CÔNG NGHỆ TRI THỨC & ỨNG DỤNG
Máy học & Khám phá tri thức: Phương pháp học dựa trên cây định danh

2. Cho bảng quan sát sau đây, hãy xây dựng một cây định danh và rút ra các luật từ bảng quan sát này.

STT	TT1	TT2	TT3	Kết luận
1	D	A	F	W
2	D	A	G	W
3	E	A	F	W
4	E	A	G	W
5	D	B	F	W
6	D	B	G	W
7	E	B	F	W
8	D	C	F	Y
9	D	C	G	Y
10	E	C	F	Y
11	E	C	G	Y
12	E	B	G	Y

Giải:

Có 2 cách giải bài toán dựa vào phương pháp chọn thuộc tính phân hoạch: Theo Quinlan hoặc độ đo hỗn loạn. Ở đây ta chọn phương pháp phân hoạch theo Quinlan.

a. Định nghĩa độ đo.

- Với thuộc tính TT1, V(TT1 = D) = (T(D,W), T(D,Y))

với T(D,W) = (tổng quan sát W có TT1 là D)/tổng quan sát có TT1 là D

T(D,Y) = (tổng quan sát Y có TT1 là D)/tổng quan sát có TT1 là D

Tương tự với các thuộc tính khác.

b. Tính toán.

1. Tính lần 1.

- TT1:

V(TT1 = D) = (2/3, 1/3)

V(TT1 = E) = (1/2, 1/2)

- TT2:

V(TT2 = A) = (1, 0) (vector đơn vị)

V(TT2 = B) = (3/4, 1/4)

V(TT2 = C) = (0, 1) (vector đơn vị)

Tổng số vector đơn vị của thuộc tính TT2 là 2.

- TT3:

V(TT3 = F) = (2/3, 1/3)

V(TT3 = G) = (1/2, 1/2)

c. Tiêu chuẩn phân hoạch.

Chọn thuộc tính có nhiều vector đơn vị nhất: chọn TT2.

Sau khi phân hoạch theo TT2 xong, chỉ có phân hoạch theo TT2 = B còn chứa kết quả W, Y nên ta tiếp tục phân hoạch theo tập này. Tập dữ liệu lúc này là:

STT	TT1	TT3	Kết luận
5	D	F	W
6	D	G	W
7	E	F	W
12	E	G	Y

Tính toán độ đo lần 2:

- TT1:

V(TT1 = D) = (1, 0) (vector đơn vị)

V(TT1 = E) = (1/2, 1/2)

- TT3:

V(TT3 = F) = (1, 0) (vector đơn vị)

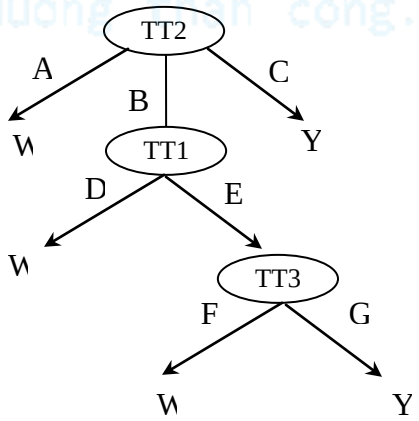
V(TT3 = G) = (1/2, 1/2)

TT1, TT3 có số vector đơn vị bằng nhau, số phân hoạch cũng bằng nhau.

i. Gia su ta chọn TT1. Tập dữ liệu lúc này còn là:

STT	TT3	Kết luận
7	F	W
12	G	Y

Ta có cây định danh cuối cùng là:



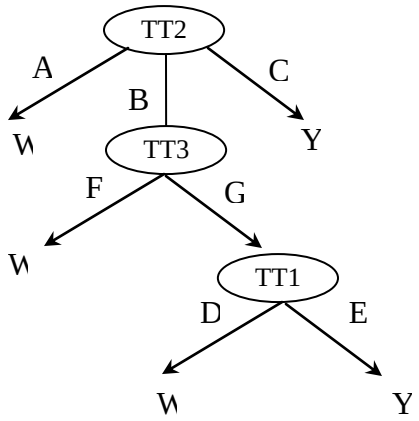
Phát sinh tập luật:

- Nếu TT2 = A hoặc (TT2 = B và TT1 = D) hoặc (TT2 = B và TT1 = E và TT3= F) thì kết luận W.
- Nếu TT2 = C hoặc (TT2 = B và TT1 = E và TT3= G) thì kết luận Y.

ii. Gia su ta chọn TT3. Tập dữ liệu lúc này còn là:

STT	TT1	Kết luận
6	D	W
12	E	Y

Ta có cây định danh cuối cùng là:



Phát sinh tập luật:

- Nếu TT2 = A hoặc (TT2 = B và TT3 = F) hoặc (TT2 = B và TT3 = G và TT1= D) thì kết luận W.
- Nếu TT2 = C hoặc (TT2 = B và TT3 = G và TT1= E) thì kết luận Y.

Rút gọn tập luật: từ i, ii ta có:

- Nếu TT2 = A hoặc (TT2 = B và TT1 = D) hoặc (TT2 = B và TT3 = F) thì kết luận W.
- Nếu TT2 = C hoặc (TT2 = B và TT3 = G và TT1= F) thì kết luận Y.

3. Sử dụng cây định danh để tìm các luật phân lớp từ bảng quyết định sau đây: (Tham khảo đề thi cao học khoa CNTT- trường ĐHKHTN)

#	TRỜI	ÁPSUẤT	GIÓ	KẾTQUẢ
1	Trong	Cao	Bắc	Không mưa
2	Mây	Cao	Nam	Mưa
3	Mây	Trung Bình	Bắc	Mưa
4	Trong	Thấp	Bắc	Không mưa
5	Mây	Thấp	Bắc	Mưa
6	Mây	Cao	Bắc	Mưa
7	Mây	Thấp	Nam	Không mưa
8	Trong	Cao	Nam	Không mưa

Giải:

a. Định nghĩa độ đo.

- Với thuộc tính TRỜI.

V(TRỜI = Trong) = (T(Trong, Không mưa), T(Trong, Mưa)) với T(Trong, Không mưa) = tổng quan sát "không mưa" khi TRỜI trong/tổng quan sát trời trong. T(Trong, Mưa) = tổng quan sát "Mưa" khi TRỜI trong/tổng quan sát trời trong.

Tương tự với các thuộc tính khác.

b. Tính toán.

Tính lần 1.

- Thuộc tính Trời:

V(Trời = Trong) = (1, 0) (vector đơn vị)
V(Trời = Mây) = (1/5, 4/5)

- Thuộc tính Áp suất:
  V(Áp suất = Cao) = (1/2, 1/2)
  V(Áp suất = Trung bình) = (0, 1) (vector đơn vị)
  V(Áp suất = Thấp) = (2/3, 1/3)
- Thuộc tính Gió:
  V(Gió = Bắc) = (2/5, 3/5)
  V(Gió = Nam) = (2/3, 1/3)

c. Tiêu chuẩn phân hoạch. Chọn thuộc tính có nhiều vector đơn vị nhất.
Thuộc tính Trời và Áp suất đều có 1 vector đơn vị. Tuy nhiên số phân hoạch của thuộc tính Trời ít hơn nên ta chọn phân hoạch theo thuộc tính này. Tập dữ liệu lúc này còn là:

#	ÁPSUẤT	GIÓ	KẾTQUẢ
2	Cao	Nam	Mưa
3	Trung Bình	Bắc	Mưa
5	Thấp	Bắc	Mưa
6	Cao	Bắc	Mưa
7	Thấp	Nam	Không mưa

Tính toán độ đo lần 2:

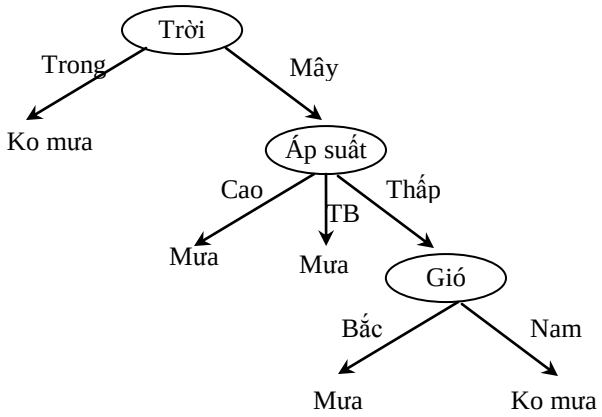
- Thuộc tính Áp suất:
  V(Áp suất = Cao) = (0, 1) (vector đơn vị)
  V(Áp suất = Trung bình) = (0, 1) (vector đơn vị)
  V(Áp suất = Thấp) = (1/2, 1/2)
- Thuộc tính Gió:
  V(Gió = Bắc) = (0, 1) (vector đơn vị)
  V(Gió = Nam) = (1/2, 1/2)

Phân hoạch:

Chọn thuộc tính có vector đơn vị nhiều nhất là Áp suất (2 vector đơn vị). Tập dữ liệu lúc này còn là

#	GIÓ	KẾTQUẢ
5	Bắc	Mưa
7	Nam	Không mưa

Ta có cây định danh cuối cùng như sau.



Phát sinh tập luật:

Nếu trời trong hoặc (trời mây và áp suất thấp và có gió Nam) thì trời không mưa.

Nếu (trời có mây và (áp suất cao hay TB)) hay (trời có mây và áp suất thấp và có gió Bắc) thì trời mưa.

4. Sử dụng cây định danh để tìm các luật phân lớp từ bảng dữ liệu sau:

#	Độ cứng	Độ ẩm	Độ pH	Kết quả
1	Trung bình	Thấp	Cao	Xấu
2	Cao	Cao	Cao	Tốt
3	Cao	Thấp	Trung bình	Tốt
4	Trung bình	Thấp	Thấp	Xấu
5	Cao	Thấp	Thấp	Tốt
6	Cao	Thấp	Cao	Tốt
7	Cao	Cao	Thấp	Xấu
8	Trung bình	Cao	Cao	Xấu
9	Cao	Trung bình	Trung bình	???

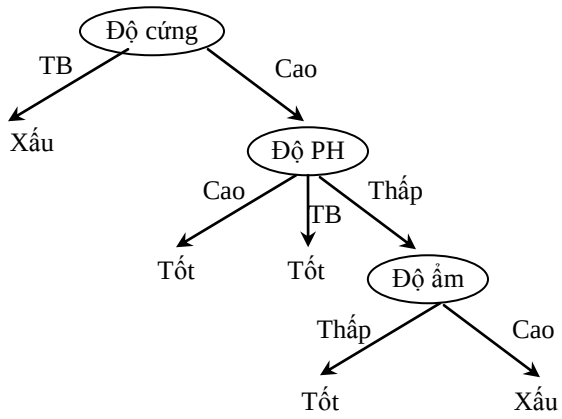
Dùng các luật phân lớp để xác định lớp của #9

Giải:

Ta sẽ tìm luật phân lớp từ bảng dữ liệu sau:

#	Độ cứng	Độ ẩm	Độ pH	Kết quả
1	Trung bình	Thấp	Cao	Xấu
2	Cao	Cao	Cao	Tốt
3	Cao	Thấp	Trung bình	Tốt
4	Trung bình	Thấp	Thấp	Xấu
5	Cao	Thấp	Thấp	Tốt
6	Cao	Thấp	Cao	Tốt
7	Cao	Cao	Thấp	Xấu
8	Trung bình	Cao	Cao	Xấu

Để ý rằng bảng dữ liệu trên giống bảng dữ liệu của bài 3 nếu thay thế thuộc tính Độ cứng = TRỜI, Độ ẩm = GIÓ, Độ PH = ÁPSUẤT đồng thời thay thế các giá trị thuộc tính tương ứng. Vì thế ta được cây định danh sau:



Vậy nếu độ cứng cao, độ ẩm trung bình và độ PH trung bình thì tốt.