Câu 1: (4 điểm)

a) So sánh kết quả các phương pháp tìm tập phổ biến về: kết quả, không gian lưu trữ và thời gian khai thác.

b) Cho cơ sở dữ liệu giao tác như sau:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Mã giao dịch | A | B | C | D | E | F |
| 1 | 0 | 0 | 1 | 0 | 1 | 1 |
| 2 | 1 | 0 | 1 | 1 | 1 | 0 |
| 3 | 0 | 0 | 1 | 1 | 0 | 0 |
| 4 | 0 | 0 | 0 | 1 | 1 | 0 |
| 5 | 1 | 1 | 1 | 0 | 1 | 0 |

1. Tìm tất cả các tập phổ biến (FIs) có trong CSDL với minSup = 40% theo phương pháp FP-Tree

**Giải:**

minSup = 40% = 40 \* 5 / 100 = 2

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Danh Mục | A | B | C | D | E | F |
| σ | 2 | 1 | 4 | 3 | 4 | 1 |

Sắp xếp theo σ

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Danh Mục | C | E | D | A | B | F |
| σ | 4 | 4 | 3 | 2 | 1 | 1 |

Xây dựng cây như sau:

C:4

E:1

D:1

D:1

E:3

D:1

A:1

A:1

{}

Chiếu trên A:2, ta có

* CSDL cục bộ: {CED:1, CE:1}
* Cây cục bộ:

{}

C:2

E:2

* Các tập con là: {∅, E:2, C:2, EC:2}
  + Các tập phổ biến là: {A:2, AE:2, AC:2, AEC:2}

Chiếu trên D:3, ta có

* CSDL cục bộ: {CE:1, C:1, E:1}
* Cây cục bộ

{}

C:2

* Các tập con là: {∅, C:2}
  + Các tập phổ biến là: {D:3, DC:2}

Chiếu trên E:4, ta có

* CSDL cục bộ: {C:3}
* Cây cục bộ:

{}

C:3

* Các tập con là: {∅, C:3}
  + Các tập phổ biến là: {E:4, EC:3}

Chiếu trên C:4, ta được {0}

* Tập phổ biến là: {C:4}

1. Tìm tất cả các tập phổ biến đóng theo thuận toán CHARM với minSup = 40%

Câu 2: (4 điểm) Cho bảng quan sát về thời tiết như sau:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Ngày | Quang cảnh | Nhiệt độ | Độ ẩm | Gió | Chơi Tennis? |
| 1 | Mưa | Lạnh | TB | Thấp | Đi |
| 2 | Nắng | TB | Cao | Thấp | Không đi |
| 3 | Nắng | Lạnh | TB | Thấp | Đi |
| 4 | Âm u | Lạnh | TB | Cao | Đi |
| 5 | Mưa | TB | TB | Thấp | Đi |
| 6 | Nắng | Nóng | Cao | Cao | Không đi |
| 7 | Nắng | TB | TB | Cao | Đi |
| 8 | Âm u | TB | Cao | Cao | Đi |
| 9 | Âm u | Nóng | TB | Thấp | Đi |
| *10* | *Mưa* | *TB* | *Cao* | *Cao* | *Không đi* |
| 11 | Nắng | Nóng | Cao | Thấp | Không đi |
| 12 | Âm u | Nóng | Cao | Thấp | Đi |
| 13 | Mưa | Lạnh | TB | Cao | Đi |
| 14 | Âm u | TB | Cao | Thấp | Không đi |
| *15* | *Mưa* | *TB* | *Cao* | *Thấp* | *Đi* |

a) Tính Ratio của thuộc tính Quang cảnh và Nhiệt độ dựa trên 10 dòng dữ liệu đầu tiên. Nếu chọn lựa nút làm gốc ta nên lựa thuộc tính nào? Tại sao?

**Giải:**

S=[7+, 3-]

Entropy(S) = -7/10 \* log2(7/10) – 3/10 \* log2(3/10) = 0.881

Ta có:

Values(Quang cảnh) = {Mưa, Nắng, Âm u}

SMua = [2+, 1-] , SNang = [2+, 2-] , SAmu = [3+, 0]

Entropy(SMua) = -2/3 \* log2(2/3) – 1/3 \* log2(1/3) = 0.918

Entropy(SNang) = 1 , Entropy(SAmu) = 0

Gain(S,Quang cảnh) =

Entropy(S) – (3/10 \* Entropy(SMua) + 4/10 \* Entropy(SNang) + 3/10 \* Entropy(SAmu))

= 0.881 – (3/10 \* 0.918 + 4/10 \* 1 + 3/10 \* 0) = 0.2056

SplitInfo(S,Quang cảnh) = -3/10 \* log2(3/10) – 4/10 \* log2(4/10) – 3/10 \* log2(3/10) = 1.571

* RatioGain(S,Quang cảnh) = Gain(S,Quang cảnh) / SplitInfo(S,Quang cảnh)

= 0.2056 / 1.571 = **0.131**

Ta có:

Values(Nhiệt độ) = {Lạnh, TB, Nóng}

SLanh = [3+, 0] , STB = [3+, 2-] , SNong = [1+, 1-]

Entropy(SLanh) = 0 , Entropy(SNong) = 1

Entropy(STB) = -3/5 \* log2(3/5) – 2/5 \* log2(2/5) = 0.971

Gain(S,Nhiệt độ) = 0.881 – (3/10 \* 0 + 5/10 \* 0.971 + 2/10 \* 1) = 0.1955

SplitInfo(S,Nhiệt độ) = -3/10 \* log2(3/10) – 5/10 \* log2(5/10) – 2/10 \* log2(2/10) = 1.485

* RatioGain(S,Nhiệt độ) = Gain(S,Nhiệt độ) / SplitInfo(S,Nhiệt độ)

= 0. 1955/ 1. 485= **0.132**

* Chọn thuộc tính Nhiệt độ làm nút gốc do RatioGain lớn nhất

b) Dự đoán mẫu của các lớp 10-15 dựa trên 10 dòng đầu tiên theo phương pháp Bayesian. Dựa vào kết quả này, cho biết độ chính xác phân lớp?

**Giải:**

**Dự đoán**

Mẫu 10: X={Mưa, TB, Cao, Cao} => Chơi Tennis(X) = ?

P(Chơi Tennis = Đi) = 7/10 , P(Chơi Tennis = Không Đi) = 3/10

P(Mưa | Đi) = 2/7 -> 3/9 , P(Mưa | Không Đi) = 1/3 -> 2/5

P(TB | Đi) = 3/7 -> 4/9 , P(TB | Không Đi) = 2/3 -> 3/5

P(Cao | Đi) = 1/7 -> 2/9 , P(Cao | Không Đi) = 3/3 -> 4/5 *(Thuộc tính : Độ ẩm)*

P(Cao | Đi) = 3/7 -> 4/9 , P(Cao | Không Đi) = 2/3 ->3/5 *(Thuộc tính : Gió)*

P(X | Đi) = 7/10 \* 3/9 \* 4/9 \* 2/9 \* 4/9 = 672/65610 = 0.0102

P(X | Không Đi) = 3/10 \* 2/5 \* 3/5 \* 4/5 \* 3/5 = 216/6250 = 0.0346 > P(X | Đi)

* Chơi Tennis(X) = Không Đi

Mẫu 15: X={Mưa, TB, Cao, Thấp} => Chơi Tennis(X) = ?

P(Chơi Tennis = Đi) = 7/10 , P(Chơi Tennis = Không Đi) = 3/10

P(Mưa | Đi) = 2/7 -> 3/9 , P(Mưa | Không Đi) = 1/3 -> 2/5

P(TB | Đi) = 3/7 -> 4/9 , P(TB | Không Đi) = 2/3 -> 3/5

P(Cao | Đi) = 1/7 -> 2/9 , P(Cao | Không Đi) = 3/3 -> 4/5

P(Thấp | Đi) = 4/7 -> 5/9 , P(Thấp | Không Đi) = 1/3 ->2/5

P(X | Đi) = 7/10 \* 3/9 \* 4/9 \* 2/9 \* 5/9 = 840/65610 = 0.0128

P(X | Không Đi) = 3/10 \* 2/5 \* 3/5 \* 4/5 \* 2/5 = 144/6250 = 0.023 > P(X | Đi)

* Chơi Tennis(X) = Không Đi

**Tính độ chính xác phân lớp**

Accuracy = ½ = 0.5