Chương 4

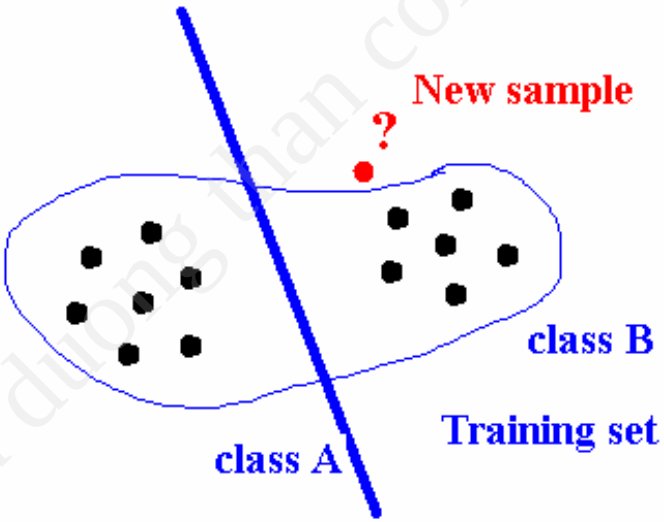
Khai phá dữ liệu

Nội dung

1. Tiền xử lý dữ liệu.
2. Phương pháp khai phá bằng luật kết hợp.
3. Phương pháp cây quyết định.
4. Các phương pháp phân cụm.
5. Các phương pháp khai phá dữ liệu phức tạp.

Phân lớp dữ liệu

- Phân lớp dữ liệu (Classification):

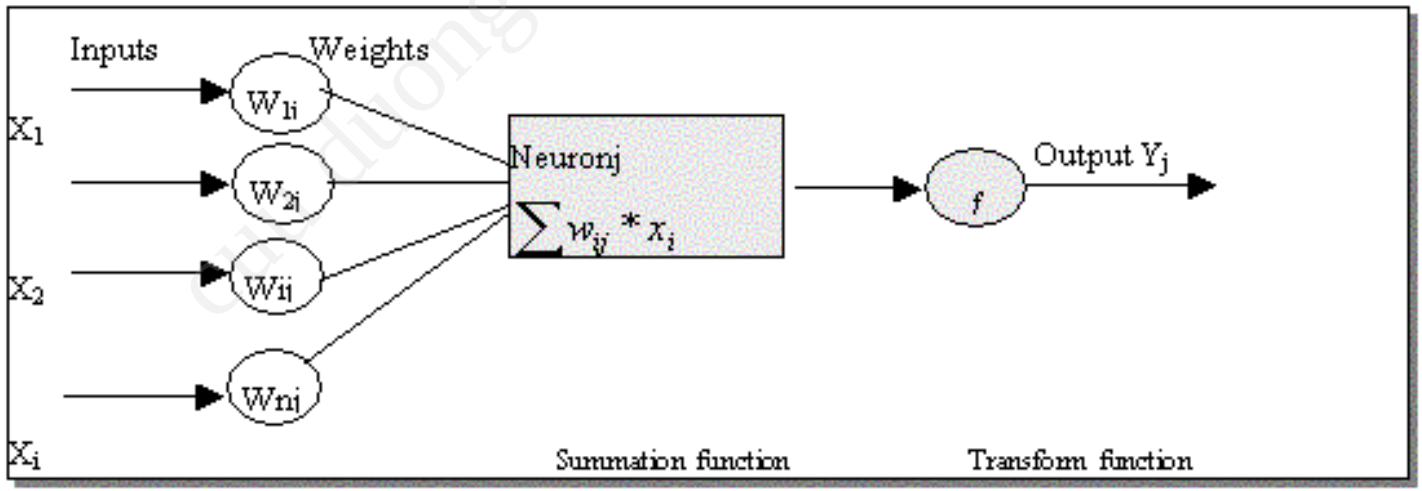


Phân lớp dữ liệu

- Phân lớp dữ liệu (Classification) là quá trình phân chia các đối tượng dữ liệu cào các lớp cho trước.
- Gồm hai bước:
  - Bước học: giai đoạn huấn luyện (training). Giai đoạn này thường áp dụng các giải thuật học có giám sát (supervised learning)
  - Bước phân loại: Phân dữ liệu mới vào các lớp đã biết.

Phân lớp dữ liệu

- Một số giải thuật dùng trong phân loại dữ liệu:
  - Mạng neural (Neural Network),



Phân lớp dữ liệu

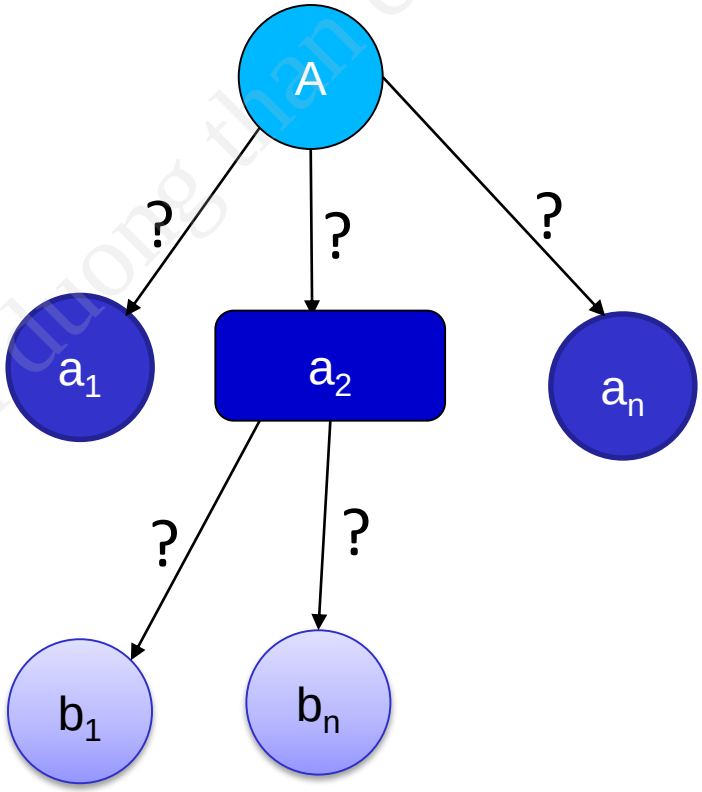
- Một số giải thuật dùng trong phân loại dữ liệu:
  - Mạng Bayesian (dạng đơn giản là Naïve Bayes).

$$P(H \mid X) = \frac{P(X \mid H)P(H)}{P(X)}$$

  - ✓ Trong đó: P(H), P(X|H), P(H) có thể được tính từ tập dữ liệu cho trước,
  - ✓ P(H|X) được tính từ định lý Bayes.

Phân lớp dữ liệu

- Một số giải thuật dùng trong phân loại dữ liệu:
  - Cây quyết định (decision tree),



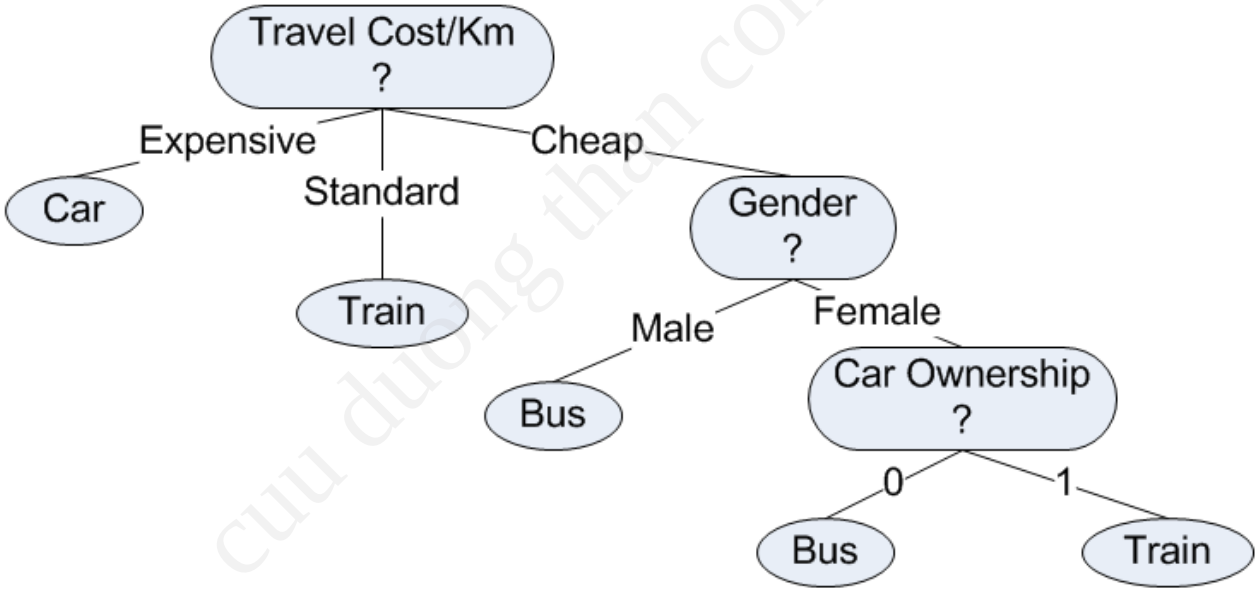
Cây quyết định

- Cây quyết định (decision tree)
  - ✓ Là một mô hình phân lớp điển hình.
  - ✓ Node trong: Kiểm thử một thuộc tính,
  - ✓ Node lá: Mô tả một lớp
  - ✓ Nhánh (từ một node trong): Kết quả của một phép thử trên thuộc tính tương ứng.
  - ✓ Có thể chuyển mô hình cây quyết định sang mô hình luật phân lớp: Đi từ node gốc tới node lá, mỗi đường đi tương ứng với một luật phân lớp.

Cây quyết định

Attributes				Classes
Gender	Car ownership	Travel Cost ($)/km	Income Level	Transportation mode
Male	0	Cheap	Low	Bus
Male	1	Cheap	Medium	Bus
Female	0	Cheap	Low	Bus
Male	1	Cheap	Medium	Bus
Female	1	Expensive	High	Car
Male	2	Expensive	Medium	Car
Female	2	Expensive	High	Car
Female	1	Cheap	Medium	Train
Male	0	Standard	Medium	Train
Female	1	Standard	Medium	Train

Cây quyết định



Cây quyết định

- Các độ đo dùng trong phân lớp bằng cây quyết định:
  - Entropy: Entropy dùng trong thông tin là một khái niệm mở rộng của entropy trong Nhiệt động lực học và Cơ học thống kê. Entropy mô tả mức độ hỗn loạn trong một tín hiệu lấy từ một sự kiện ngẫu nhiên.

$$Entropy = \sum_{j} -p_j log_2 p_j$$

Trong đó: p_i là xác suất xuất hiện một thông tin trong tập dữ liệu.

Cây quyết định

- Các độ đo dùng trong phân lớp bằng cây quyết định:
  - Gini Index: Độ đo về độ không tinh khiết của thông tin.

$$Gini\ Index = 1 - \sum_j p_j^2$$

Cây quyết định

- Các độ đo dùng trong phân lớp bằng cây quyết định:
  - Information Gain (Độ lợi thông tin): Là độ sai biệt giữa trị thông tin trước phân hoạch (Info(D)) và trị thông tin sau phân hoạch với A ($Info_A(D)$).

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

Cây quyết định

- Entropy:

Attributes				Classes
Gender	Car ownership	Travel Cost ($)/km	Income Level	Transportation mode
Male	0	Cheap	Low	Bus
Male	1	Cheap	Medium	Bus
Female	0	Cheap	Low	Bus
Male	1	Cheap	Medium	Bus
Female	1	Expensive	High	Car
Male	2	Expensive	Medium	Car
Female	2	Expensive	High	Car
Female	1	Cheap	Medium	Train
Male	0	Standard	Medium	Train
Female	1	Standard	Medium	Train

- Pro(Bus) = 4/10
- Pro(Car) = 3/10
- Pro(Train) = 3/10
- Entropy = $-0.4\log_2(0.4) - 0.3\log_2(0.3) - 0.3\log_2(0.3)$

  $= 1.571$
- Gini Index = $1 - (0.4^2 + 0.3^2 + 0.3^2) = 0.66$

Cây quyết định

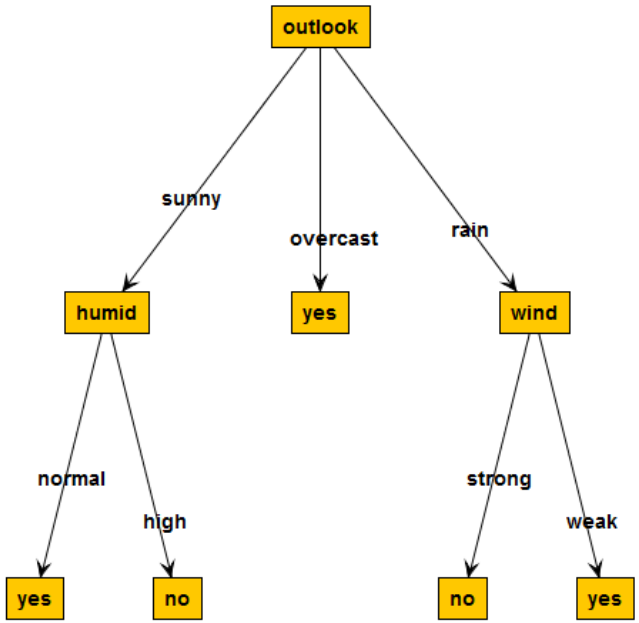
Outlook	Temperature	Humidity	Wind	Play ball
Sunny	Hot	High	Weak (false)	No
Sunny	Hot	High	Strong (true)	No
Overcast	Hot	High	Weak	Yes
Rain	Mild	High	Weak	Yes
Rain	Cool	Normal	Weak	Yes
Rain	Cool	Normal	Strong	No
Overcast	Cool	Normal	Strong	Yes
Sunny	Mild	High	Weak	No
Sunny	Cool	Normal	Weak	Yes
Rain	Mild	Normal	Weak	Yes
Sunny	Mild	Normal	Strong	Yes
Overcast	Mild	High	Strong	Yes
Overcast	Hot	Normal	Weak	Yes
Rain	Mild	High	Strong	No
			Total	**14**

Cây quyết định

- Gain information:
  - **Entropy(S)** = $-(9/14)\log_2(9/14) -(5/14)\log_2(5/14)$ = **0.940**
  - **Gain(S, Windy) =** Entropy(S) - (8/14)Entropy(S_{false}) - (6/14)Entropy(S_{true}) = **0.048**
  - Windy: Weak=8→(6+,2-), Strong=6→(3+,3-)
  - Entropy(S_{false})=$-6/8\text{Log}_2(6/8)-2/8\text{Log}_2(2/8)$=0.811
  - Entropy(S_{true}) =$-3/6\text{Log}_2(3/6)-3/6\text{Log}_2(3/6)$=1
  - **Gain(S,Windy) =** 0.940-(8/14)(0.811)-(6/14)(1)=**0.048**

Cây quyết định

- Gain information:
  - Tính tương tự ta được:
    - ✓ Gain(S, Windy) = 0.048
    - ✓ Gain(S, Humidity) = 0.151
    - ✓ Gain(S, Temperature) = 0.029
    - ✓ Gian(S, Outlook) = **0.246**



Cây quyết định

Method:

```
(1)  create a node N;
(2)  if tuples in D are all of the same class, C then
(3)      return N as a leaf node labeled with the class C;
(4)  if attribute_list is empty then
(5)      return N as a leaf node labeled with the majority class in D; // majority voting
(6)  apply Attribute_selection_method(D, attribute_list) to find the "best" splitting_criterion;
(7)  label node N with splitting_criterion;
(8)  if splitting_attribute is discrete-valued and
         multiway splits allowed then // not restricted to binary trees
(9)      attribute_list ← attribute_list − splitting_attribute; // remove splitting_attribute
(10) for each outcome j of splitting_criterion
     // partition the tuples and grow subtrees for each partition
(11)     let D_j be the set of data tuples in D satisfying outcome j; // a partition
(12)     if D_j is empty then
(13)         attach a leaf labeled with the majority class in D to node N;
(14)     else attach the node returned by Generate_decision_tree(D_j, attribute_list) to node N;
     endfor
(15) return N;
```

Cây quyết định

```
procedure Build_tree(Records, Attributes)
begin
    Tạo nút N;
    if (tất cả các bản ghi thuộc về một lớp Ci nào đó) then
    begin
        N.Label = Ci;
        return N;
    end;
    if (Attributes = ∅) then
    begin
        Tìm lớp Cj mà phần lớn các bản ghi r ∈ Records thuộc về lớp đó.
        N.Label = Cj;
        return N;
    end;
    Chọn Ai ∈ Attribute sao cho Gain(Ai)→max;
    N.Label = Ai;
    for each giá trị vi đã biết của Ai do
    begin
        Thêm một nhánh mới vào nút N ứng với Ai = vj ;
        Sj = Tập con của Records có Ai = vj;
        if (Sj = ∅) then
                Thêm một nút lá L với nhãn là lớp mà phần lớn các bản ghi r ∈ Records thuộc về lớp đó;
                Return L;
        else
                Thêm vào nút được trả về bởi Build_Tree(Sj, Attribute \{Ai});
    end ;
end;
```

Continuous attributes

Day	Outlook	Temperature	Humidity	Wind	Play ball
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

Thuộc tính giá trị liên tục

$Entropy_{Day}(S) = (1/14)Entropy(S_{D1}) + (1/14)Entropy(S_{D2}) + \dots + (1/14)Entropy(S_{D14})$

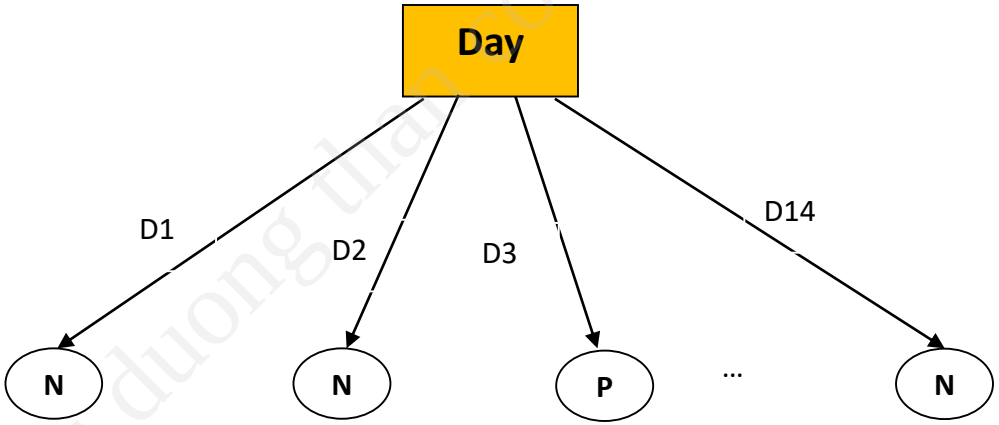
$Entropy(S_{D1}) = Entropy(S_{D2}) = \dots = Entropy(S_{D14}) = 0$

$\rightarrow Entropy_{Day}(S) = 0$

$Entropy(S) = -(9/14)\log_2(9/14) -(5/14)\log_2(5/14) = \mathbf{0.940}$

$Gain(S, Day) = Entropy(S) - Entropy_{Day}(S) = 0.940$

Thuộc tính giá trị liên tục



Thuộc tính giá trị liên tục

- Vấn đề:
  - ✓ Thuộc tính ngày có độ thu thập thông tin cao → có độ ưu tiên trong lựa chọn quyết định.
  - ✓ Nếu ý nghĩa của thuộc tính Day không cao thì sự lựa chọn quyết định này là không hiệu quả → tính dự đoán kém.
- Giải quyết vấn đề: nguyên tắc lựa chọn phân tách:
  - ✓ Tỷ lệ tăng thêm thông tin (GainRatio) cao,
  - ✓ Có Entropy của thuộc tính lớn hơn Entropy trung bình của tất cả các thuộc tính

Thuộc tính giá trị liên tục

Outlook	Temperature	Humidity	Wind	Play ball
Sunny	Hot	0.9	Weak	No
Sunny	Hot	0.87	Strong	No
Overcast	Hot	0.93	Weak	Yes
Rain	Mild	0.89	Weak	Yes
Rain	Cool	0.80	Weak	Yes
Rain	Cool	0.59	Strong	No
Overcast	Cool	0.77	Strong	Yes
Sunny	Mild	0.91	Weak	No
Sunny	Cool	0.68	Weak	Yes
Rain	Mild	0.84	Weak	Yes
Sunny	Mild	0.72	Strong	Yes
Overcast	Mild	0.49	Strong	Yes
Overcast	Hot	0.74	Weak	Yes
Rain	Mild	0.86	Strong	No
			Total	**14**

Thuộc tính giá trị liên tục

- SplitInfomation: Thông tin tiềm ẩn được tạo ra bằng cách chia tập dữ liệu trong một số tập con nào đó.

$$SplitInformation(S,A) = \sum_{i=1}^{c} \frac{|S_i|}{|S|} \log_2 \frac{|S_i|}{|S|}$$

  - ✓ S_i là tập con của S chứa các thể hiện của thuộc tính A mang giá trị V_i.
  - ✓ Splitinfomation thực sự chính là Entropy của S với sự liên quan trên những giá trị của thuộc tính A

Thuộc tính giá trị liên tục

- GainRatio*:* Đánh giá sự thay đổi các giá trị của thuộc tính.

$$GainRatio(S,A) = \frac{Gain(S,A)}{SplitInformation(S,A)}$$

- Tất cả các thuộc tính sẽ được tính toán độ đo tỷ lệ Gain, thuộc tính nào có độ đo tỷ lệ Gain lớn nhất sẽ được chọn làm thuộc tính phân chia

Thuộc tính giá trị liên tục

- Các bước tính:
  1. Tính Entropy,
  2. Tính Gain,
  3. Tính SplitInformation,
  4. Tính GainRatio,
  5. Tính Entropy trung bình,
  6. So sánh các Entropy với Entropy trung bình + so sánh GainRation để chọn thuộc tính phân tách.