KHAI THÁC DỮ LIỆU & ỨNG DỤNG *(DATA MINING)*

GV : NGUYỄN HOÀNG TÚ ANH

1

BÀI 4 – PHẦN 1
PHÂN LỚP DỮ LIỆU

2

NỘI DUNG

1. ***Giới thiệu***
2. **Phương pháp dựa trên cây quyết định**
3. **Phương pháp dựa trên luật**

3

GIỚI THIỆU

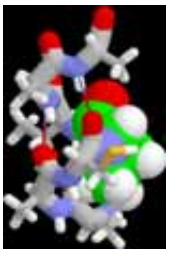
1. **Phân lớp :**
   - **Cho tập các mẫu đã phân lớp trước, xây dựng mô hình cho từng lớp**
   - ***Mục đích : Gán các mẫu mới vào các lớp với độ chính xác cao nhất có thể.***
   - **Cho CSDL $D=\{t_1,t_2,\ldots,t_n\}$ và tập các lớp $C=\{C_1,\ldots,C_m\}$, *phân lớp* là bài toán xác định ánh xạ $f : D \rightarrow C$ sao cho mỗi t_i được gán vào một lớp.**

4

GIỚI THIỆU

Ví dụ Phân lớp :

- **Phân lớp khách hàng (trong ngân hàng) để cho vay hay không**
- ***Dự đoán tế bào khối u là lành tính hay ác tính***
- **Phân loại giao dịch thẻ tín dụng là hợp pháp hay gian lận**
- ***Phân loại tin tức thuộc lĩnh vực tài chính, thời tiết, giải trí, thể thao, …***
- **Dự đoán khi nào sông có lũ**
- ***Chuẩn đoán y khoa***



5

GIỚI THIỆU

2. Qui trình phân lớp :

- ***Bước 1 : Xây dựng mô hình***
  - *Mô tả tập các lớp xác định trước*
    - Tập huấn luyện : các mẫu / bộ dành cho xây dựng mô hình
      - Mỗi mẫu/ bộ thuộc về một lớp đã định nghĩa trước
    - ***Tìm luật phân lớp, cây quyết định hoặc công thức toán mô tả lớp***

6

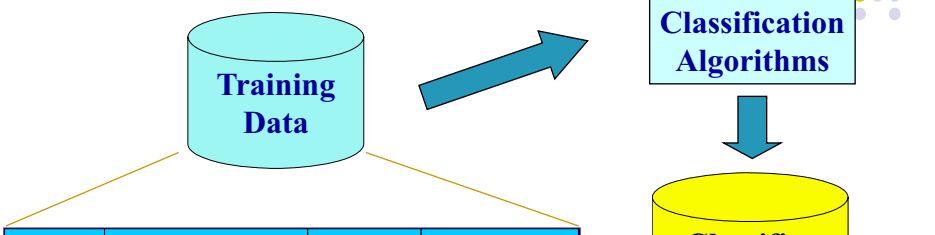
GIỚI THIỆU

2. Qui trình phân lớp (tt) :

- ***Bước 2 : Sử dụng mô hình***
  - *Phân lớp các đối tượng chưa biết*
    - Xác định độ chính xác của mô hình, sử dụng tập DL kiểm tra độc lập
    - Độ chính xác chấp nhận được -> áp dụng mô hình để phân lớp các mẫu/bộ chưa xác định được nhãn lớp

7

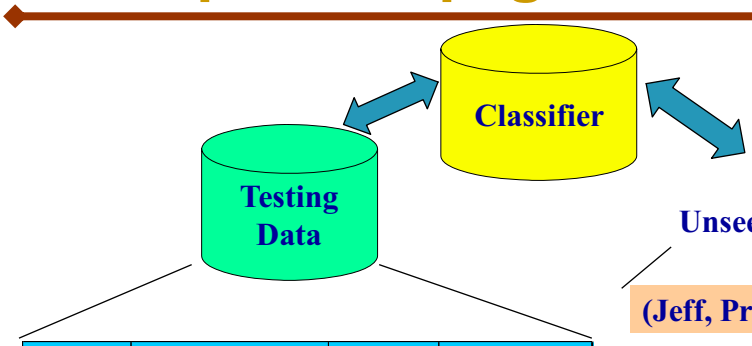
Ví dụ : XD mô hình



NAME	RANK	YEARS	TENURED
Mike	Assistant Prof	3	no
Mary	Assistant Prof	7	yes
Bill	Professor	2	yes
Jim	Associate Prof	7	yes
Dave	Assistant Prof	6	no
Anne	Associate Prof	3	no

8

Ví dụ : Sử dụng mô hình



NAME	RANK	YEARS	TENURED
Tom	Assistant Prof	2	no
Merlisa	Associate Prof	7	no
George	Professor	5	yes
Joseph	Assistant Prof	7	yes

(Jeff, Professor, 4)

Tenured? Yes

9

GIỚI THIỆU

3. Các kỹ thuật phân lớp :

- *Phương pháp dựa trên cây quyết định*
- **Phương pháp dựa trên luật**
- ***Phương pháp Naïve Bayes***
- **Phương pháp dựa trên thể hiện**
- **Mạng Nơron**
- **SVM (support vector machine)**
- **Tập thô**

10

NỘI DUNG

1. Giới thiệu
2. ***Phương pháp dựa trên cây quyết định***
3. Phương pháp dựa trên luật

11

CÂY QUYẾT ĐỊNH

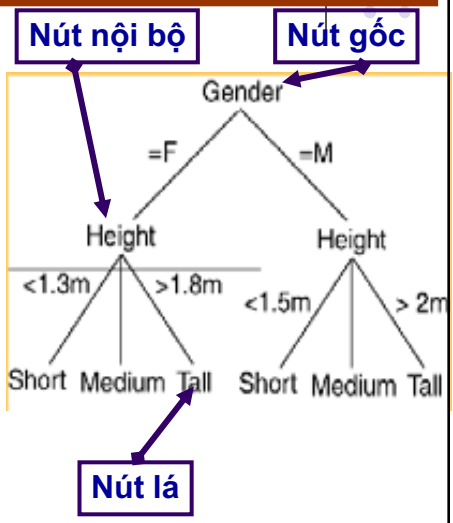
1. Định nghĩa
2. Xây dựng cây quyết định
3. Thuật toán xây dựng cây quyết định
4. Cách phân chia mẫu
   - Độ đo để lựa chọn thuộc tính
5. Vấn đề quá phù hợp với DL
6. Ưu điểm

12

CÂY QUYẾT ĐỊNH

1. Định nghĩa

- Cây quyết định là một cấu trúc phân cấp của các nút và các nhánh
- 3 loại nút trên cây:
  - Nút gốc
  - Nút nội bộ : mang tên thuộc tính của CSDL
  - Nút lá : mang tên lớp C_i
- **Nhánh : mang giá trị của thuộc tính**



13

CÂY QUYẾT ĐỊNH

2. Xây dựng cây quyết định

- **Gồm 2 bước :**
  - ***Bước 1 : Thiết lập cây quyết định***
    - Bắt đầu từ gốc
    - Kiểm tra các giá trị của thuộc tính và phân chia các mẫu đệ qui
  - ***Bước 2 : Tỉa bớt cây***
    - Xác định và loại bỏ bớt các nhánh không ổn định hoặc cá biệt

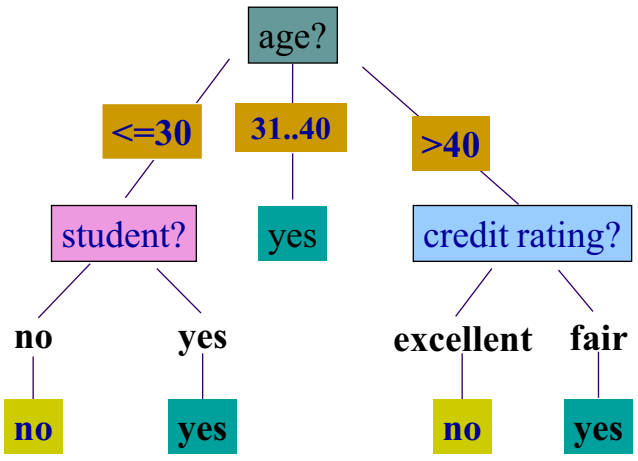
14

VÍ DỤ 1: Dữ liệu huấn luyện

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

15

VÍ DỤ 1 : CÂY QUYẾT ĐỊNH



16

CÂY QUYẾT ĐỊNH

3. Thuật toán xây dựng cây quyết định

- Hunt's Algorithm
- CART
- ID3, C4.5
- SLIQ, SPRINT

17

CÂY QUYẾT ĐỊNH

3. Thuật toán xây dựng cây quyết định

- **Ý tưởng chính :**
  - *Phương pháp "tham lam" (greedy)*
  - *Phân chia tập mẫu dựa trên thuộc tính cho kết quả tối ưu hóa tiêu chuẩn*
- ***Vấn đề :***
  - *Xác định cách phân chia các mẫu*
    - *Dựa trên độ đo sự đồng nhất của dữ liệu*
  - *Điều kiện dừng*

18

CÂY QUYẾT ĐỊNH

3. Thuật toán xây dựng cây quyết định (tt)

- ***Điều kiện dừng*** :
  - Tất cả các mẫu rơi vào một nút thuộc về cùng một lớp
  - *Không còn thuộc tính nào có thể dùng để phân chia mẫu nữa*
  - Không còn lại mẫu nào tại nút

19

CÂY QUYẾT ĐỊNH

4. Cách phân chia các mẫu

- Tiêu chuẩn phân chia : tạo ra các nhóm sao cho một lớp chiếm ưu thế trong từng nhóm
- *Thuộc tính được chọn là thuộc tính cho độ đo tốt nhất, có lợi nhất cho quá trình phân lớp*
- Độ đo để đánh giá chất lượng phân chia là độ đo sự đồng nhất
  - Entropy (Information Gain)
  - Information Gain Ratio
  - Gini Index

20

CÂY QUYẾT ĐỊNH

- ***Độ lợi thông tin (Information gain) :*** **ID3 / C4.5**
  - ***Chọn thuộc tính có độ lợi thông tin cao nhất***
  - **Giả sử :**
    - **D : tập huấn luyện**
    - **$C_{i,D}$: tập các mẫu của D thuộc lớp C_i với i = {1, …, m}**
    - **$|C_{i,D}|$, |D| : lực lượng của tập $C_{i,D}$ và D tương ứng**
  - ***p_i là xác suất để một mẫu bất kỳ của D thuộc về lớp C_i***
  - **Thông tin kỳ vọng để phân lớp một mẫu trong D là :**

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$

$$p_i = \frac{|C_{i,D}|}{|D|}$$

21

VÍ DỤ 1: Dữ liệu huấn luyện

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

22

CÂY QUYẾT ĐỊNH

- ***Độ lợi thông tin (Information gain) :***
  - **Trong VD1 : 14 mẫu tin, trong đó có 9 mua máy tính**
    - **$|D| = 14$; $m = 2$; C_1 ="Mua "; C_2 ="Không mua"**
    - **$|C_{1,D}| = 9$, $|C_{2,D}| = 5$**
  - **Thông tin kỳ vọng để phân lớp một mẫu trong D là :**

$$Info(D) = I(9,5) = -\frac{9}{14}\log_2\frac{9}{14} - \frac{5}{14}\log_2\frac{5}{14} = 0.940$$

23

CÂY QUYẾT ĐỊNH

- **Thuộc tính A có các giá trị :{a_1, a_2, ...,a_v}**
- ***Dùng thuộc tính A để phân chia tập huấn luyện D thành v tập con {D_1, D_2, ..., D_v}***
- **Thông tin cần thiết để phân chia D theo thuộc tính A :**

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} I(D_j)$$

  - **Độ lợi thông tin (information gain) dựa trên phân chia theo thuộc tính A :**

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

24

VÍ DỤ 1 : INFORMATION GAIN

- Ký hiệu :
  - Lớp P: buys_computer = "Yes"
  - Lớp N: buys_computer = "No"
- *Info(D) = I(9, 5) =0.940*
- *Tính độ lợi thông tin cho thuộc tính "age" ?*

age	p_j	n_j	$I(p_j, n_j)$
<=30	2	3	0.971
31…40	4	0	0
>40	3	2	0.971

$$I(2,3) = -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5} = 0.971$$

$$I(4,0) = -\frac{4}{4}\log_2\frac{4}{4} - \frac{0}{4}\log_2\frac{0}{4} = 0$$

$$I(3,2) = -\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5} = 0.971$$

VÍ DỤ 1 : INFORMATION GAIN

- *Tính độ lợi thông tin cho thuộc tính "age" ?*
- Khi đó :

$$Info_{age}(D) = \frac{5}{14}I(2,3) + \frac{4}{14}I(4,0) + \frac{5}{14}I(3,2) = 0.694$$

- ***Suy ra :***

$$Gain(age) = Info(D) - Info_{age}(D) = 0.246$$

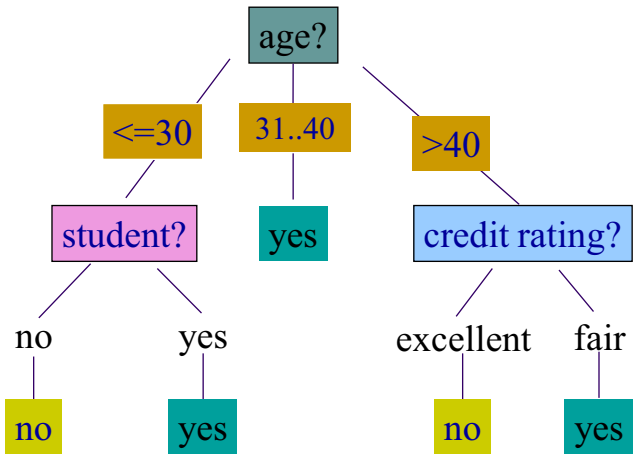
26

BÀI TẬP

- **Thời gian : 10’**
- **Cho tập DL như trong ví dụ 1**
- **Ký hiệu :**
  - **Lớp P: buys_computer = “Yes”**
  - **Lớp N: buys_computer = “No”**
- **Tính độ lợi thông tin dựa trên phân chia theo thuộc tính**
  - ***“income” : dãy giữa***
  - ***“student” : dãy trái***
  - ***“credit_rating” : dãy phải***

27

VÍ DỤ 1 : IG



28

CÂY QUYẾT ĐỊNH

Information Gain Ratio: C4.5

- Độ đo Gain có xu hướng thiên vị cho các thuộc tính có nhiều giá trị -> cần chuẩn hóa độ đo Gain
- ***Chọn thuộc tính có độ đo Gain Ratio lớn nhất***
- **GainRatio(A) = Gain(A)/SplitInfo$_A$(D)**

$$SplitInfo_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \log_2\left(\frac{|D_j|}{|D|}\right)$$

29

CÂY QUYẾT ĐỊNH

Chỉ mục Gini (Gini index) : CART, SLIQ, SPRINT

- **Tập huấn luyện *D* chứa các mẫu của *m* lớp.**
- ***Chỉ mục Gini* của tập D – gini(D) là :**

$$gini(D) = 1 - \sum_{i=1}^{m} p_i^2$$

với p_i là tần suất của lớp C_i trong D

- **Cho tập DL của ví dụ 1, ta có *gini(D)* là :**

$$gini(D) = 1 - \left(\frac{9}{14}\right)^2 - \left(\frac{5}{14}\right)^2 = 0.459$$

30

CÂY QUYẾT ĐỊNH

- ***Chỉ mục Gini (Gini index) :***
  - Thuộc tính *A* có các giá trị :{a_1, a_2, …,a_v}
  - Dùng thuộc tính A để phân chia tập huấn luyện D thành *v* tập con {D_1, D_2, …, D_v}
  - ***Chỉ mục Gini*** của phân chia D theo thuộc tính A :

$$gini_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} gini(D_j)$$

  - ***Tại mỗi cấp, chúng ta chọn thuộc tính có chỉ mục Gini nhỏ nhất để phân chia tập dữ liệu***

31

VÍ DỤ 1 : GINI INDEX

- Lớp P: buys_computer = "Yes"
- Lớp N: buys_computer = "No"
- *gini(D) = 0.459*
- *Tính chỉ mục gini cho thuộc tính "age" ?*
- ***Suy ra :***

age	p_j	n_j	gini(p_j, n_j)
<=30	2	3	0.48
31…40	4	0	0
>40	3	2	0.48

$$gini_{age}(D) = \frac{5}{14} gini(2,3) + \frac{4}{14} gini(4,0) + \frac{5}{14} gini(3,2)$$
$$= 0.343$$

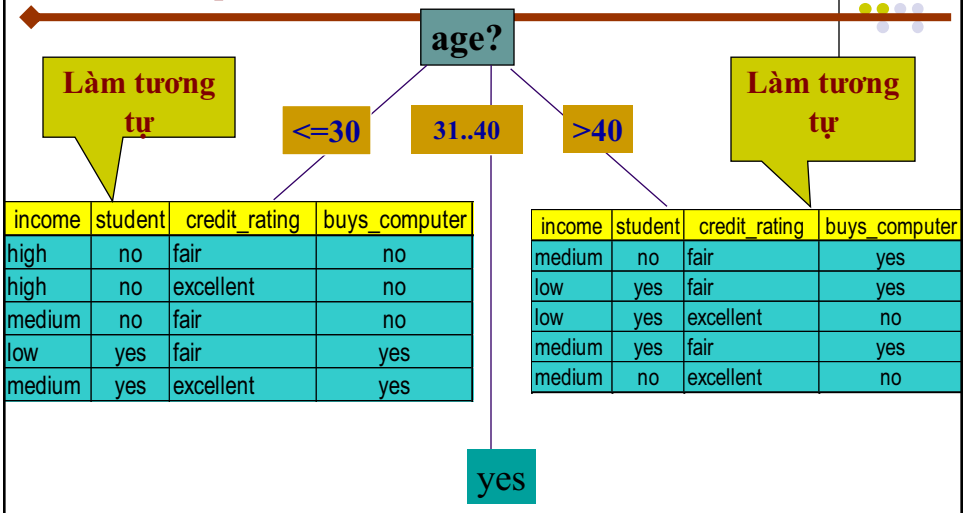
32

Ví dụ : GINI INDEX

- Sau khi tính độ đo chỉ mục Gini dựa trên phân chia theo thuộc tính :
  - $Gini_{age}(D) = 0.343$
  - $Gini_{income}(D) = 0.44$
  - $Gini_{student}(D) = 0.367$
  - $Gini_{credit\_rating}(D) = 0.429$
- ***Độ đo chỉ mục Gini dựa trên phân chia theo thuộc tính "age" là nhỏ nhất nên ta sẽ chia DL theo thuộc tính "age"***

33

Ví dụ : GINI INDEX



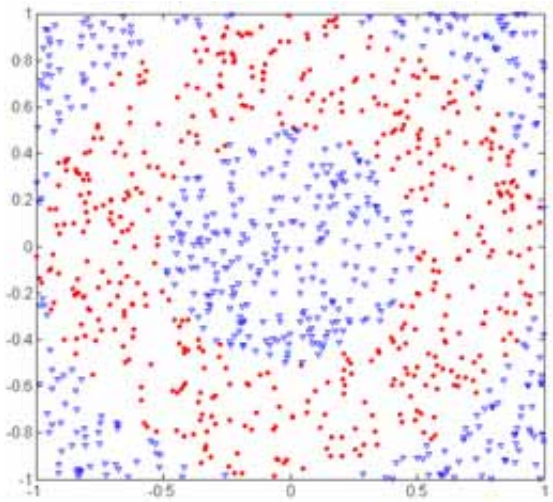
income	student	credit_rating	buys_computer
high	no	fair	no
high	no	excellent	no
medium	no	fair	no
low	yes	fair	yes
medium	yes	excellent	yes

income	student	credit_rating	buys_computer
medium	no	fair	yes
low	yes	fair	yes
low	yes	excellent	no
medium	yes	fair	yes
medium	no	excellent	no

34

CÂY QUYẾT ĐỊNH

5. Vấn đề quá phù hợp với DL (overfitting)



- Có 500 điểm DL hình tròn and 500 điểm hình tam giác.

Các điểm hình tròn :

$0.5 \leq \text{sqrt}(x_1^2+x_2^2) \leq 1$

Các điểm hình tam giác:

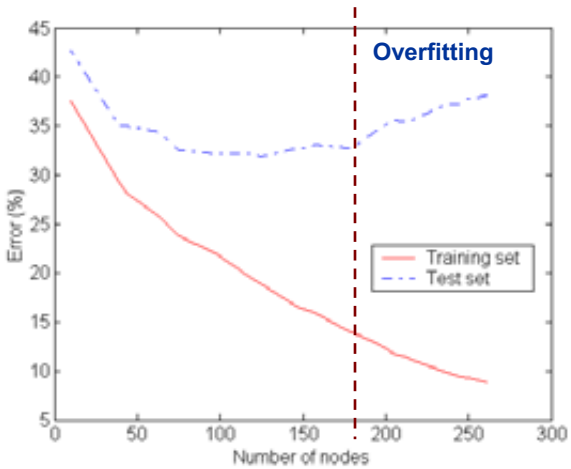
$\text{sqrt}(x_1^2+x_2^2) > 0.5$ or

$\text{sqrt}(x_1^2+x_2^2) < 1$

35

CÂY QUYẾT ĐỊNH

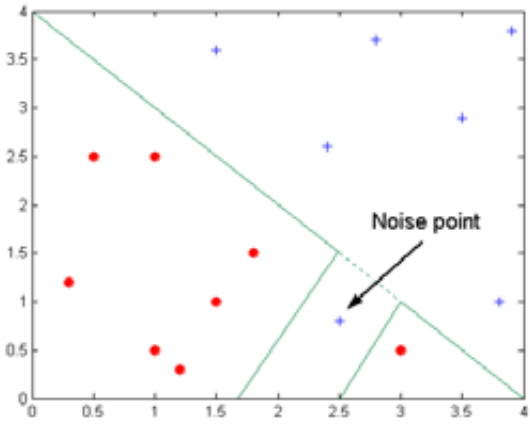
5. Vấn đề quá phù hợp với DL (overfitting)



36

CÂY QUYẾT ĐỊNH

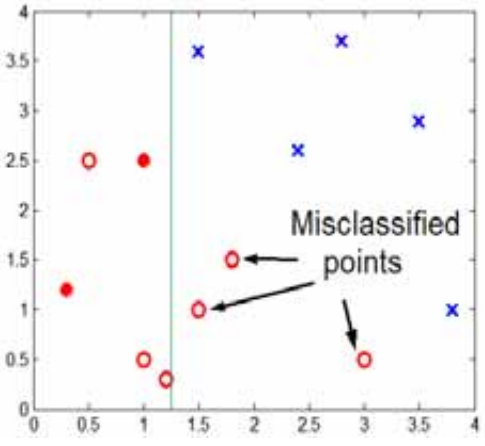
- **Cây tạo ra có thể quá phù hợp với DL huấn luyện :**
  - **Quá nhiều nhánh do nhiễu hoặc cá biệt**



37

CÂY QUYẾT ĐỊNH

- Cây tạo ra có thể quá phù hợp với DL huấn luyện do thiếu mẫu
  - *Thiếu các điểm DL ở nửa dưới của biểu đồ gây khó khăn cho việc dự đoán lớp chính xác của vùng này.*



38

CÂY QUYẾT ĐỊNH

- **Kết quả :** độ chính xác kém khi phân lớp cho mẫu mới
- Hai phương pháp tránh quá **PHÙ HỢP DL** :
  - ***Loại bỏ trước :*** Dừng thêm nhánh cây sớm, ngay khi nó có thể tạo ra độ đo dưới ngưỡng nào đó
    - Rất khó chọn ngưỡng thích hợp
  - ***Loại bỏ sau :*** Loại bớt nhánh từ cây hoàn chỉnh (từ dưới lên)
    - Sử dụng tập DL độc lập để kiểm tra và loại bớt
- Xác định chính xác kích thước cây kết quả như thế nào ?
  - Phân chia : tập huấn luyện (2/3), tập test (1/3)
  - Sử dụng đánh giá chéo (cross-validation)

39

CÂY QUYẾT ĐỊNH

6. Ưu điểm :

- **Dễ dàng xây dựng cây**
- ***Phân lớp mẫu mới nhanh***
- **Dễ dàng diễn giải cho các cây có kích thước nhỏ**
- ***Độ chính xác chấp nhận được so với các kỹ thuật phân lớp khác trên nhiều tập DL đơn***

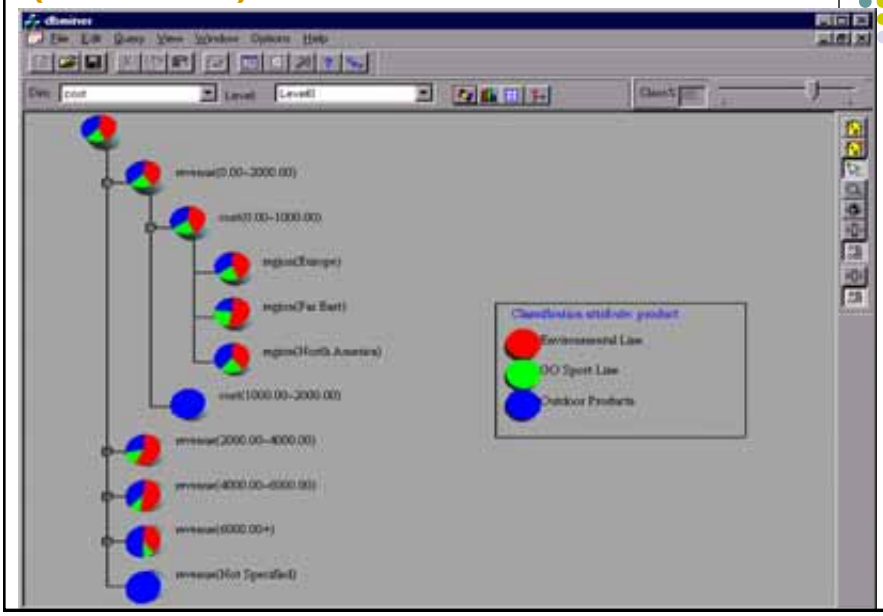
Ví dụ : C4.5

- ***Dùng độ lợi thông tin***
- ***Sắp xếp thuộc tính số tại từng nút***
- ***Yêu cầu toàn bộ DL chứa vừa trong bộ nhớ***
- ***Không phù hợp với các tập dữ liệu lớn***

http://www.cse.unsw.edu.au/~quinlan/c4.5r8.tar.gz

40

VD : Biểu diễn kết quả phân loại (DBMiner)



41

NỘI DUNG

1. Giới thiệu
2. Phương pháp dựa trên cây quyết định
3. ***Phương pháp dựa trên luật***

42

GIỚI THIỆU

- *Sử dụng các luật IF-THEN để phân loại*
- *Luật có dạng : IF (Điều kiện) Then Y*
  - *Với "Điều kiện " : kết hợp các thuộc tính*
  - *Y là nhãn lớp*
  - *Ví dụ :*

  IF *age* = "youth" AND *student* = "*yes*" THEN *buys_computer* = "*yes*"
- *Luật R phủ một mẫu x nếu các thuộc tính của mẫu thỏa mãn điều kiện của luật*

43

GIỚI THIỆU

- *Độ phủ của luật : coverage(R)*
  - *Tỷ lệ các mẫu thỏa mãn điều kiện (vế trái) của luật*
- *Độ chính xác của luật : accuracy(R)*
  - *Tỷ lệ các mẫu thỏa mãn cả điều kiện và kết luận (2 vế trái, phải) của luật*
- **Ví dụ :**
  - *IF* **(Marital Status=Single) → No**
  - **Coverage(R) = 40%**
  - **Accuracy(R) = 50%**

Tid	Refund	Marital Status	Taxable Income	Class
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

44

VÍ DỤ 2

Cho tập DL huấn luyện sau :

Name	Blood Type	Give Birth	Can Fly	Live in Water	Class
human	warm	yes	no	no	mammals
python	cold	no	no	no	reptiles
salmon	cold	no	no	yes	fishes
whale	warm	yes	no	yes	mammals
frog	cold	no	no	sometimes	amphibians
komodo	cold	no	no	no	reptiles
bat	warm	yes	yes	no	mammals
pigeon	warm	no	yes	no	birds
cat	warm	yes	no	no	mammals
leopard shark	cold	yes	no	yes	fishes
turtle	cold	no	no	sometimes	reptiles
penguin	warm	no	no	sometimes	birds
porcupine	warm	yes	no	no	mammals
eel	cold	no	no	yes	fishes
salamander	cold	no	no	sometimes	amphibians
gila monster	cold	no	no	no	reptiles
platypus	warm	no	no	no	mammals
owl	warm	no	yes	no	birds
dolphin	warm	yes	no	yes	mammals
eagle	warm	no	yes	no	birds

45

VÍ DỤ 2

- ***Tập luật :***

  R1: (Give Birth = no) $\wedge$ (Can Fly = yes) $\rightarrow$ Birds

  R2: (Give Birth = no) $\wedge$ (Live in Water = yes) $\rightarrow$ Fishes

  R3: (Give Birth = yes) $\wedge$ (Blood Type = warm) $\rightarrow$ Mammals

  R4: (Give Birth = no) $\wedge$ (Can Fly = no) $\rightarrow$ Reptiles

  R5: (Live in Water = sometimes) $\rightarrow$ Amphibians

- ***Sử dụng tập luật để xác định lớp cho các mẫu mới sau :***

Name	Blood Type	Give Birth	Can Fly	Live in Water	Class
lemur	warm	yes	no	no	?
turtle	cold	no	no	sometimes	?
dogfish shark	cold	yes	no	yes	?

46

VÍ DỤ 2

Name	Blood Type	Give Birth	Can Fly	Live in Water	Class
lemur	warm	yes	no	no	?
turtle	cold	no	no	sometimes	?
dogfish shark	cold	yes	no	yes	?

- Mẫu "lemur" phủ bởi luật R3, nên được phân vào lớp "Mammals"
- Mẫu "turtle" phủ bởi cả luật R4 và R5
- Mẫu "dogfish shark" không được phủ bởi bất kỳ luật nào.
- ***Cách giải quyết ?***

47

GIỚI THIỆU

- ***Cách giải quyết ?***
  - ***Xếp hạng các luật theo độ ưu tiên:***
    - ***Theo kích thước của luật : các luật có tập điều kiện lớn hơn sẽ có độ ưu tiên cao hơn***
    - **Theo luật : các luật được xếp hạng theo độ đo chất lượng luật hoặc theo ý kiến chuyên gia**
    - ***Theo lớp : gom các luật thuộc cùng một lớp***
  - **Nếu một mẫu được phủ bởi nhiều luật thì chọn luật có thứ hạng cao nhất**
  - **Nếu không phủ bởi bất kỳ luật nào thì gán vào lớp mặc định**

48

GIỚI THIỆU

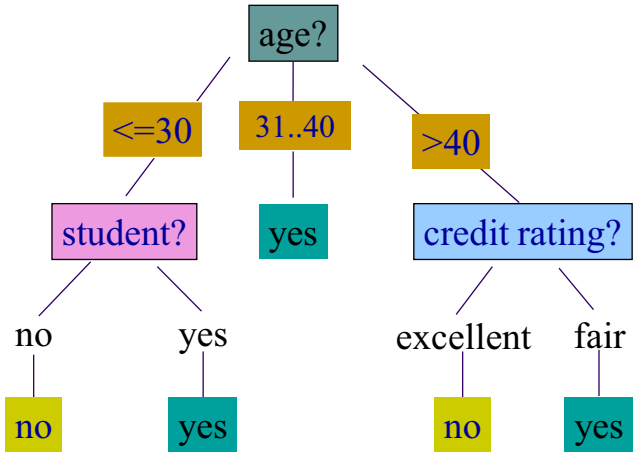
- **Xây dựng luật phân lớp** *:*
  - *Phương pháp trực tiếp :*
    - *Rút các luật trực tiếp từ dữ liệu*
    - *RIPPER, CN2, ILA, FOIL, AQ, …*
  - **Phương pháp gián tiếp :**
    - **Rút luật từ các mô hình phân lớp khác như cây quyết định, mạng nơron, …**
    - **Luật C4.5**

GiỚI THIỆU

Phương pháp gián tiếp :

- **Rút trích luật từ cây quyết định**
- ***Luật tạo ra từ từng đường dẫn từ gốc đến lá***
- **Mỗi cặp giá trị thuộc tính dọc theo đường dẫn tạo nên phép kết**
- ***Các nút lá mang tên của lớp***
- **Ví dụ 1**

VÍ DỤ 1 : CÂY QUYẾT ĐỊNH



51

GiỚI THIỆU

Phương pháp gián tiếp :

- Các luật của Ví dụ 1:

IF *age* = "<=30" AND *student* = "*no*"
THEN *buys_computer* = "*no*"
IF *age* = "<=30" AND *student* = "*yes*"
THEN *buys_computer* = "*yes*"
IF *age* = "31…40"
THEN *buys_computer* = "*yes*"
IF *age* = ">40" AND *credit_rating* = "*excellent*" THEN *buys_computer* = "no"
IF *age* = ">40" AND *credit_rating* = "*fair*"
THEN *buys_computer* = "yes"

52

GiỚI THIỆU

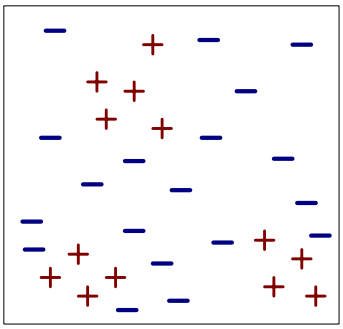
Phương pháp trực tiếp :

- Rút trích luật trực tiếp từ DL : thật toán phủ tuần tự
- *Các luật được học tuần tự. Mỗi luật trong lớp C_i sẽ phủ nhiều mẫu của C_i nhưng không phủ (hoặc phủ ít) mẫu của các lớp khác*
- **Xây dựng luật :**
  - Bắt đầu từ luật rỗng
  - Sử dụng hàm Learn-One-Rule để phát triển luật
    - Thêm thuộc tính làm tăng chất lượng của luật (độ phủ, độ chính xác)
  - Loại các mẫu bị phủ bởi luật ra khỏi DL
  - *Lặp lại quá trình trên cho đến khi gặp điều kiện dừng (không còn mẫu hoặc độ đo chất lượng thấp hơn ngưỡng do người dùng xác định)*

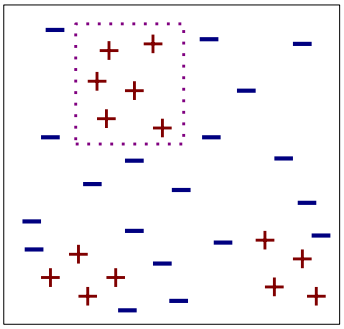
53

GiỚI THIỆU

Ví dụ : Phủ tuần tự



(i) Original Data

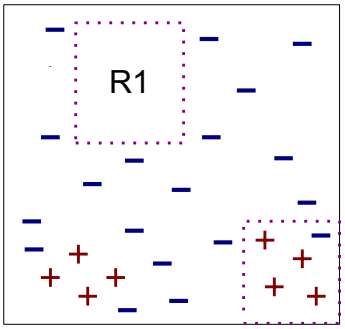


(ii) Step 1

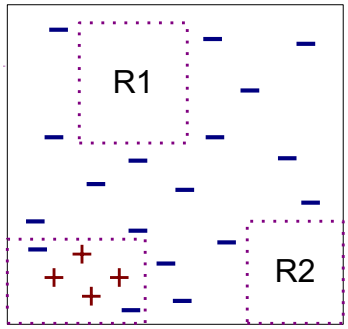
54

GiỚI THIỆU

Ví dụ : Phủ tuần tự



(iii) Step 2



(iv) Step 3

55

GiỚI THIỆU

Ưu điểm của phương pháp dựa trên luật :

- **Dễ hiểu như cây quyết định**
- **Dễ xây dựng luật**
- **Phân lớp mẫu mới nhanh**
- **Thời gian thi hành tương tự như phương pháp dựa trên cây quyết định**

56

LUẬT QUI NẠP - ILA

- M.Tolun, 1998, ILA – inductive learning algorithm
- *Xác định các luật IF-THEN trực tiếp từ tập huấn luyện (phát triển luật theo hướng từ tổng quát -> cụ thể)*
- Chia tập huấn luyện thành các bảng con theo từng giá trị của lớp.
- *Thực hiện việc so sánh các giá trị của thuộc tính trong từng bảng con và tính số lần xuất hiện.*
- Thuộc tính có dạng phi số, giá trị rời rạc

57

THUẬT TOÁN ILA

- Bước 1 : Chia bảng có chứa m mẫu thành n bảng con (*ứng với n giá trị của thuộc tính lớp)*

  (Bước 2 đến bước 8 sẽ được lặp lại cho mỗi bảng con)
- Bước 2 : Khởi tạo số lượng thuộc tính kết hợp j=1
- *Bước 3 : Xét từng bảng con, tạo danh sách các thuộc tính kết hợp (phần tử danh sách có j thuộc tính)*
- Bước 4 : Với mỗi phần tử trong danh sách trên, đếm số lần xuất hiện các giá trị của thuộc tính ở các dòng chưa đánh dấu của bảng con đang xét, nhưng giá trị không được xuất hiện ở những bảng con khác.

Chọn phần tử kết hợp đầu tiên có số lần xuất hiện của giá trị thuộc tính nhiều nhất và đặt tên là max-combination.

58

THUẬT TOÁN ILA

- **Bước 5 : Nếu max-combination = 0 thì j=j+1 và quay lại bước 3**
- ***Bước 6 : Trong bảng con đang xét, đánh dấu các dòng có xuất hiện giá trị của max-combination***
- **Bước 7 : tạo luật**
  - **IF AND(thuộc tính = giá trị) (*thuộc max-combination*) THEN giá trị của thuộc tính lớp ứng với bảng con đang xét**
- **Bước 8 :**
  - **Nếu tất cả các dòng đều đánh dấu**
    - **Nếu còn bảng con thì chuyển qua bảng con tiếp theo và lập lại từ bước 2**
    - **Ngược lại : chấm dứt thuật toán**
  - **Ngược lại (còn dòng chưa đánh dấu) thì quay lại bước 4**

59

VÍ DỤ 3 : THUẬT TOÁN ILA

No	Size	Color	Shape	Decision
1	Vừa	Xanh dương	Hộp	Yes
2	Nhỏ	đỏ	Nón	No
3	Nhỏ	đỏ	Cầu	Yes
4	Lớn	đỏ	Nón	No
5	Lớn	Xanh lá cây	Trụ	Yes
6	Lớn	đỏ	Trụ	No
7	Lớn	Xanh lá cây	Cầu	Yes

60

VÍ DỤ 3 : *BƯỚC 1*

Bảng con 1				
No	Size	Color	Shape	Decision
1 (1)	Vừa	Xanh dương	Hộp	Yes
2 (3)	Nhỏ	đỏ	Cầu	Yes
3 (5)	Lớn	Xanh lá cây	Trụ	Yes
4 (7)	Lớn	Xanh lá cây	Cầu	Yes
Bảng con 2				
No	Size	Color	Shape	Decision
1 (2)	Nhỏ	đỏ	Nón	No
2 (4)	Lớn	đỏ	Nón	No
3 (6)	Lớn	đỏ	Trụ	No

61

VÍ DỤ 3 : THUẬT TOÁN ILA

Bảng con 1				
No	Size	Color	Shape	Decision
1 (1)	Vừa	Xanh dương	Hộp	Yes
2 (3)	Nhỏ	đỏ	Cầu	Yes
3 (5)	Lớn	Xanh lá cây	Trụ	Yes
4 (7)	Lớn	Xanh lá cây	Cầu	Yes

- B2 : j = 1
- B3 : {[size], [color], [shape]}
- B4 : max-combination = “xanh lá cây”
- B6 : đánh dấu dòng 3,4
- B7 : R1 : IF color = “xanh lá cây” THEN decision =“Yes”
- B8 : Quay lại B4

62

VÍ DỤ 3 : THUẬT TOÁN ILA

Bảng con 1				
No	Size	Color	Shape	Decision
1 (1)	**Vừa**	**Xanh dương**	**Hộp**	**Yes**
2 (3)	**Nhỏ**	**đỏ**	**Cầu**	**Yes**
3 (5)	Lớn	Xanh lá cây	Trụ	Yes
4 (7)	Lớn	Xanh lá cây	Cầu	Yes

- **B4 : max-combination = "vừa"**
- **B6 : đánh dấu dòng 1**
- **B7 : R2 : IF size = "vừa" THEN decision ="Yes"**
- **B8 : quay lại B4**

63

VÍ DỤ 3 : THUẬT TOÁN ILA

Bảng con 1				
No	Size	Color	Shape	Decision
1 (1)	Vừa	Xanh dương	Hộp	Yes
2 (3)	**Nhỏ**	**đỏ**	**Cầu**	**Yes**
3 (5)	Lớn	Xanh lá cây	Trụ	Yes
4 (7)	Lớn	Xanh lá cây	Cầu	Yes

- **B4 : max-combination = "cầu"**
- **B6 : đánh dấu dòng 2**
- **B7 : R3 : IF shape = "cầu" THEN decision ="Yes"**
- **B8 : chuyển qua bảng con 2 vàs bắt dầu từ B2**

64

VÍ DỤ 3 : THUẬT TOÁN ILA

Bảng con 2				
No	Size	Color	Shape	Decision
1 (2)	Nhỏ	đỏ	Nón	No
2 (4)	Lớn	đỏ	Nón	No
3 (6)	Lớn	đỏ	Trụ	No

65

VÍ DỤ 3 : THUẬT TOÁN ILA

Bảng con 2				
No	Size	Color	Shape	Decision
1 (2)	Nhỏ	đỏ	Nón	No
2 (4)	Lớn	đỏ	Nón	No
3 (6)	Lớn	đỏ	Trụ	No

- B2 : j = 1
- B3 : {[size], [color], [shape]}
- B4 : max-combination = "Nón"
- B6 : đánh dấu dòng 1,2
- B7 : R4 : IF shape = "Nón" THEN decision ="No"
- B8 : Quay lại B4

66

VÍ DỤ 3 : THUẬT TOÁN ILA

Bảng con 2				
No	Size	Color	Shape	Decision
1 (2)	Nhỏ	đỏ	Nón	No
2 (4)	Lớn	đỏ	Nón	No
3 (6)	**Lớn**	**đỏ**	**Trụ**	**No**

- **B4 : max-combination = {}**
- **B5 : j = j +1 =2 và quay lại B3**
- **B3 : {[size, color], [size, shape], [color, shape]}**
- **B4 : max-combination = "Lớn đỏ"**
- **B6 : đánh dấu dòng 3**
- **B7 : R5 : IF size = "Lớn" AND color = 'đỏ" THEN decision ="No"**
- **B8 : kết thúc**

67

VÍ DỤ 3 : TẬP LUẬT

- **R1 : IF color = "xanh lá cây"**
  THEN decision ="Yes"
- **R2 : IF size = "vừa" THEN decision ="Yes"**
- **R3 : IF shape = "cầu"**
  THEN decision ="Yes"
- **R4 : IF shape = "Nón" THEN decision ="No"**
- **R5 : IF size = "Lớn" AND color = 'đỏ"**
  THEN decision ="No"

68

CÁC CÔNG VIỆC CẦN LÀM

1. Thực hiện bài tập nhóm chương 4 – Phần 1.
   - **Nộp bài qua Moodle trước 23h00 ngày chủ nhật – 25/10/2009**
2. Chuẩn bị bài 4 : Phân lớp dữ liệu
   - Xem nội dung bài 4 – Phần 2.
   - Chuẩn bị BT.
   - **Cách thực hiện :**
     - **Đọc slide, xem các ví dụ**
     - **Tham khảo trên Internet và tài liệu tham khảo**

69

BÀI TẬP PHẦN 1

1. Hãy cho biết chi tiết các bước của phương pháp phân lớp dựa trên cây quyết định, dựa trên luật (ILA).
2. Cho cây quyết định, bạn có 2 lựa chọn :
   a) Biến đổi cây thành luật, sau đó loại bớt luật kết quả
   b) Loại bớt nhánh của cây, sau đó biến đổi cây thành luật.

   Hãy cho biết các ưu thế của a) so với b) ?
3. Cho tập huấn luyện như trong ví dụ 1 (“mua’, “không mua máy tính”)
   a) Sử dụng chi mục gini để xây dựng cây quyết định. So sánh kết quả với cây sử dụng độ lợi thông tin.
   b) Áp dụng thuật toán ILA cho ví dụ 1, so sánh kết quả với tập luật rút ra từ phương pháp cây quyết định
   c) Xác định lớp cho mẫu ***X* = < age =<=30, income = medium, student =yes, credit_rating =fair >** dựa trên tập luật thu được.

70

BÀI TẬP PHẦN 1

4. Cho tập huấn luyện như trong ví dụ 2.

a) Áp dụng phương pháp cây quyết định lên ví dụ 2 ***(không sử dụng cột thuộc tính name để phân chia DL***) và rút tập luật từ cây. So sánh với tập luật đã có, nhận xét.

b) Sử dụng tập luật thu được từ cây quyết định để xác định lớp cho các mẫu mới sau. So sánh kết quả với việc sử dụng tập luật đã có trong bài giảng.

Name	Blood Type	Give Birth	Can Fly	Live in Water	Class
hawk	warm	no	yes	no	?
grizzly bear	warm	yes	no	no	?

Name	Blood Type	Give Birth	Can Fly	Live in Water	Class
lemur	warm	yes	no	no	?
turtle	cold	no	no	sometimes	?
dogfish shark	cold	yes	no	yes	?

71

BÀI TẬP PHẦN 1

5. Cho tập huấn luyện sau. Giả sử "Chơi Tennis" là thuộc tính lớp.

a) Sử dụng lần lượt độ đo Gain, chỉ mục gini để xây dựng cây quyết định. Biến đổi cây thành luật. So sánh.

b) Sử dụng phương pháp ILA để xác định luật. So sánh với các luật tthu được ở câu a). Nhận xét ?

c) Sử dụng lần lượt các tập luật thu được từ câu a) , b) để xác định lớp cho mẫu mới. So sánh kết quả.

Quang cảnh	Nhiệt độ	Độ ẩm	Sức gió	Chơi Tennis
Mưa	TB	BT	Mạnh	?
Nắng	TB	Cao	Mạnh	?

72

BÀI TẬP PHẦN 1

Quang cảnh	Nhiệt độ	Độ ẩm	Sức gió	Chơi tennis
Nắng	Nóng	Cao	Yếu	Không
Nắng	Nóng	Cao	Mạnh	Không
Mây	Nóng	Cao	Yếu	Có
Mưa	TB	Cao	Yếu	Có
Mưa	Lạnh	BT	Yếu	Có
Mưa	Lạnh	BT	Mạnh	Không
Mây	Lạnh	BT	Mạnh	Có
Nắng	TB	Cao	Yếu	Không
Nắng	Lạnh	BT	Yếu	Có
Mưa	TB	BT	Yếu	Có
Nắng	TB	BT	Mạnh	Có
Mây	TB	Cao	Mạnh	Có
Mây	Nóng	BT	Yếu	Có
Mưa	TB	Cao	Mạnh	Không

73

TÀI LIỆU THAM KHẢO

1. C. Apte and S. Weiss. ***Data mining with decision trees and decision rules. Future Generation Computer Systems***, 13, 1997.
2. M. Kamber, L. Winstone, W. Gong, S. Cheng, and J. Han. ***Generalization and decision tree induction: Efficient classification in data mining***. In Proc. 1997 Int. Workshop Research Issues on Data Engineering (RIDE'97), pages 111-120, Birmingham, England, April 1997.
3. Mehmet R. Tolun, Saleh M. Abu-Soud. ***ILA, an inductive learning algorithm for rule extraction***. ESA 14(3), 4/1998, 361-370

74

Q & A