toán ID3**

Giờ chúng ta hãy cùng tìm hiểu cách thức hoạt động của thuật toán cây quyết định thông qua thuật toán đơn giản ID3.

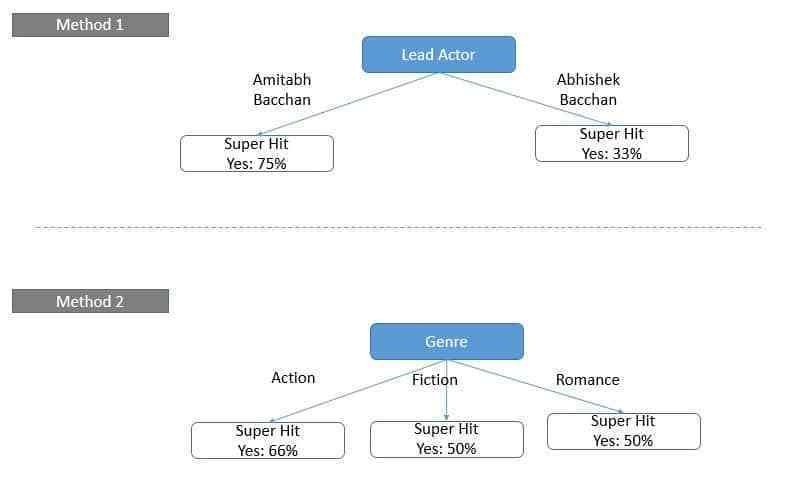
ID3 (J. R. Quinlan 1993) sử dụng phương pháp tham lam tìm kiếm từ trên xuống thông qua không gian của các nhánh có thể không có backtracking. ID3 sử dụng [Entropy](https://trituenhantao.io/tu-dien-thuat-ngu/entropy/) và Information Gain để xây dựng một cây quyết định.

**Ta xét ví dụ 2:**

Bạn muốn xem xét sự thành công của một bộ phim thông qua hai yếu tố: diễn viên chính của phim và thể loại phim:

| **Lead Actor** | **Genre** | **Hit(Y/N)** |
| --- | --- | --- |
| Amitabh Bacchan | Action | Yes |
| Amitabh Bacchan | Fiction | Yes |
| Amitabh Bacchan | Romance | No |
| Amitabh Bacchan | Action | Yes |
| *Abhishek Bacchan* | *Action* | *No* |
| *Abhishek Bacchan* | *Fiction* | *No* |
| *Abhishek Bacchan* | *Romance* | *Yes* |

Giả sử, bạn muốn xác định độ thành công của bộ phim chỉ trên 1 yếu tố, bạn sẽ có hai cách thực hiện sau: qua diễn viên chính của phim và qua thể loại phim.



Qua sơ đồ, ta có thể thấy rõ ràng ràng, với phương pháp thứ nhất, ta phân loại được rõ ràng, trong khi phương pháp thứ hai, ta có một kết quả lộn xộn hơn. Và tương tự, cây quyết định sẽ thực hiện như trên khi thực hiện việc chọn các biến.

Có rất nhiều hệ số khác nhau mà phương pháp cây quyết định sử dụng để phân chia. Dưới đây, tôi sẽ đưa ra hai hệ số phổ biến là ***Information Gain*** và **Gain Ratio** (ngoài ra còn hệ số Gini).

***Entropy trong Cây quyết định (Decision Tree)***

[Entropy](https://trituenhantao.io/tu-dien-thuat-ngu/entropy/) là thuật ngữ thuộc Nhiệt động lực học, là thước đo của sự biến đổi, hỗn loạn hoặc ngẫu nhiên. Năm 1948, Shannon đã mở rộng khái niệm [Entropy](https://trituenhantao.io/tu-dien-thuat-ngu/entropy/) sang lĩnh vực nghiên cứu, thống kê với công thức như sau:

Với một phân phối xác suất của một biến rời rạc x có thể nhận n giá trị khác nhau x1,x2,…,xn.

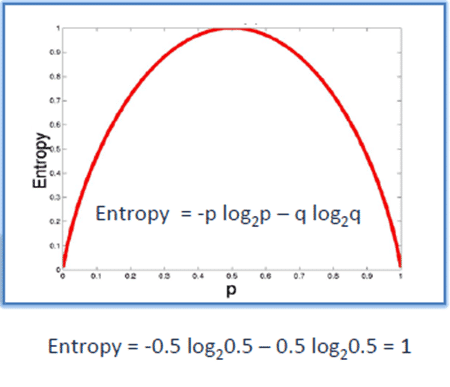
Giả sử rằng xác suất để x nhận các giá trị này là pi=p(x=xi).

Ký hiệu phân phối này là p=(p1 ,p2 ,…,pn). [Entropy](https://trituenhantao.io/tu-dien-thuat-ngu/entropy/) của phân phối này được định nghĩa là:

H(p)=  – ∑nn=1 pi log(pi)

Giả sử bạn tung một đồng xu, [entropy](https://trituenhantao.io/tu-dien-thuat-ngu/entropy/) sẽ được tính như sau:

H = -[0.5 ln(0.5) + 0.5 ln(0.5)]

Hàm [Entropy](https://trituenhantao.io/tu-dien-thuat-ngu/entropy/)

Hình vẽ trên biểu diễn sự thay đổi của hàm [entropy](https://trituenhantao.io/tu-dien-thuat-ngu/entropy/). Ta có thể thấy rằng, [entropy](https://trituenhantao.io/tu-dien-thuat-ngu/entropy/) đạt tối đa khi xác suất xảy ra của hai lớp bằng nhau.

P tinh khiết: pi = 0 hoặc pi = 1

P vẩn đục: pi = 0.5, khi đó hàm [Entropy](https://trituenhantao.io/tu-dien-thuat-ngu/entropy/) đạt đỉnh cao nhất

***Information Gain* *trong Cây quyết định (Decision Tree)***

Information Gain dựa trên sự giảm của hàm [Entropy](https://trituenhantao.io/tu-dien-thuat-ngu/entropy/) khi tập dữ liệu được phân chia trên một thuộc tính. Để xây dựng một cây quyết định, ta phải tìm tất cả thuộc tính trả về Infomation gain cao nhất.

Để xác định các nút trong mô hình cây quyết định, ta thực hiện tính Infomation Gain tại mỗi nút theo trình tự sau:

•**Bước 1**: Tính toán hệ số [Entropy](https://trituenhantao.io/tu-dien-thuat-ngu/entropy/) của biến mục tiêu S có N phần tử với Nc phần tử thuộc lớp c cho trước:

H(S)=  – ∑cc=1 (Nc/N) log(Nc/N)

•**Bước 2**: Tính hàm số [Entropy](https://trituenhantao.io/tu-dien-thuat-ngu/entropy/) tại mỗi thuộc tính: với thuộc tính x, các điểm dữ liệu trong S được chia ra K child node S1, S2, …, SK với số điểm trong mỗi child node lần lượt là m1, m2 ,…, mK , ta có:

H(x, S) = ∑Kk=1 (mk / N) \* H(Sk )

**Bước 3**: Chỉ số Gain Information được tính bằng:

**G(x, S) = H(S) – H(x,S)**

Với ví dụ 2 trên, ta tính được hệ số [Entropy](https://trituenhantao.io/tu-dien-thuat-ngu/entropy/) như sau:

[*Entropy*](https://trituenhantao.io/tu-dien-thuat-ngu/entropy/)*Parent= -(0.57\*ln(0.57) + 0.43\*ln(0.43))* = 0.68

Hệ số [Entropy](https://trituenhantao.io/tu-dien-thuat-ngu/entropy/) theo phương pháp chia thứ nhất:

[*Entropy*](https://trituenhantao.io/tu-dien-thuat-ngu/entropy/)*left= -(.75\*ln(0.75) + 0.25\*ln(0.25))  = 0.56*[*Entropy*](https://trituenhantao.io/tu-dien-thuat-ngu/entropy/)*right= -(.33\*ln(0.33) + 0.67\*ln(0.67)) = 0.63*

Ta có thể tính hệ số ***Information Gain*** như sau:

*Information Gain* = *0.68 – (4\*0.56 + 3\*0.63)/7 = 0.09*

Hệ số [Entropy](https://trituenhantao.io/tu-dien-thuat-ngu/entropy/) với phương pháp chia thứ hai như sau:

[*Entropy*](https://trituenhantao.io/tu-dien-thuat-ngu/entropy/)*left= -(.67\*ln(0.67) + 0.33\*ln(0.33))  = 0.63*[*Entropy*](https://trituenhantao.io/tu-dien-thuat-ngu/entropy/)*middle= -(.5\*ln(0.5) + 0.5\*ln(0.5))  = 0.69*[*Entropy*](https://trituenhantao.io/tu-dien-thuat-ngu/entropy/)*right= -(.5\*ln(0.5) + 0.5\*ln(0.5))  = 0.69*

Hệ số ***Information Gain***:

*Information Gain = 0.68 – (3\*0.63 + 2\*0.69 + 2\*0.69)/7 = 0.02*

So sánh kết quả, ta thấy nếu chia theo phương pháp 1 thì ta được giá trị hệ số Information Gain lớn hơn gấp 4 lần so với phương pháp 2. Như vậy, giá trị thông tin ta thu được theo phương pháp 1 cũng nhiều hơn phương pháp 2.