



Phương pháp học cây quyết định Decision Tree

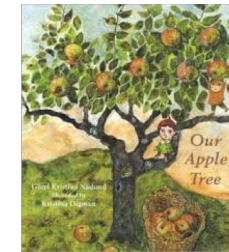


Đỗ Thành Nghị - dtnghi@cit.ctu.edu.vn
Trần Nguyễn Minh Thư - tnmthu@cit.ctu.edu.vn

Cần Thơ - 2015

Nội dung

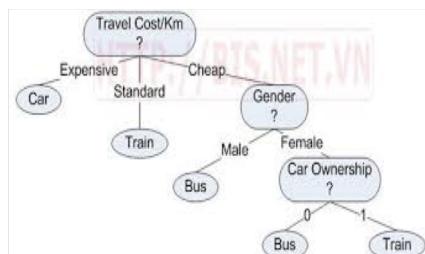
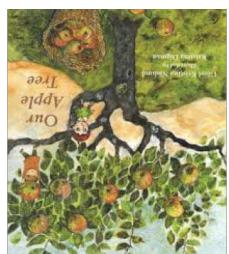
- Giới thiệu về cây quyết định
- Giải thuật học của cây quyết định
- Kết luận và hướng phát triển



6

Nội dung

- Giới thiệu về cây quyết định
- Giải thuật học của cây quyết định
- Kết luận và hướng phát triển



7

- Giới thiệu về cây quyết định
- Giải thuật học cây quyết định
- Kết luận và hướng phát triển

Cây quyết định

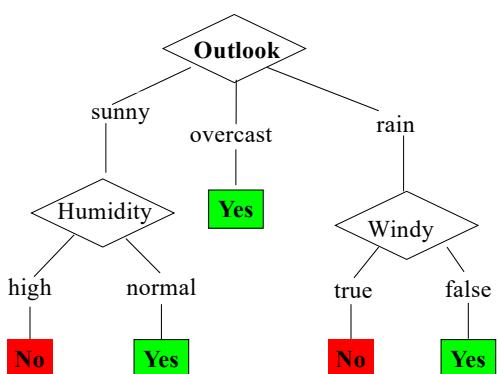
- lớp các giải thuật học
 - kết quả sinh ra dễ dịch (**if ... then ...**)
 - khá đơn giản, nhanh, hiệu quả được sử dụng nhiều
 - liên tục trong nhiều năm qua, cây quyết định được bình chọn là giải thuật được sử dụng nhiều nhất và thành công nhất
 - giải quyết các vấn đề của phân loại, hồi quy
 - làm việc cho **dữ liệu số** và **kiểu liệt kê**
 - được ứng dụng thành công trong hầu hết các lĩnh vực về phân tích dữ liệu, phân loại text, spam, phân loại gien, etc

8

Cây quyết định

- Có rất nhiều giải thuật sẵn dùng
 - ID3 (Quinlan 79)
 - **CART – Classification and Regression Trees (Brieman et al. 84)**
 - Assistant (Cestnik et al. 87)
 - **C4.5 (Quinlan 93)**
 - See5 (Quinlan 97)
 - ...
 - Orange (Demšar, Zupan 98-03)

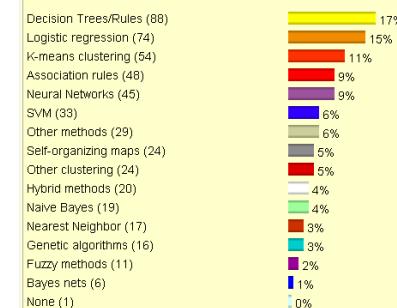
Example Decision Tree



Kỹ thuật DM thành công trong ứng dụng thực (2004)

KDnuggets : Polls : Deployed data mining techniques

Poll
Which data mining techniques you used in a successfully deployed application?
[173 voters, 509 votes total]



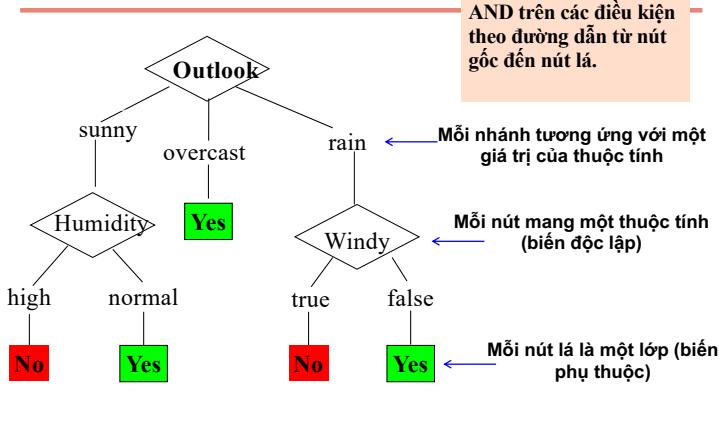
10

Cây quyết định

- **Nút trong** : được tích hợp với điều kiện để kiểm tra rẽ nhánh
- **Nút lá** : được gán nhãn tương ứng với lớp của dữ liệu
- **1 nhánh** : trình bày cho dữ liệu thỏa mãn điều kiện kiểm tra, ví dụ : age < 25.
- Ở mỗi nút, 1 thuộc tính được chọn để phân hoạch dữ liệu học sao cho tách rời các lớp tốt nhất có thể
- Một luật quyết định có dạng IF-THEN được tạo ra từ việc thực hiện AND trên các điều kiện theo đường dẫn từ nút gốc đến nút lá.
- Dữ liệu mới đến được phân loại bằng cách duyệt từ nút gốc của cây cho đến khi dừng đến nút lá, từ đó rút ra lớp của đối tượng cần xét

12

Ví dụ Decision Tree



Dữ liệu weather, dựa trên các thuộc tính (Outlook, Temp, Humidity, Windy), quyết định (play/no)

Outlook	Temp	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No

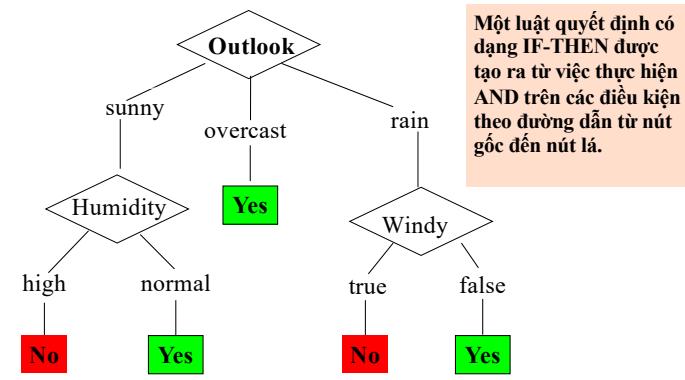
15

Nội dung

- Giới thiệu về cây quyết định
- Giải thuật học cây quyết định
- Kết luận và hướng phát triển

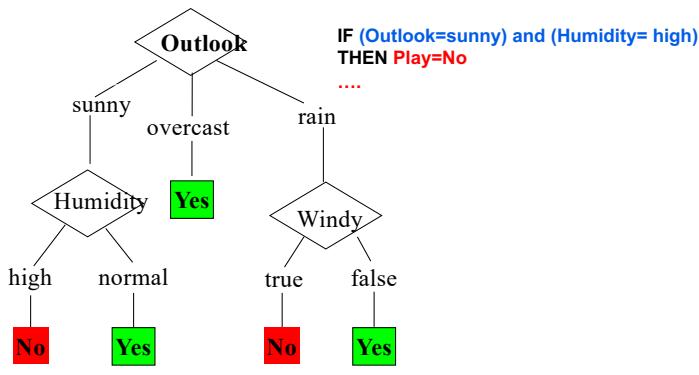
14

Cây quyết định cho tập dữ liệu weather, dựa trên các thuộc tính (Outlook, Temp, Humidity, Windy)



16

Cây quyết định cho tập dữ liệu weather,
dựa trên các thuộc tính (Outlook, Temp, Humidity, Windy)



17

Chọn thuộc tính phân hoạch

- ở mỗi nút, các thuộc tính được đánh giá dựa trên phân tách dữ liệu học **tốt nhất** có thể
 - việc đánh giá dựa trên các heuristics
 - độ lợi thông tin** (chọn thuộc tính có **chi số lớn**) - information gain (ID3/C4.5 - Quinlan)
 - Tỉ số độ lợi thông tin (information gain ratio)
 - chi số gini** (chọn thuộc tính có **chi số nhỏ**) - gini index (CART - Breiman)

19

Giải thuật cây quyết định

■ xây dựng cây Top-down

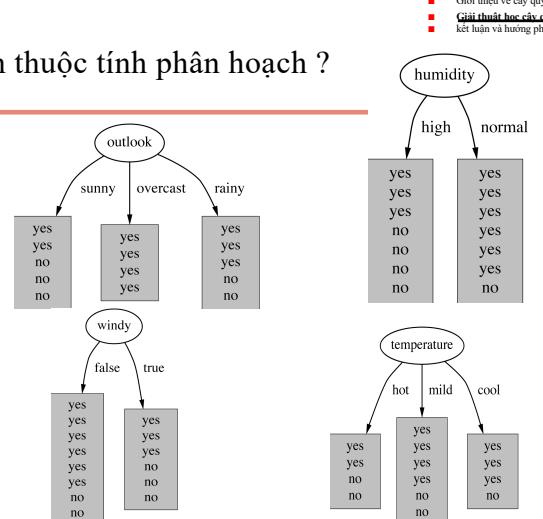
- bắt đầu nút gốc, tất cả các dữ liệu học ở nút gốc
 - Nếu dữ liệu tại 1 nút có cùng lớp -> nút lá (nhân của nút chính là nhân của các phần tử trong nút lá); Nếu dữ liệu ở nút chứa các phần tử có lớp rất khác nhau (không thuần nhất) thì phân hoạch dữ liệu một cách để quy bằng việc chọn 1 thuộc tính để thực hiện phân hoạch **tốt nhất** có thể => kết quả thu được cây nhỏ nhất

■ cắt nhánh Bottom-up

- cắt những cây con hoặc các nhánh từ dưới lên trên, để tránh học vẹt (overfitting, over learning)

18

Chọn thuộc tính phân hoạch ?



20

Chọn thuộc tính phân hoạch ?

- thuộc tính nào tốt ?
 - cho ra kết quả là cây nhỏ nhất
 - heuristics: chọn thuộc tính sinh ra các nút “purest” (thuần khiết)
- độ lợi thông tin
 - tăng với giá trị trung bình thuần khiết của các tập con của dữ liệu mà thuộc tính sinh ra
- chọn thuộc tính có độ lợi thông tin lớn nhất

21

Giới thiệu về cây quyết định
Giải thuật học cây quyết định
kết luận và hướng phát triển

Độ lợi thông tin

- Độ đo hỗn loạn trước khi phân hoạch trừ cho sau khi phân hoạch
- thông tin được đo lường bằng *bits*
 - cho 1 phân phối xác suất, thông tin cần thiết để dự đoán 1 sự kiện là *entropy* Ⓢ
- công thức tính entropy – độ hỗn loạn thông tin trước khi phân hoạch

$$Info(D) = \text{entropy}(p_1, p_2, \dots, p_n) = -p_1 \log p_1 - p_2 \log p_2 - \dots - p_n \log p_n$$

p_i: xác suất mà phần tử trong dữ liệu D thuộc lớp C_i

22

Giới thiệu về cây quyết định
Giải thuật học cây quyết định
kết luận và hướng phát triển

*Claude Shannon

Born: 30 April 1916
Died: 23 February 2001

"Father of information theory"



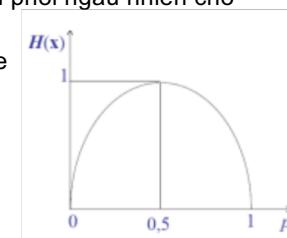
23

Claude Shannon © Stan Rowin

Giới thiệu về cây quyết định
Giải thuật học cây quyết định
kết luận và hướng phát triển

Entropy

- Entropy là một đại lượng toán học dùng để đo lượng thông tin không chắc chắn (hay lượng ngẫu nhiên) của một sự kiện hay một phân phối ngẫu nhiên cho trước
- Entropy – uncertainty measure
- Entropy luôn ≥ 0
 - Entropy = 0?
 - Entropy = 1?



$$Info(D) = \text{entropy}(p_1, p_2, \dots, p_n) = -p_1 \log p_1 - p_2 \log p_2 - \dots - p_n \log p_n$$

■ p_i: xác suất mà phần tử trong dữ liệu D thuộc lớp C_i

24

Entropy

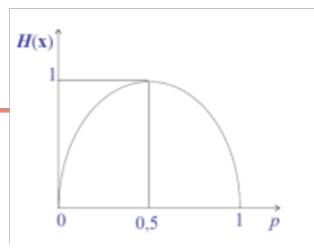
p: # phần tử có nhãn +
n: # phần tử có nhãn -

$$I\left(\frac{p}{p+n}, \frac{n}{p+n}\right) = -\frac{p}{p+n} \log_2\left(\frac{p}{p+n}\right) - \frac{n}{p+n} \log_2\left(\frac{n}{p+n}\right)$$

$p = n = 6$;
Entropy (0.5,0.5) = $-0.5 \log_2(0.5) - 0.5 \log_2(0.5) = 1$

Entropy = 1
(cực đại khi xác suất xuất hiện của các thành phần bằng nhau 50/50)

25



- Giới thiệu về cây quyết định
- Giải thuật học cây quyết định
- Kết luận và hướng phát triển

Độ lợi thông tin

- Độ hỗn loạn thông tin **trước** khi phân hoạch

$$Info(D) = \text{entropy}(p_1, p_2, \dots, p_n) = -p_1 \log p_1 - p_2 \log p_2 - \dots - p_n \log p_n$$

p_i : xác suất mà phần tử trong dữ liệu D thuộc lớp C_i

- Độ hỗn loạn thông tin **sau** khi phân hoạch

$$Info_A(D) = D_1 / |D| * Info(D_1) + D_2 / |D| * Info(D_2) + \dots + D_v / |D| * Info(D_v)$$

Thuộc tính A phân hoạch dữ liệu D thành v phần

- Độ lợi thông tin khi chọn thuộc tính A phân hoạch dữ liệu D thành v phần

$$Gain(A) = Info(D) - Info_A(D)$$

26

Ví dụ : thuộc tính outlook

- Giới thiệu về cây quyết định
- Giải thuật học cây quyết định
- Kết luận và hướng phát triển

Outlook	Temperature	Humidity	Windy	Play?
sunny	hot	high	false	No
sunny	hot	high	true	No
overcast	hot	high	false	Yes
rain	mild	high	false	Yes
rain	cool	normal	false	Yes
rain	cool	normal	true	No
overcast	cool	normal	true	Yes
sunny	mild	high	false	No
sunny	cool	normal	false	Yes
rain	mild	normal	false	Yes
sunny	mild	normal	true	Yes
overcast	mild	high	true	Yes
overcast	hot	normal	false	Yes
rain	mild	high	true	No

27

$$Info_A(D) = D_1 / |D| * Info(D_1) + D_2 / |D| * Info(D_2) + \dots + D_v / |D| * Info(D_v)$$

Ví dụ : thuộc tính outlook

- Độ hỗn loạn thông tin sau khi chọn thuộc tính A= Outlook phân hoạch dữ liệu D thành v=3 phần

- “Outlook” = “Sunny”:

$$\text{info}([2,3]) = \text{entropy}(2/5, 3/5) = -2/5 \log(2/5) - 3/5 \log(3/5) = 0.971 \text{ bits}$$

chú ý : $\log(0)$
không xác định

$$\text{info}([4,0]) = \text{entropy}(1,0) = -1 \log(1) - 0 \log(0) = 0 \text{ bits}$$

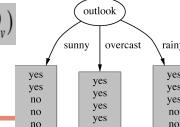
nhưng $0 * \log(0)$

$$\text{info}([3,2]) = \text{entropy}(3/5, 2/5) = -3/5 \log(3/5) - 2/5 \log(2/5) = 0.971 \text{ bits}$$

là 0

$$\text{info}([3,2], [4,0], [3,2]) = (5/14) \times 0.971 + (4/14) \times 0 + (5/14) \times 0.971$$

$$= 0.693 \text{ bits}$$



28

Độ lợi thông tin

- Độ hỗn loạn thông tin trước khi phân hoạch

$$\text{info}([9,5]) = \text{entropy}(9/14,5/14) = -9/14 \log(9/14) - 5/14 \log(5/14) = 0.940 \text{ bits}$$

- độ lợi thông tin của outlook

(trước khi phân hoạch) – (sau khi phân hoạch)

$$\begin{aligned} \text{gain("Outlook")} &= \text{info}([9,5]) - \text{info}([2,3],[4,0],[3,2]) = 0.940 - 0.693 \\ &= 0.247 \text{ bits} \end{aligned}$$

29

Giới thiệu về cây quyết định
Giải thuật học cây quyết định
kết luận và hướng phát triển

Ví dụ : thuộc tính outlook

- Độ hỗn loạn thông tin sau khi chọn thuộc tính A
phân hoạch dữ liệu D thành v=3 phần

- "Outlook" = "Sunny":**

$$\text{info}([2,3]) = \text{entropy}(2/5,3/5) = -2/5 \log(2/5) - 3/5 \log(3/5)$$

- "Outlook" = "Overcast":**

$$\text{info}([4,0]) = \text{entropy}(1,0) = -1 \log(1) - 0 \log(0) = 0 \text{ bits}$$

- "Outlook" = "Rainy":**

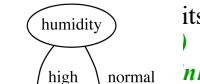
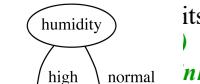
$$\text{info}([3,2]) = \text{entropy}(3/5,2/5) = -3/5 \log(3/5) - 2/5 \log(2/5) = 0.971 \text{ bits}$$

- thông tin của thuộc tính outlook:**

$$\begin{aligned} \text{info}([3,2],[4,0],[3,2]) &= (5/14) \times 0.971 + (4/14) \times 0 + (5/14) \times 0.971 \\ &= 0.693 \text{ bits} \end{aligned}$$

30

Giới thiệu về cây quyết định
Giải thuật học cây quyết định
kết luận và hướng phát triển



its
+
**nh
(0)**

its

its

Ví dụ : thuộc tính outlook

- Độ hỗn loạn thông tin sau khi chọn thuộc tính A
phân hoạch dữ liệu D thành v=3 phần

- "Outlook" = "Sunny":**

$$\text{info}([2,3]) = \text{entropy}(2/5,3/5) = -2/5 \log(2/5) - 3/5 \log(3/5)$$

- "Outlook" = "Overcast":**

$$\text{info}([4,0]) = \text{entropy}(1,0) = -1 \log(1) - 0 \log(0) = 0 \text{ bits}$$

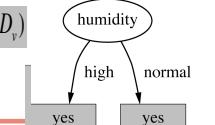
- "Outlook" = "Rainy":**

$$\text{info}([3,2]) = \text{entropy}(3/5,2/5) = -3/5 \log(3/5) - 2/5 \log(2/5) = 0.971 \text{ bits}$$

- thông tin của thuộc tính outlook:**

$$\begin{aligned} \text{info}([3,2],[4,0],[3,2]) &= (5/14) \times 0.971 + (4/14) \times 0 + (5/14) \times 0.971 \\ &= 0.693 \text{ bits} \end{aligned}$$

31



Thuộc tính humidity

- “Humidity” = “High”:

$$\text{info}([3,4]) = \text{entropy}(3/7,4/7) = -3/7\log(3/7) - 4/7\log(4/7) = 0.985 \text{ bits}$$

- “Humidity” = “Normal”:

$$\text{info}([6,1]) = \text{entropy}(6/7,1/7) = -6/7\log(6/7) - 1/7\log(1/7) = 0.592 \text{ bits}$$

- thông tin của thuộc tính humidity

$$\text{info}([3,4],[6,1]) = (7/14) \times 0.985 + (7/14) \times 0.592 = 0.788 \text{ bits}$$

- độ lợi thông tin của thuộc tính humidity**

$$\text{info}([9,5]) - \text{info}([3,4],[6,1]) = 0.940 - 0.788 = 0.152$$

33

Độ lợi thông tin

- độ lợi thông tin của các thuộc tính
 (trước khi phân hoạch) – (sau khi phân hoạch)

$$\text{gain("Outlook")} = 0.247 \text{ bits}$$

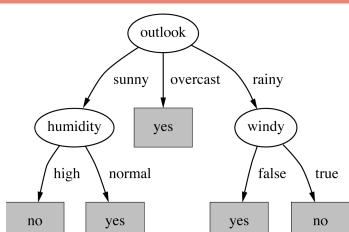
$$\text{gain("Temperature")} = 0.029 \text{ bits}$$

$$\text{gain("Humidity")} = 0.152 \text{ bits}$$

$$\text{gain("Windy")} = 0.048 \text{ bits}$$

34

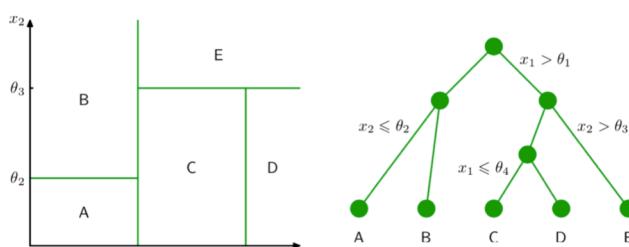
Kết quả



- chú ý : có thể có nút lá không thuần khiết

⇒ phân hoạch dừng khi dữ liệu không thể phân hoạch, nhãn được gán cho lớp lớn nhất chứa trong nút lá

35



Chỉ số gini (CART)

- nếu dữ liệu T có n lớp, chỉ số gini(T) được định nghĩa như sau :

p_j là xác suất của lớp j trong T

$$gini(T) = 1 - \sum_{j=1}^n p_j^2$$

- gini(T) là nhỏ nhất nếu những lớp trong T bị lệch

37

Chỉ số gini (CART)

- sau khi phân hoạch T thành 2 tập con T1 & T2 với kích thước N1 & N2, chỉ số gini

$$gini_{split}(T) = \frac{N_1}{N} gini(T_1) + \frac{N_2}{N} gini(T_2)$$

- thuộc tính có **gini_{split}(T) nhỏ nhất** được chọn để phân hoạch

39

Chỉ số gini (CART)

- nếu dữ liệu T có n lớp, chỉ số gini(T) được định nghĩa như sau :

$$gini(T) = 1 - \sum_{j=1}^n p_j^2$$

p_j là xác suất của lớp j trong T

- gini(T) là nhỏ nhất nếu những lớp trong T bị lệch

$$gini_{split}(T) = \frac{N_1}{N} gini(T_1) + \frac{N_2}{N} gini(T_2)$$

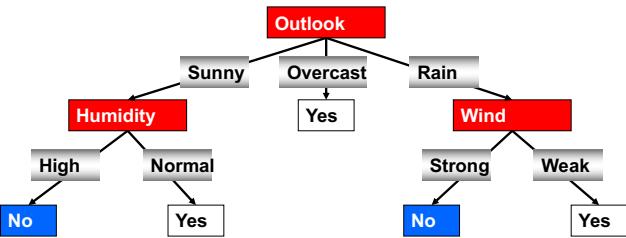
- sau khi phân hoạch T thành 2 tập con T1 & T2 với kích thước N1 & N2, chỉ số gini
- thuộc tính có **gini_{split}(T) nhỏ nhất** được chọn để phân hoạch

38

Biến đổi cây quyết định thành luật

- **Biểu diễn tri thức dưới dạng luật IF-THEN**
- **Mỗi luật tạo ra từ mỗi đường dẫn từ gốc đến lá**
- **Mỗi cặp giá trị thuộc tính đọc theo đường dẫn tạo nên phép kết (phép AND – và)**
- **Các nút lá mang tên của lớp**

Biến đổi cây quyết định thành luật



- R₁: If (Outlook=Sunny) ∧ (Humidity=High) Then Play=No
- R₂: If (Outlook=Sunny) ∧ (Humidity=Normal) Then Play=Yes
- R₃: If (Outlook=Overcast) Then Play=Yes
- R₄: If (Outlook=Rain) ∧ (Wind=Strong) Then Play=No
- R₅: If (Outlook=Rain) ∧ (Wind=Weak) Then Play=Yes

Giới thiệu về cây quyết định
Giải thuật học cây quyết định
 kết luận và hướng phát triển

Giải thuật

- giải thuật ID3/C4.5 (Quinlan, 1993)
 - sử dụng Gain ratio
 - xử lý dữ liệu số, liệt kê, nhiều
- CART (Breiman et al., 1984)
 - sử dụng chỉ số Gini
 - xử lý dữ liệu số, liệt kê, nhiều

42

Giải thuật C4.5, dữ liệu kiểu số

Giới thiệu về cây quyết định
Giải thuật học cây quyết định
 kết luận và hướng phát triển

- phân hoạch nhị phân
 - ví dụ : temp < 45
- không như dữ liệu liệt kê, dữ liệu kiểu số có nhiều nhánh phân hoạch
- phương pháp
 - tính độ lợi thông tin cho mọi giá trị phân nhánh của thuộc tính
 - chọn giá trị phân nhánh tốt nhất

43

Tập Weather, dữ liệu kiểu số

Giới thiệu về cây quyết định
Giải thuật học cây quyết định
 kết luận và hướng phát triển

Outlook	Temperature	Humidity	Windy	Play
Sunny	85	85	False	No
Sunny	80	90	True	No
Overcast	83	86	False	Yes
Rainy	75	80	False	Yes
...

```

If outlook = sunny and humidity > 83 then play = no
If outlook = rainy and windy = true then play = no
If outlook = overcast then play = yes
If humidity < 85 then play = yes
If none of the above then play = yes
  
```

44

Tập Weather, dữ liệu kiểu số

■ phân hoạch trên thuộc tính temperature

64 65 68 69 70 71 | 72 72 75 75 80 81 83 85
Yes No Yes Yes Yes No No Yes Yes Yes No Yes Yes No

- ví dụ $\text{temperature} < 71.5$: yes/4, no/2
 $\text{temperature} \geq 71.5$: yes/5, no/3

- $\text{Info}([4,2],[5,3]) = 6/14 \text{ info}([4,2]) + 8/14 \text{ info}([5,3])$
 $= 0.939 \text{ bits}$

- điểm phân hoạch : giữa
- có thể tính tất cả với 1 lần pass!
- cần sắp xếp dữ liệu

45

Giới thiệu về cây quyết định
Giải thuật học cây quyết định
 kết luận và hướng phát triển

Cải tiến

■ chỉ cản tính entropy tại các điểm thay đổi lớp (Fayyad & Irani, 1992)

giá trị lớp 64 | 65 | 68 69 70 | 71 72 | 72 75 75 | 80 | 81 83 | 85
lớp Yes | No | Yes Yes Yes | No No Yes | Yes Yes Yes | No Yes Yes | No

điểm giữa của cùng lớp không phải điểm tối ưu

46

Giới thiệu về cây quyết định
Giải thuật học cây quyết định
 kết luận và hướng phát triển

Day	Outlook	Temp.	Humidity	Wind	Golf Players
1	Sunny	Hot	High	Weak	25
2	Sunny	Hot	High	Strong	30
3	Overcast	Hot	High	Weak	46
4	Rain	Mild	High	Weak	45
5	Rain	Cool	Normal	Weak	52
6	Rain	Cool	Normal	Strong	23
7	Overcast	Cool	Normal	Strong	43
8	Sunny	Mild	High	Weak	35
9	Sunny	Cool	Normal	Weak	38
10	Rain	Mild	Normal	Weak	46
11	Sunny	Mild	Normal	Strong	48
12	Overcast	Mild	High	Strong	52
13	Overcast	Hot	Normal	Weak	44
14	Rain	Mild	High	Strong	30

Chọn thuộc tính phân hoạch ?

❖ Bài toán phân lớp

- độ lợi thông tin

- Chi số Gini*

❖ Bài toán hồi quy

- Standard deviation (độ lệch chuẩn)

$$\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2}$$

- The residual sum of squares

$$RSS = \sum_{i=1}^n (y_i - f(x_i))^2$$

<https://www.mathsisfun.com/data/standard-deviation.html>

48

Cây quyết định cho bài toán hồi quy

Day	Outlook	Temp.	Humidity	Wind	Golf Players
1	Sunny	Hot	High	Weak	25
2	Sunny	Hot	High	Strong	30
8	Sunny	Mild	High	Weak	35
9	Sunny	Cool	Normal	Weak	38
11	Sunny	Mild	Normal	Strong	48

- Số lượng người chơi golf trung bình
- Độ lệch chuẩn (Standard deviation) số lượng người chơi

Cây quyết định cho bài toán hồi quy

Day	Outlook	Temp.	Humidity	Wind	Golf Players
1	Sunny	Hot	High	Weak	25
2	Sunny	Hot	High	Strong	30
8	Sunny	Mild	High	Weak	35
9	Sunny	Cool	Normal	Weak	38
11	Sunny	Mild	Normal	Strong	48

- Số lượng người chơi golf trung bình
 $= (25 + 30 + 35 + 38 + 48)/5 = 35.2$
- Độ lệch chuẩn (Standard deviation) số lượng người chơi
 $= \sqrt{((25 - 35.2)^2 + (30 - 35.2)^2 + (35 - 35.2)^2 + (48 - 35.2)^2)/5} = 7.78$

Cây quyết định cho bài toán hồi quy

Day	Outlook	Temp.	Humidity	Wind	Golf Players
3	Overcast	Hot	High	Weak	46
7	Overcast	Cool	Normal	Strong	43
12	Overcast	Mild	High	Strong	52
13	Overcast	Hot	Normal	Weak	44

- Số lượng người chơi golf trung bình
- Độ lệch chuẩn (Standard deviation) số lượng người chơi

Cây quyết định cho bài toán hồi quy

Day	Outlook	Temp.	Humidity	Wind	Golf Players
3	Overcast	Hot	High	Weak	46
7	Overcast	Cool	Normal	Strong	43
12	Overcast	Mild	High	Strong	52
13	Overcast	Hot	Normal	Weak	44

- Số lượng người chơi golf trung bình
 $= (46 + 43 + 52 + 44)/4 = 46.25$
- Độ lệch chuẩn (Standard deviation) số lượng người chơi
 $= \sqrt{((46-46.25)^2+(43-46.25)^2+...)= 3.49}$

Cây quyết định cho bài toán hồi quy

Day	Outlook	Temp.	Humidity	Wind	Golf Players
4	Rain	Mild	High	Weak	45
5	Rain	Cool	Normal	Weak	52
6	Rain	Cool	Normal	Strong	23
10	Rain	Mild	Normal	Weak	46
14	Rain	Mild	High	Strong	30

- Số lượng người chơi golf trung bình
- Độ lệch chuẩn (Standard deviation) số lượng người chơi

Cây quyết định cho bài toán hồi quy

Day	Outlook	Temp.	Humidity	Wind	Golf Players
4	Rain	Mild	High	Weak	45
5	Rain	Cool	Normal	Weak	52
6	Rain	Cool	Normal	Strong	23
10	Rain	Mild	Normal	Weak	46
14	Rain	Mild	High	Strong	30

- Số lượng người chơi golf trung bình

$$= (45+52+23+46+30)/5 = 39.2$$
- Độ lệch chuẩn (Standard deviation) số lượng người chơi

$$= \sqrt{((45 - 39.2)^2 + (52 - 39.2)^2 + \dots)/5} = 10.87$$

Cây quyết định cho bài toán hồi quy

- Số lượng người chơi golf trung bình

$$= (25 + 30 + 46 + 45 + 52 + 23 + 43 + 35 + 38 + 46 + 48 + 52 + 44 + 30)/14 = 39.78$$
- Độ lệch chuẩn (Standard deviation) số lượng người chơi (Toàn bộ tập dữ liệu)

$$= \sqrt{((25 - 39.78)^2 + (30 - 39.78)^2 + (46 - 39.78)^2 + \dots + (30 - 39.78)^2)/14} = 9.32$$

Golf Players
25
30
46
45
52
23
43
35
38
46
48
52
44
30

Cây quyết định cho bài toán hồi quy

outlook	Stddev of Golf Players	Instances
Overcast	3.49	4
Rain	10.87	5
Sunny	7.78	5

Độ lệch chuẩn của thuộc tính Outlook

$$= (4/14) \times 3.49 + (5/14) \times 10.87 + (5/14) \times 7.78 = 7.66$$

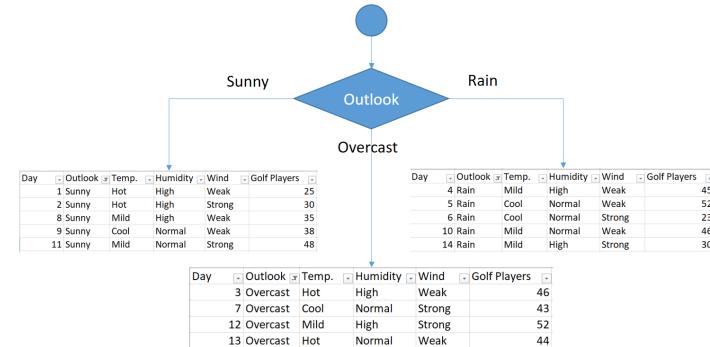
Độ chênh lệch giữa độ lệch chuẩn của toàn bộ dữ liệu và độ lệch chuẩn của thuộc tính outlook

$$= 9.32 - 7.66 = 1.66$$

Cây quyết định cho bài toán hồi quy

	Standard Deviation Reduction
Outlook	1.66
Temperature	0.47
Humidity	0.27
Wind	0.29

Cây quyết định cho bài toán hồi quy

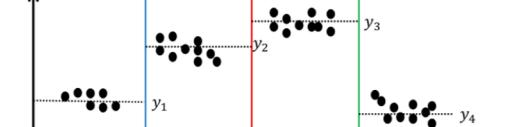


Cây quyết định cho bài toán hồi quy

Day	Outlook	Temp.	Humidity	Wind	Golf Players
1	Sunny	Hot	High	Weak	25
2	Sunny	Hot	High	Strong	30
8	Sunny	Mild	High	Weak	35
9	Sunny	Cool	Normal	Weak	38
11	Sunny	Mild	Normal	Strong	48

- Golf players for sunny outlook = {25, 30, 35, 38, 48}
- Standard deviation for sunny outlook = 7.78
- Sử dụng độ lệch chuẩn này như là độ lệch chuẩn cho toàn bộ dữ liệu của bước trước đó.

Feature Space



Regression Tree

