THUẬT TOÁN ID3

Phân lớp(classification)

Tạo ra bộ phân lớp/ mô hình phân lớp từ dữ liệu.

Phân lớp dữ liệu là tiến trình có 2 bước

- **Huấn luyện**: Dữ liệu huấn luyện được phân tích bởi thuật tóan phân lớp (có thuộc tính nhãn lớp) để tạo ra bộ phân lớp. Ví dụ CSDL

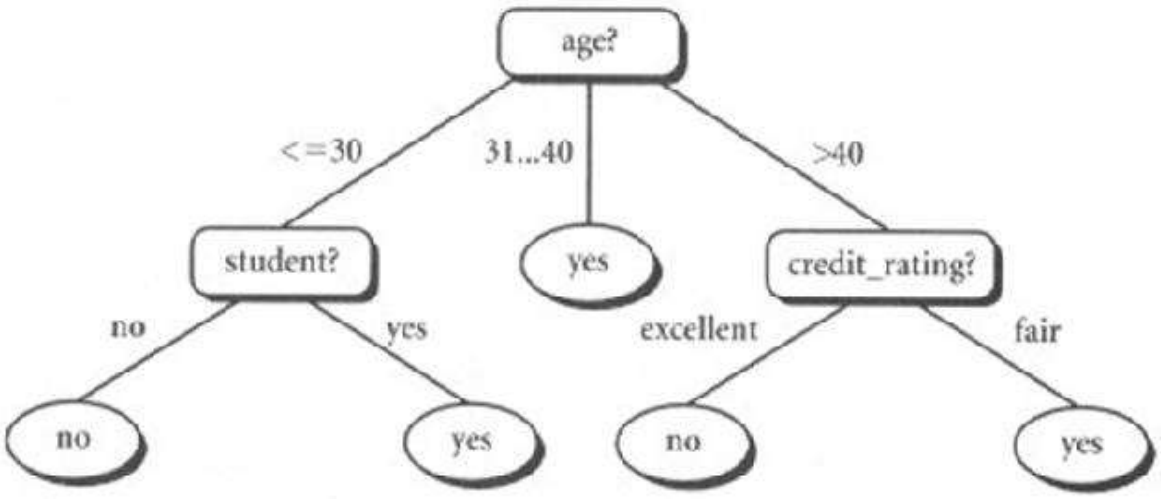
Ngày	Quang cảnh	Nhiệt độ	Độ ẩm (%)	Gió to	Kết quả
N1	Nắng	24	70	Không	Thi đấu
N2	Nắng	27	90	Có	Không thi đấu
N3	Nắng	30	85	Không	Không thi đấu
N4	Nắng	22	95	Không	Không thi đấu
N5	Nắng	20	70	Không	Thi đấu
N6	Nhiều mây	22	90	Có	Thi đấu
N7	Nhiều mây	28	75	Không	Thi đấu
N8	Nhiều mây	18	65	Có	Thi đấu
N9	Nhiều mây	28	75	Không	Thi đấu
N10	Mưa	21	80	Có	Không thi đấu
N11	Mưa	18	70	Có	Không thi đấu
N12	Mưa	24	80	Không	Thi đấu
N13	Mưa	20	80	Không	Thi đấu
N14	Mưa	21	96	Không	Thi đấu

- **Phân lớp:** Dữ liệu kiểm tra được dùng để ước lượng độ chính xác của bộ phân lớp. Nếu độ chính xác là chấp nhận được thì có thể dùng bộ phân lớp để phân lớp các mẫu dữ liệu mới.

Cây quyết định

Là cấu trúc cây sao cho:

- Mỗi nút trong ứng với một phép kiểm tra trên một thuộc tính
- Mỗi nhánh biểu diễn kết quả phép kiểm tra
- Các nút lá biểu diễn các lớp hay các phân bố lớp
- Nút cao nhất trong cây là nút gốc.



Cây quyết định: có mua computer? Dựa vào các thuộc tính tuổi, sinh viên, Uy tín.

Sườn chung về quy nạp trên cây quyết định

1. Chọn thuộc tính "tốt nhất" theo một độ đo chọn lựa cho trước
2. Mở rộng cây bằng cách thêm các nhánh mới cho từng giá trị thuộc tính
3. Sắp xếp các ví dụ học vào nút lá
4. Nếu các ví dụ được phân lớp rõ thì Stop ngược lại lặp lại các bước 1-4 cho mỗi nút lá
5.

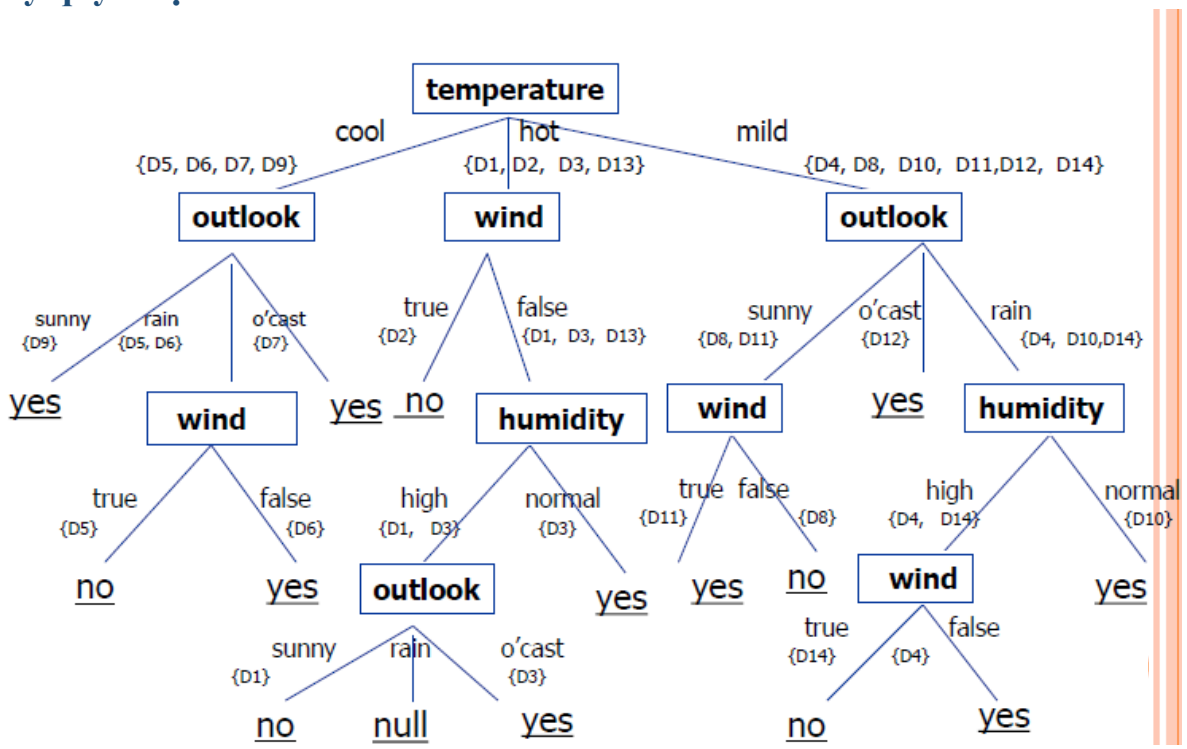
Chiến lược cơ bản

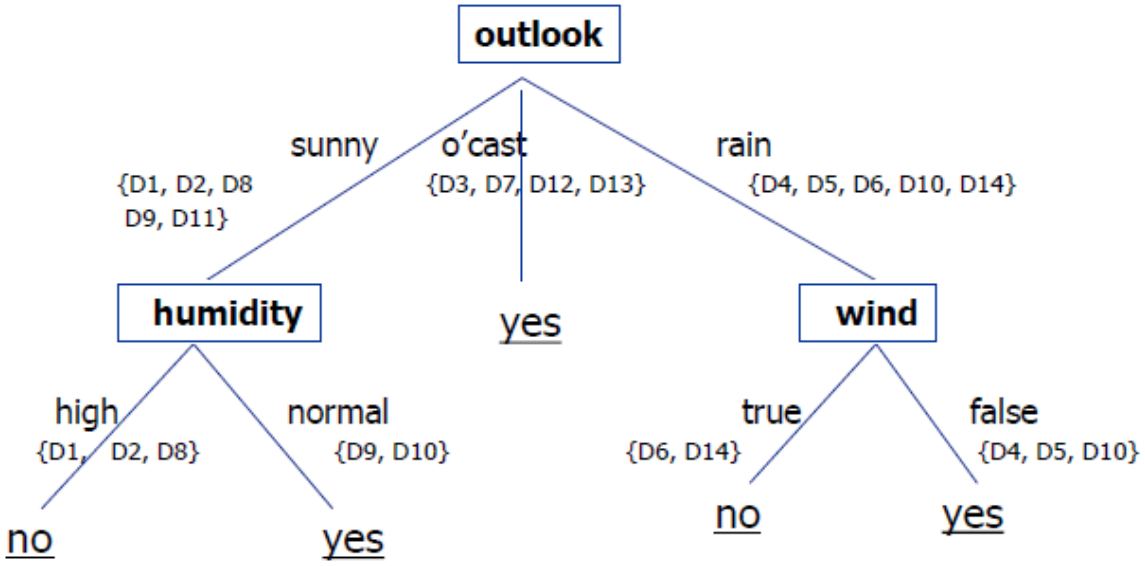
- Bắt đầu từ nút đơn biểu diễn tất cả các mẫu
- Nếu các mẫu thuộc về cùng một lớp, nút trở thành nút lá và được gán nhãn bằng lớp đó
- Ngược lại, dùng độ đo thuộc tính để chọn thuộc tính sẽ phân tách tốt nhất các mẫu vào các lớp
- Một nhánh được tạo cho từng giá trị của thuộc tính được chọn và các mẫu được phân hoạch theo
- Dùng đệ quy cùng một quá trình để tạo cây quyết định
- Tiến trình kết thúc chỉ khi bất kỳ điều kiện nào sau đây là đúng
  - Tất cả các mẫu cho một nút cho trước đều thuộc về cùng một lớp.
  - Không còn thuộc tính nào mà mẫu có thể dựa vào để phân hoạch xa hơn.
  - Không còn mẫu nào cho nhánh test_attribute = ai

Bảng dữ liệu huấn luyện

Day	Outloook	Temp	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Cây quyết định cho bài toán chơi Tennis





Cây sẽ đơn giản hơn nếu "outlook" được chọn làm gốc. Cách chọn thuộc tính tốt để tách nút quyết định?

Thuật toán ID3 (Quinlan86)

Là một trong những thuật toán xây dựng cây quyết định sử dụng information gain để lựa chọn thuộc tính phân lớp các đối tượng. Nó xây dựng cây theo cách từ trên xuống, bắt đầu từ một tập các đối tượng và một đặc tả của các thuộc tính. Tại mỗi đỉnh của cây, một thuộc tính có *information gain* lớn nhất sẽ được chọn để phân chia tập đối tượng. Quá trình này được thực hiện một cách đệ qui cho đến khi một tập đối tượng tại một cây con đã cho trở nên thuần nhất, tức là nó chỉ chứa các đối tượng thuộc về cùng một lớp. Lớp này sẽ trở thành một lá của cây.

Ngày	Quang cảnh	Nhiệt độ	Độ ẩm (%)	Gió to	Kết quả
N1	Nắng	24	70	Không	Thi đấu
N2	Nắng	27	90	Có	Không thi đấu
N3	Nắng	30	85	Không	Không thi đấu
N4	Nắng	22	95	Không	Không thi đấu
N5	Nắng	20	70	Không	Thi đấu
N6	Nhiều mây	22	90	Có	Thi đấu
N7	Nhiều mây	28	75	Không	Thi đấu
N8	Nhiều mây	18	65	Có	Thi đấu
N9	Nhiều mây	28	75	Không	Thi đấu
N10	Mưa	21	80	Có	Không thi đấu
N11	Mưa	18	70	Có	Không thi đấu
N12	Mưa	24	80	Không	Thi đấu
N13	Mưa	20	80	Không	Thi đấu
N14	Mưa	21	96	Không	Thi đấu

Bảng 1. Bảng dữ liệu

N15	**Mưa**	**25**	**80**	**Không**	**?**
N16	**Nắng**	**30**	**90**	**Có**	**?**

*dữ liệu rời rạc

Giá trị thuộc tính	Lớp	
	Thi đấu	Không thi đấu
Nắng	2	3
Nhiều mây	4	0
Mưa	3	2

Bảng 2. Thông tin phân bố lớp của thuộc tính Quang cảnh

Bảng 2 cho thấy thông tin phân lớp của thuộc tính **Quang cảnh**. **Đối với một thuộc tính liên tục, chúng ta phải chuẩn hóa chúng thành dữ liệu rời rạc trước khi xây dựng cây**. Bảng 3 chỉ ra thông tin phân lớp của thuộc tính **Độ ẩm**.

*dữ liệu liên tục

Giá trị thuộc tính	Lớp	
	Thi đấu	Không thi đấu
<=75%	5	1
>75%	4	4

Hàm Entropy

Hàm Entropy xác định tính không thuần khiết của một tập các ca dữ liệu bất kỳ. Chúng ta gọi *S* là tập các ca dương tính (ví dụ Thi đấu) và âm tính (ví dụ Không thi đấu), P_A là tỉ lệ các ca dương tính trong S, P_Q là tỉ lệ các ca âm tính trong *S*.

$$Entropy(S) = -P_A \log_2 P_A - P_Q \log_2 P_Q$$

Ví dụ 1. Trong Bảng 1 của ví dụ thi đấu tennis, tập S có 9 ca dương và 5 ca âm (ký hiệu là [9+,5-]).

$$Entropy(S) = Entropy([9+,5-]) = -\frac{9}{14}\log_2\frac{9}{14} - \frac{5}{14}\log_2\frac{5}{14} = 0.940$$

Nhận xét. Entropy bằng 0 nếu tất cả các ca trong S đều thuộc về cùng một lớp. Chẳng hạn như, nếu tất cả các ca đều dương thì và , do vậy: $P_A = 1$ và $P_Q = 0$.

$$Entropy(S) = -1\log_2(1) - 0\log_2(0) = 0$$

Entropy bằng 1 nếu tập *S* chứa số ca dương và âm bằng nhau. Nếu số các ca này khác nhau thì Entropy nằm giữa 0 và 1.

Trường hợp tổng quát, nếu *S* bao gồm *c* lớp, thì Entropy của *S* được tính bằng công thức sau:

$$\text{Entropy}(S) = \sum_{i=1}^{c} -p_i \log_2 p_i$$

trong đó *Pi* là tỉ lệ của các ca thuộc lớp *i* trong tập *S*.

Độ đo (Informatic Gain):

Độ đo, đo mức độ hiệu quả của một thuộc tính trong bài toán phân lớp dữ liệu. Đó chính là sự rút gọn mà ta mong đợi khi phân chia các ca dữ liệu theo thuộc tính này. Nó được tính theo công thức sau đây:

$$\text{Gain}(S, A) = \text{Entropy}(S) - \sum_{v \in Value(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

trong đó Value(A) là tập tất cả các giá trị có thể có đối với thuộc tính *A*, và *Sv* là tập con của *S* mà *A* có giá trị là *v*.

Ví dụ 2. Value(Gió to)={true,false}, S=[9+,5-] *Strue*, là đỉnh con với giá trị là "true", bằng [2+,3-] *Sfalse*, là đỉnh con với giá trị là "false", bằng [7+,2-]

$$\text{Gain}(S,\text{Gió to}) = \text{Entropy}(S) - \sum_{v \in Value(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

$$= Entropy(S) - \frac{5}{14} * Entropy(S_{true}) - \frac{9}{14} * Entropy(S_{false})$$

$$= 0.940 - \frac{5}{14} * 0.97 - \frac{9}{14} * 0.764$$

$$= 0.1024$$

Tương tự như vậy, ta có thể tính được độ đo cho các thuộctính còn lại của ví dụ trong Bảng 1. Đối với thuộc tính **Độ ẩm**, ta lấy độ ẩm 75% để chia các ca thành hai phần, một phần ứng với các ca có độ ẩm 75% được gọi là có độ ẩm Bình thường ([5+,1-]), phần còn lại được gọi là có độ ẩm Cao([4+,4-]). Còn đối với thuộc tính Nhiệt độ, ta sẽ chia thành ba mức, các ngày có nhiệt độ nhỏ hơn 21° được gọi là Lạnh (4 ngày), các ngày có nhiệt độ lớn hơn hay bằng 21° đến nhỏ hơn 27° được gọi là Ấm (6 ngày), và còn lại là những ngày có nhiệt độ lớn hơn hoặc bằng 27° được gọi là Nóng (4 ngày).

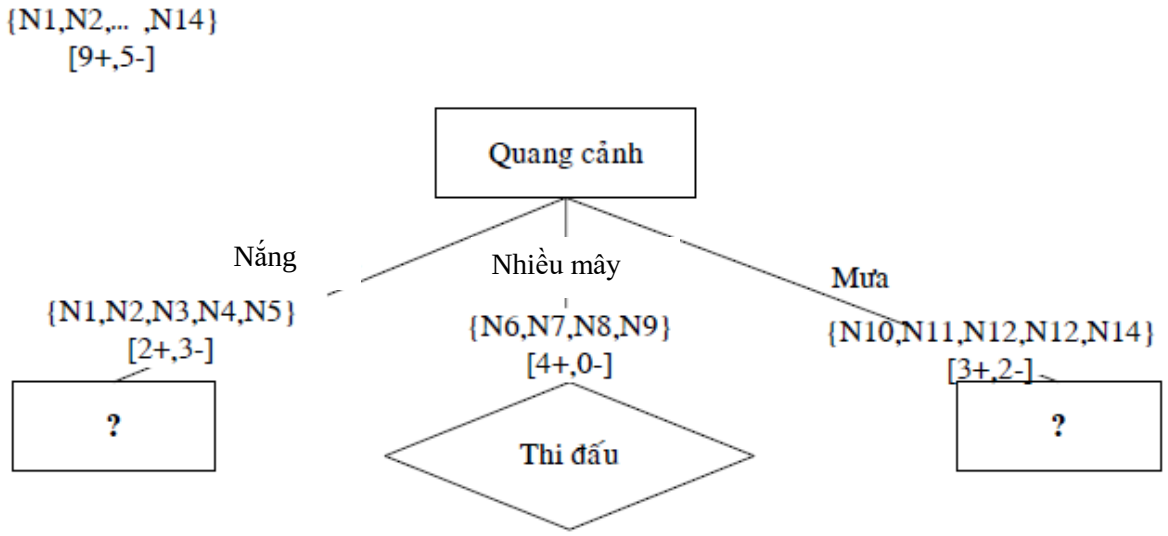
Gain(S,Quang cảnh)=0.246

Gain(S,Gió to) =0.1024.

Gain(S,Nhiệt độ)=0.029

Gain(S,Độ ẩm)=0.045

Từ đây ta thấy rằng **độ đo** của S đối với thuộc tính **Quang cảnh** là **lớn nhất** trong số 4 thuộc tính. Như vậy, có thể quyết định chọn **Quang cảnh** làm thuộc tính đầu tiên để khai triển cây. Hình 3 là khai triển của cây quyết định theo thuộc tính **Quang cảnh**.



Hình 3. Khai triển cây theo thuộc tính đã chọn

Tương tự như vậy, ta có thể tiến hành triển khai các nút ở mức tiếp theo.

$S_{nắng}=\{N1,N2,N3,N4,N5\}$

$$Entropy(S_{nắng})=-\frac{2}{5}*\log_2\frac{2}{5}-\frac{3}{5}*\log_2\frac{3}{5}=0.097$$

$$Gain(S_{nắng},\ Độ\ ẩm)=0.970-\frac{3}{5}*0.0-\frac{2}{5}*0.0=0.970$$

$$Gain(S_{nắng},\ Nhiệt\ độ)=0.970-\frac{2}{5}*0.0-\frac{2}{5}*1.0-\frac{1}{5}*0.0=0.570$$

$$Gain(S_{nắng},\ Gió\ to)=0.970-\frac{2}{5}*1.0-\frac{3}{5}*0.918=0.019$$

Từ các giá trị của Entropy Gain, ta thấy Độ ẩm là thuộc tính tốt nhất cho đỉnh nằm dưới nhánh Nắng của thuộc tính Quang cảnh. Tiếp tục quá trình trên cho tất cả các đỉnh và sẽ dừng khi không còn đỉnh nào có thể khai triển được nữa.

Bài Tập

1. Hoàn thành cây ID3 của bảng dữ liệu 1
2. Dựa vào cây ID3 vừa hoàn thành, trả lời 2 dòng N15, N16

N15	**Mưa**	**25**	**80**	**Không**	**?**
N16	**Nắng**	**30**	**90**	**Có**	**?**

Ví dụ:Nếu Quang Cảnh = “Nhiều Mây” và Nhiệt Độ =”<=21” thi Thi đấu =”Có”

3. Sử dụng ID3 xây dựng cây quyết định cho CSDL trượt tuyết bên dưới

Snow	**Weather**	**Seasion**	**Physical_Condition**	**Go_skiing**
Sticky	Foggy	Low	Rested	No
Fresh	Sunny	Low	Injured	No
Fresh	Sunny	Low	Rested	Yes
Fresh	Sunny	High	Rested	Yes
Fresh	Sunny	Mid	Rested	Yes
Frosted	Windy	High	Tired	No
Sticky	Sunny	Low	Rested	Yes
Frosted	Foggy	Mid	Rested	No
Fresh	Windy	Low	Rested	Yes
Fresh	Windy	Low	Rested	Yes
Fresh	Foggy	Low	Rested	Yes
Fresh	Foggy	Low	Rested	Yes
Sticky	Sunny	Mid	Rested	Yes
Frosted	Foggy	Low	Injured	No