[**https://towardsdatascience.com/what-is-the-c4-5-algorithm-and-how-does-it-work-2b971a9e7db0**](https://towardsdatascience.com/what-is-the-c4-5-algorithm-and-how-does-it-work-2b971a9e7db0)

[**https://sefiks.com/2018/05/13/a-step-by-step-c4-5-decision-tree-example/**](https://sefiks.com/2018/05/13/a-step-by-step-c4-5-decision-tree-example/)

[**https://sefiks.com/2017/11/20/a-step-by-step-id3-decision-tree-example/**](https://sefiks.com/2017/11/20/a-step-by-step-id3-decision-tree-example/)

**C4.5** là giải thuật xây dựng 1 phân lớp (classifier) dưới dạng 1 **decision tree** **classifier(**bằng cách sinh ra các decision dựa trên tập dữ liệu mẫu**).**

**C4.5** là cải tiến của **ID3**, khắc phục các nhược điểm của **ID3** như : **ID3** ko hỗ trợ thuộc tính là 1 số (chỉ hỗ trợ nominal values), dataset của ID3 ko được chứa missing data, và giải thuật ID3 có xu hướng bị overfitting. Với những cải tiến mới, **C4.5** có thể tạo ra 1 model khái quát hơn, có thể xử lý thuộc tính kiểu số và các missing data và có thể làm việc với cả dữ liệu liên tục và rời rạc.

**Ví dụ dataset của ID3 :**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Day** | **Outlook** | **Temp.** | **Humidity** | **Wind** | **Decision** |
| 1 | Sunny | Hot | High | Weak | No |
| 2 | Sunny | Hot | High | Strong | No |
| 3 | Overcast | Hot | High | Weak | Yes |
| 4 | Rain | Mild | High | Weak | Yes |
| 5 | Rain | Cool | Normal | Weak | Yes |
| 6 | Rain | Cool | Normal | Strong | No |

**Ví dụ dataset của C4.5**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Day** | **Outlook** | **Temp.** | **Humidity** | **Wind** | **Decision** |
| 1 | Sunny | 85 | 85 | Weak | No |
| 2 | Sunny | 80 | 90 | Strong | No |
| 3 | Overcast | 83 | 78 | Weak | Yes |
| 4 | Rain | 70 | 96 | Weak | Yes |
| 5 | Rain | 68 | 80 | Weak | Yes |
| 6 | Rain | 65 | 70 | Strong | No |

Ta cần hiểu 2 khái niệm : **Information Gain** and **Entropy**

**Information Gain:**

Vì trong tập dữ liệu, mỗi record gồm nhiều thuộc tính và label. IG giúp ta biết được thuộc tính nào quan trọng hơn thuộc tính nào.

GainRatio(A) = Gain(A) / SplitInfo(A)

SplitInfo(A) = -∑ |Dj|/|D| x log2|Dj|/|D|

Gain(S, A) = Entropy(S) – ∑ [ p(S|A) . Entropy(S|A) ]

**Entropy**

Là phép đo sự ko chắc chắn. Ví dụ nếu tung 1 đồng xu hoàn hảo, có 2 mặt. Kết quả (h,h,t,t) xác xuất 50%. Với kết quả trên thu được tính ko chắc chắn cao nhất. H(X) = 1

Entropy quyết định việc split data của Decision tree.

H(X)=−∑nn=1 P(xi)log2P(xi)