|  |  |
| --- | --- |
| **TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**  **ĐẠI HỌC QUỐC GIA HỒ CHÍ MINH** | **IE403.E21.CN1.CNTT**  **Khai thác dữ liệu truyền thông xã hội** |

**CHƯƠNG 1: TẬP PHỔ BIẾN**

**Bài 1:** Cho ngữ cảnh khai thác dữ liệu:

|  |  |
| --- | --- |
| **Mã hóa đơn** | **Mã hàng** |
| o1 | i1 |
| o1 | i2 |
| o1 | i3 |
| o2 | i2 |
| o2 | i3 |
| o2 | i4 |
| o3 | i2 |
| o3 | i3 |
| o3 | i4 |
| o4 | i1 |
| o4 | i2 |
| o4 | i3 |
| o5 | i3 |
| o5 | i4 |

Tìm tập phổ biến thỏa ngưỡng minsupp = 0.4.

**ĐÁP ÁN**

**Bước 1: Thành lập ma trận nhị phân**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | i1 | i2 | i3 | i4 |
| o1 | 1 | 1 | 1 | 0 |
| o2 | 0 | 1 | 1 | 1 |
| o3 | 0 | 1 | 1 | 1 |
| o4 | 1 | 1 | 1 | 0 |
| o5 | 0 | 0 | 1 | 1 |

**Bước 2: Tìm tập phổ biến thỏa ngưỡng**

* Các tập ứng cử viên có 1 phần tử:
  + F1 = {{i1},{i2},{i3},{i4}}
    - SP({i1}) = 2/5 = 0.4; Phổ biến
    - SP({i2}) = 4/5 = 0.8; Phổ biến
    - SP({i3}) = 5/5 = 1.0; Phổ biến
    - SP({i4}) = 3/5 = 0.6; Phổ biến
  + Các tập phổ biến có 1 phần tử gồm:

C1={{i1},{i2},{i3},{i4}}

* Các tập ứng cử viên có 2 phần tử từ tập C1:
  + F2 = {{i1,i2},{i1,i3},{i1,i4},{i2,i3},{i2,i4},{i3,i4}}
    - SP({i1,i2}) = 2/5 = 0.4; Phổ biến
    - SP({i1,i3}) = 2/5 = 0.4; Phổ biến
    - SP({i1,i4}) = 0/5 = 0.0; Không phổ biến
    - SP({i2,i3}) = 4/5 = 0.8; Phổ biến
    - SP({i2,i4}) = 2/5 = 0.4; Phổ biến
    - SP({i3,i4}) = 3/5 = 0.6; Phổ biến
  + Các tập phổ biến có 2 phần tử gồm:

C2 = {{i1,i2},{i1,i3},{i2,i3},{i2,i4},{i3,i4}}

* Các tập ứng cử viên có 3 phần tử:
  + F3 = {{i1,i2,i3},{i2,i3,i4}}
    - SP({i1,i2,i3}) = 2/5 = 0.4; Phổ biến
    - SP({i2,i3,i4}) = 2/5 = 0.4; Phổ biến
  + Các tập phổ biến có 3 phần tử gồm:

C3 = {{i1,i2,i3},{i2,i3,i4}}

* Các tập ứng cử viên có 4 phần tử:
  + F3 = {i1,i2,i3,i4}
    - {i1,i2,i3,i4} không phổ biến vì {i1,i4} không phổ biến
* Các tập phổ biến thỏa ngưỡng gồm: {i1},{i2},{i3},{i4},{i1,i2},{i1,i3},{i2,i3},{i2,i4},{i3,i4},{i1,i2,i3},{i2,i3,i4}

**Bước 3: Tìm tập phổ biến tối đại**

* Các tập phổ biến tối đại gồm: {i1,i2,i3},{i2,i3,i4}

**Bước 4: Phát biểu luật kết hợp**

|  |  |
| --- | --- |
| i1 -> i2,i3  i2 -> i1,i3  i3 -> i1,i2  i1,i2 -> i3  i1,i3 -> i2  i2,i3 -> i1 | i2 -> i3,i4  i3 -> i2,i4  i4 -> i2,i3  i2,i3 -> i4  i2,i4 -> i3  i3,i4 -> i2 |

**Bước 5: Tính độ tin cậy của luật kết hợp**

|  |  |
| --- | --- |
| R1: i1 -> i2,i3  CF(R1) = SP(i1,i2,i3) / SP(i1) = 0.4/0.4 = 1.0 ⬄ 100% | R7: i1 -> i2,i3  CF(R7) = SP(i1,i2,i3) / SP(i1) = 0.4/0.4 = 1.0 ⬄ 100% |
| R2: i2 -> i1,i3  CF(R2) = SP(i1,i2,i3) / SP(i2) = 0.4/0.8 = 0.5 ⬄ 50% | R8: i2 -> i1,i3  CF(R8) = SP(i1,i2,i3) / SP(i2) = 0.4/0.8 = 0.5 ⬄ 50% |
| R3: i3 -> i2,i3  CF(R3) = SP(i1,i2,i3) / SP(i3) = 0.4/1.0 = 0.4 ⬄ 40% | R9: i3 -> i2,i3  CF(R9) = SP(i1,i2,i3) / SP(i3) = 0.4/1.0 = 0.4 ⬄ 40% |
| R4: i1,i2 -> i3  CF(R4) = SP(i1,i2,i3) / SP(i1,i2) = 0.4/0.4 = 1.0 ⬄ 100% | R10: i1,i2 -> i3  CF(R10) = SP(i1,i2,i3) / SP(i1,i2) = 0.4/0.4 = 1.0 ⬄ 100% |
| R5: i1,i3 -> i2  CF(R5) = SP(i1,i2,i3) / SP(i1,i3) = 0.4/0.4 = 1.0 ⬄ 100% | R11: i1,i3 -> i2  CF(R11) = SP(i1,i2,i3) / SP(i1,i3) = 0.4/0.4 = 1.0 ⬄ 100% |
| R6: i2,i3 -> i1  CF(R6) = SP(i1,i2,i3) / SP(i2,i3) = 0.4/0.8 = 0.5 ⬄ 50% | R12: i2,i3 -> i1  CF(R12) = SP(i1,i2,i3) / SP(i2,i3) = 0.4/0.8 = 0.5 ⬄ 50% |

**Bài 2:** Cho bối cảnh gồm các giao tác:

o1 = {d1,d3,d4}; o2 = {d1,d3,d4}; o3 = {d3,d5}; o4 = {d4,d5}; o5 = {d2,d3,d5}

* Tìm các tập phổ biến tối đại minsupp = 0.3.
* Liệt kê 3 luật từ tập phổ biến tối đại và tính độ tin cậy của 3 luật này.

**ĐÁP ÁN**

**Bước 1: Thành lập ma trận nhị phân**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | d1 | d2 | d3 | d4 | d5 |
| o1 | 1 | 0 | 1 | 1 | 0 |
| o2 | 1 | 0 | 1 | 1 | 0 |
| o3 | 0 | 0 | 1 | 0 | 1 |
| o4 | 0 | 0 | 0 | 1 | 1 |
| o5 | 0 | 1 | 1 | 0 | 1 |

**Bước 2: Tìm tập phổ biến thỏa ngưỡng**

* Các tập ứng cử viên có 1 phần tử:
  + F1 = {{d1},{d2},{d3},{d4},{d5}}
    - SP({d1}) = 2/5 = 0.4; Phổ biến
    - SP({d2}) = 1/5 = 0.2; Không phổ biến
    - SP({d3}) = 4/5 = 0.8; Phổ biến
    - SP({d4}) = 3/5 = 0.6; Phổ biến
    - SP({d5}) = 3/5 = 0.6; Phổ biến
  + Các tập phổ biến có 1 phần tử gồm:

C1={{d1},{d3},{d4},{d5}}

* Các tập ứng cử viên có 2 phần tử từ tập C1:
  + F2 = {{d1,d3},{d1,d4},{d1,d5},{d3,d4},{d3,d5},{d4,d5}}
    - SP({d1,d3}) = 2/5 = 0.4; Phổ biến
    - SP({d1,d4}) = 2/5 = 0.4; Phổ biến
    - SP({d1,d5}) = 0/5 = 0.0; Không phổ biến
    - SP({d3,d4}) = 2/5 = 0.4; Phổ biến
    - SP({d3,d5}) = 2/5 = 0.4; Phổ biến
    - SP({d4,d5}) = 1/5 = 0.2; Không phổ biến
  + Các tập phổ biến có 2 phần tử gồm:

C2 = {{d1,d3},{d1,d4},{d3,d4},{d3,d5}}

* Các tập ứng cử viên có 3 phần tử:
  + F3 = {d1,d3,d4}
    - SP({d1,d3,d4}) = 2/5 = 0.4; Phổ biến
  + Các tập phổ biến có 3 phần tử gồm:

C3 = {d1,d3,d4}

* Các tập phổ biến thỏa ngưỡng gồm: {d1},{d3},{d4},{d5},{d1,d3},{d1,d4},{d3,d4},{d3,d5},{d1,d3,d4}

**Bước 3: Tìm tập phổ biến tối đại**

* Các tập phổ biến tối đại gồm: {d3,d5},{d1,d3,d4}

**Bước 4: Phát biểu luật kết hợp và tính độ tin cậy của luật kết hợp**

|  |  |  |
| --- | --- | --- |
| Luật | Độ tin cậy luật | Thoả ngưỡng |
| R1: d1 -> d3,d4 | CF(R1) = SP(d1,d3,d4) / SP(d1) = 0.4/0.4 = 1.0 ⬄ 100% | Có |
| R2: d3 -> d1,d4 | CF(R2) = SP(d1,d3,d4) / SP(d3) = 0.4/0.8 = 0.5 ⬄ 50% | Có |
| R3: d4 -> d1,d3 | CF(R3) = SP(d1,d3,d4) / SP(d4) = 0.4/0.6 = 0.7 ⬄ 70% | Có |
| R4: d1,d3 -> d4 | CF(R4) = SP(d1,d3,d4) / SP(d1,d3) = 0.4/0.4 = 1.0 ⬄ 100% | Có |
| R5: d1,d4 -> d3 | CF(R5) = SP(d1,d3,d4) / SP(d1,d4) = 0.4/0.4 = 1.0 ⬄ 100% | Có |
| R6: d3,d4 -> d1 | CF(R6) = SP(d1,d3,d4) / SP(d3,d4) = 0.4/0.4 = 1.0 ⬄ 100% | Có |

**Bài 3:** Cho tập hoá đơn O{o1,o2,o3,o4,o5,o6}, tập các mặt hàng I{i1,i2,i3,i4,i5,i6} và tập các giao tác mua hàng sau:

o1(i1,i2,i3,i4)

o2(i1,i2,i3)

o3(i2,i3,i4,i5,i6)

o4(i1,i2,i4,i5,i6)

o5(i2,i4,i6)

o6(i1,i2,i4,i5,i6)

1. Tìm tất cả các tập phổ biến có ngưỡng minsupp = 0.5.
2. Tìm tất cả các tập phổ biến tối đại có ngưỡng minsupp = 0.5.
3. Liệt kê 3 luật từ tập phổ biến tối đại và tính độ tin cậy của 3 luật này.

**ĐÁP ÁN**

**Bước 1: Thành lập ma trận nhị phân**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | i1 | i2 | i3 | i4 | i5 | i6 |
| o1 | 1 | 1 | 1 | 1 | 0 | 0 |
| o2 | 1 | 1 | 1 | 0 | 0 | 0 |
| o3 | 0 | 1 | 1 | 1 | 1 | 1 |
| o4 | 1 | 1 | 0 | 1 | 1 | 1 |
| o5 | 0 | 1 | 0 | 1 | 0 | 1 |
| o6 | 1 | 1 | 0 | 1 | 1 | 1 |

**Bước 2: Tìm tập phổ biến thỏa ngưỡng**

* Các tập ứng cử viên có 1 phần tử:
  + F1 = {{i1},{i2},{i3},{i4},{i5},{i6}}
    - SP({i1}) = 4/6 = 0.7; Phổ biến
    - SP({i2}) = 6/6 = 1.0; Phổ biến
    - SP({i3}) = 3/6 = 0.5; Phổ biến
    - SP({i4}) = 5/6 = 0.8; Phổ biến
    - SP({i5}) = 3/6 = 0.5; Phổ biến
    - SP({i6}) = 4/6 = 0.7; Phổ biến
  + Các tập phổ biến có 1 phần tử gồm:

C1={{i1},{i2},{i3},{i4},{i5},{i6}}

* Các tập ứng cử viên có 2 phần tử từ tập C1:
  + F2 = {{i1,i2},{i1,i3},{i1,i4},{i1,i5},{i1,i6},{i2,i3},{i2,i4},{i2,i5},{i2,i6},{i3,i4},{i3,i5},{i3,i6},{i4,i5},{i4,i6},{i5,i6}}
    - SP({i1,i2}) = 4/6 = 0.7; Phổ biến
    - SP({i1,i3}) = 2/6 = 0.3; Không phổ biến
    - SP({i1,i4}) = 3/6 = 0.5; Phổ biến
    - SP({i1,i5}) = 2/6 = 0.3; Không phổ biến
    - SP({i1,i6}) = 2/6 = 0.3; Không phổ biến
    - SP({i2,i3}) = 3/6 = 0.5; Phổ biến
    - SP({i2,i4}) = 5/6 = 0.8; Phổ biến
    - SP({i2,i5}) = 3/6 = 0.5; Phổ biến
    - SP({i2,i6}) = 4/6 = 0.7; Phổ biến
    - SP({i3,i4}) = 2/6 = 0.3; Không phổ biến
    - SP({i3,i5}) = 1/6 = 0.2; Không phổ biến
    - SP({i3,i6}) = 1/6 = 0.2; Không phổ biến
    - SP({i4,i5}) = 3/6 = 0.5; Phổ biến
    - SP({i4,i6}) = 4/6 = 0.7; Phổ biến
    - SP({i5,i6}) = 3/6 = 0.5; Phổ biến
  + Các tập phổ biến có 2 phần tử gồm:

C2 = {{i1,i2},{i1,i4},{i2,i3},{i2,i4},{i2,i5},{i2,i6},{i4,i5},{i4,i6},{i5,i6}}

* Các tập ứng cử viên có 3 phần tử:
  + F3 = {{i1,i2,i4},{i2,i4,i5},{i2,i4,i6},{i2,i5,i6},{i4,i5,i6}}
    - SP({i1,i2,i4}) = 3/6 = 0.5; Phổ biến
    - SP({i2,i4,i5}) = 3/6 = 0.5; Phổ biến
    - SP({i2,i4,i6}) = 4/6 = 0.7; Phổ biến
    - SP({i2,i5,i6}) = 3/6 = 0.5; Phổ biến
    - SP({i4,i5,i6}) = 3/6 = 0.5; Phổ biến
  + Các tập phổ biến có 3 phần tử gồm:

C3 = {{i1,i2,i4},{i2,i4,i5},{i2,i4,i6},{i2,i5,i6},{i4,i5,i6}}

* Các tập ứng cử viên có 4 phần tử:
  + F4 = {{i1,i2,i4,i5},{i1,i2,i4,i6},{i2,i4,i5,i6}}
    - SP({i1,i2,i4,i5}) = 2/6 = 0.3; Không phổ biến
    - SP({i1,i2,i4,i6}) = 2/6 = 0.3; Không phổ biến
    - SP({i2,i4,i5,i6}) = 3/6 = 0.5; Phổ biến
  + Các tập phổ biến có 4 phần tử gồm:

C4 = {i2,i4,i5,i6}

* Các tập phổ biến thỏa ngưỡng gồm: {i1},{i2},{i3},{i4},{i5},{i6},{i1,i2},{i1,i4},{i2,i3},{i2,i4},{i2,i5},{i2,i6},{i4,i5},{i4,i6},{i5,i6},{i1,i2,i4},{i2,i4,i5},{i2,i4,i6},{i2,i5,i6},{i4,i5,i6},{i2,i4,i5,i6}

**Bước 3: Tìm tập phổ biến tối đại**

* Các tập phổ biến tối đại gồm: {i2,i3},{i1,i2,i4},{i2,i4,i5,i6}

**Bước 4: Phát biểu luật kết hợp và tính độ tin cậy của luật kết hợp**

|  |  |  |
| --- | --- | --- |
| Luật | Độ tin cậy luật | Thoả ngưỡng min-conf = 0.75 |
| R1: i1 -> i2,i4 | CF(R1) = SP(i1,i2,i4) / SP(i1) = 0.5/(4/6) = 0.75 | Có |
| R2: i2 -> i1,i4 | CF(R2) = SP(i1,i2,i4) / SP(i2) = 0.5/1.0 = 0.5 | Không |
| R3: i4 -> i1,i2 | CF(R3) = SP(i1,i2,i4) / SP(i4) = 0.5/(5/6) = 0.6 | Không |
| R4: i1,i2 -> i4 | CF(R4) = SP(i1,i2,i4) / SP(i1,i2) = 0.5/(4/6) = 0.75 | Có |
| R5: i1,i4 -> i2 | CF(R5) = SP(i1,i2,i4) / SP(i1,i4) = 0.5/0.5 = 1.0 | Có |
| R6: i2,i4 -> i1 | CF(R6) = SP(i1,i2,i4) / SP(i2,i4) = 0.5/(5/6) = 0.6 | Không |

|  |  |
| --- | --- |
| **TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**  **ĐẠI HỌC QUỐC GIA HỒ CHÍ MINH** | **IE403.E21.CN1.CNTT**  **Khai thác dữ liệu truyền thông xã hội** |

**CHƯƠNG 2: TẬP THÔ**

**Bài 1:** Cho bảng quyết định của lớp Anh Văn với:

* 7 đối tượng { X1… X7}
* 2 thuộc tính điều kiện {Độ tuổi, Số buổi}
* 1 thuộc tính quyết định {Thi đậu}

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Độ tuổi** | **Số buổi** | **Thi đậu** |
| **X1** | 16-30 | 50 | Yes |
| **X2** | 16-30 | 0 | No |
| **X3** | 31-45 | 1-25 | No |
| **X4** | 31-45 | 1-25 | Yes |
| **X5** | 46-60 | 26-49 | No |
| **X6** | 16-30 | 26-49 | Yes |
| **X7** | 46-60 | 26-49 | No |

**ĐÁP ÁN**

**Bước 1: Biểu diễn quan hệ bất khả phân biệt**

* IND{Độ tuổi} = {{X1,X2,X6},{X3,X4},{X5,X7}}
* IND{Số buổi} = {{X1},{X2},{X3,X4},{X5,X6,X7}}
* IND{Độ tuổi, Số buổi} = {{X1},{X2},{X3,X4},{X5,X7},{X6}}
* {X3,X4} gọi là tập thô vì không thể có một định nghĩa quyết định là đối tượng đậu hay không đậu.

**Bước 2: Biểu diễn xấp xỉ tập hợp**

Gọi: Tập các đối tượng W = {X | Thi đậu(X) = Yes} ⬄ W = {X1,X4,X6}

Và B = {Độ tuổi, Số buổi}

U = IND{Độ tuổi, Số buổi} = {{X1},{X2},{X3,X4},{X5,X7},{X6}}

* BW = {X1,X6} là những cặp IND là tập con hoàn toàn của W
* BW = {X1,X3,X4,X6} là những cặp IND có chứa các phần tử thuộc W
* Vùng B-biên của W: BNA(W) = BW - BW = {X3,X4}
* Vùng B-ngoài của W: U – BW = {X2,X5,X7}
* Một tập được gọi là thô (rough) nếu vùng biên của nó khác rỗng, ngược lại gọi là tập rõ. Như vậy, lớp quyết định Thi đậu là thô vì vùng biên khác rỗng.

**Bước 3: Tính độ chính xác của tập thô**

αB(W) = |B(W)| / |B(W)| = 2/4 = 0.5 thì W là thô so với B.

Lưu ý: Nếu αB(W) = 1 thì W mới được gọi là rõ so với B.

**Bài 2:** Ví dụ về rút gọn:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Bằng\_cấp  (d) | Kinh\_nghiệm  (e) | Tiếng\_Anh  (f) | Giới\_thiệu  (r) | Tuyển\_dụng |
| X1 | MBA | Vừa | Tốt | Xuất\_sắc | Chấp\_nhận |
| X4 | MSC | Nhiều | Tốt | Trung\_bình | Chấp\_nhận |
| X6 | MSC | Nhiều | Tốt | Xuất\_sắc | Chấp\_nhận |
| X7 | MBA | Nhiều | Không | Tốt | Chấp\_nhận |
| X2 | MBA | Thấp | Tốt | Trung\_bình | Từ\_chối |
| X3 | MCE | Thấp | Tốt | Tốt | Từ\_chối |
| X5 | MSC | Vừa | Tốt | Trung\_bình | Từ\_chối |
| X8 | MCE | Thấp | Không | Xuất\_sắc | Từ\_chối |

**ĐÁP ÁN**

Đặt:

* Bằng cấp: d
* Kinh nghiệm: e
* Tiếng Anh: f
* Giới thiệu: r

**Bước 1: Lập ma trận phân biệt**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | [x1] | [x4] | [x6] | [x7] | [x2] | [x3] | [x5] | [x8] |
| [x1] | ∅ |  |  |  |  |  |  |  |
| [x4] | ∅ | ∅ |  |  |  |  |  |  |
| [x6] | ∅ | ∅ | ∅ |  |  |  |  |  |
| [x7] | ∅ | ∅ | ∅ | ∅ |  |  |  |  |
| [x2] | e,r | d,e | d,e,r | e,f,r | ∅ |  |  |  |
| [x3] | d,e,r | d,e,r | d,e,r | d,e,f | ∅ | ∅ |  |  |
| [x5] | d,r | e | e,r | d,e,f,r | ∅ | ∅ | ∅ |  |
| [x8] | d,e,f | d,e,f,r | d,e,f | d,e,r | ∅ | ∅ | ∅ | ∅ |

*Lưu ý: Chỉ tính những thuộc tính có giá trị quyết định khác nhau.*

**Bước 2: Xác định hàm phân biệt và rút gọn hàm**

f = (e v r) ^ (d v e) ^ (d v e v r) ^ (e v f v r) ^ (d v e v f) ^ (d v r) ^ e ^ (d v e v f v r) = e ^ (d v r) = (e ^ d ) v (e ^ r)

Vậy, hệ quyết định có 2 rút gọn là:

* Bằng cấp, Kinh nghiệm
* Kinh nghiệm, Giới thiệu

**Bài 3:** Cho hệ quyết định như sau:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Trời  (a) | Gió  (b) | Áp suất  (c) | Kết quả |
| O1 | Trong | Bắc | Cao | Không mưa |
| O2 | Mây | Nam | Cao | Mưa |
| O3 | Mây | Bắc | Trung bình | Mưa |
| O4 | Trong | Bắc | Thấp | Không mưa |
| O5 | Mây | Bắc | Thấp | Mưa |
| O6 | Mây | Bắc | Cao | Mưa |
| O7 | Mây | Nam | Thấp | Không mưa |
| O8 | Trong | Nam | Cao | Không mưa |

**Yêu cầu:**

a. Tính xấp xỉ tập X = {O1,O3,O4} qua tập thuộc tính B = {Trời,Gió}

b. Khảo sát sự phụ thuộc thuộc tính của C = {Kết quả} vào B = {Trời,Gió}

**ĐÁP ÁN**

*a. Tính xấp xỉ tập X = {O1,O3,O4} qua tập thuộc tính B = {Trời,Gió}*

**Bước 1: Biểu diễn quan hệ bất khả phân biệt**

IND{Trời,Gió} = {{O1,O4},{O2,O7},{O3,O5,O6},{O8}}

**Bước 2: Biểu diễn xấp xỉ tập hợp**

Với X = {O1,O3,O4} qua tập thuộc tính B = {Trời,Gió}, ta có:

* BX = {O1,O4} là những cặp IND là tập con hoàn toàn của X
* BX = {O1,O3,O4,O5,O6} là những cặp IND có chứa các phần tử thuộc X
* Vùng B-biên của X: BNA(X) = BX - BX = {O3,O5,O6}
* Vùng B-ngoài của X: U – BX = {O2,O7,O8}

**Bước 3: Tính độ chính xác của tập thô**

αB(X) = |B(X)| / |B(X)| = 2/5 = 0.4 thì X là thô so với B.

Lưu ý: Nếu αB(W) = 1 thì W mới được gọi là rõ so với B.

*b. Khảo sát sự phụ thuộc thuộc tính của C = {Kết quả} vào B = {Trời,Gió}*

Với C = {Kết quả} vào B = {Trời,Gió}, ta có:

* X1 = {O1,O4,O7,O8}
* X2 = {O2,O3,O5,O6}
* BX1 = {O1,O4,O8}
* BX2 = {O3,O5,O6}

Do vậy:

*k* = ( |BX1| + | BX2| ) / |O| = 6 / 8 = 0.75

**Bài 4:** Cho bảng quyết định sau:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| O | Kích thước  (a) | Màu sắc  (b) | Hình dạng  (c) | Lớp |
| 1 | Vừa | Xanh | Viên gạch | A |
| 2 | Nhỏ | Đỏ | Hình nêm | B |
| 3 | Nhỏ | Đỏ | Hình cầu | A |
| 4 | Lớn | Đỏ | Hình nêm | B |
| 5 | Lớn | Lục | Hình trụ | A |
| 6 | Lớn | Đỏ | Hình trụ | B |
| 7 | Lớn | Lục | Hình cầu | A |

Thuộc tính: Kích thước, Màu sắc, Hình dạng là thuộc tính điều kiện. Thuộc tính Lớp là thuộc tính quyết định.

**Yêu cầu:**

a. Tính xấp xỉ tập X gồm các đối tượng thuộc lớp A qua tập thuộc tính B = {Kích thước, Màu sắc}.

b. Tính độ chính xác của tập thô theo {Kích thước,Màu sắc}.

c. Khảo sát sự phụ thuộc thuộc tính của C = {Lớp} vào B = {Kích thước,Màu sắc}

d. Tìm các rút gọn của hệ quyết định trên.

**ĐÁP ÁN**

*a. Tính xấp xỉ tập X = {O1,O3,O5,O7} qua tập thuộc tính B = {Kích thước,Màu sắc}*

**Bước 1: Biểu diễn quan hệ bất khả phân biệt**

IND{Kích thước,Màu sắc} = {{O1},{O2,O3},{O4,O6},{O5,O7}}

**Bước 2: Biểu diễn xấp xỉ tập hợp**

Với X = {O1,O3,O5,O7} qua tập thuộc tính B = {Kích thước,Màu sắc}, ta có:

* BX = {O1,O5,O7} là những cặp IND là tập con hoàn toàn của X
* BX = {O1,O2,O3,O5,O7} là những cặp IND có chứa các phần tử thuộc X
* Vùng B-biên của X: BNA(X) = BX - BX = {O2,O3}
* Vùng B-ngoài của X: U – BX = {O4,O6}

*b. Tính độ chính xác của tập thô theo {Kích thước,Màu sắc}*

**Bước 3: Tính độ chính xác của tập thô**

αB(X) = |B(X)| / |B(X)| = 3/5 = 0.6 thì X là thô so với B.

Lưu ý: Nếu αB(W) = 1 thì W mới được gọi là rõ so với B.

*c. Khảo sát sự phụ thuộc thuộc tính của C = {Lớp} vào B = {Kích thước,Màu sắc}*

Với C = {Lớp} vào B = {Kích thước,Màu sắc}, ta có:

* X1 = {O1,O3,O5,O7}
* X2 = {O2,O4,O6}
* BX1 = {O1,O5,O7} là những cặp IND là tập con hoàn toàn của X
* BX2 = {O4,O6} là những cặp IND là tập con hoàn toàn của X

Do vậy:

*k* = ( |BX1| + | BX2| ) / |O| = 5 / 7 = 0.71

*d. Tìm các rút gọn của hệ quyết định trên*

Đặt:

* Kích thước: a
* Màu sắc: b
* Hình dạng: c

**Bước 1: Lập ma trận phân biệt**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | [O1] | [O2] | [O3] | [O4] | [O5] | [O6] | [O7] |
| [O1] | ∅ |  |  |  |  |  |  |
| [O2] | a,b,c | ∅ |  |  |  |  |  |
| [O3] | ∅ | c | ∅ |  |  |  |  |
| [O4] | a,b,c | ∅ | a,c | ∅ |  |  |  |
| [O5] | ∅ | a,b,c | ∅ | b,c | ∅ |  |  |
| [O6] | a,b,c | ∅ | a,c | ∅ | b | ∅ |  |
| [O7] | ∅ | a,b,c | ∅ | b,c | ∅ | b,c | ∅ |

*Lưu ý: Chỉ tính những thuộc tính có giá trị quyết định khác nhau.*

**Bước 2: Xác định hàm phân biệt và rút gọn hàm**

f = (a v b v c) ^ c ^ (a v c) ^ (b v c) ^ b = b ^ c

Vậy, hệ quyết định có 1 rút gọn (Reduct) là: {b,c} ⬄ {Màu sắc,Hình dạng}

|  |  |
| --- | --- |
| **TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**  **ĐẠI HỌC QUỐC GIA HỒ CHÍ MINH** | **IE403.E21.CN1.CNTT**  **Khai thác dữ liệu truyền thông xã hội** |

**PHÂN LỚP DỮ LIỆU (CLASSIFICATION)**

**Bài 1:** Xây cây quyết định dùng giải thuật ID3 bằng cách tính hệ số Gain.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Day** | **Outlook** | **Temperature** | **Humidity** | **Wind** | **Play?** |
| 1 | Sunny | Hot | High | Weak | No |
| 2 | Sunny | Hot | High | Strong | No |
| 3 | Overcast | Hot | High | Weak | Yes |
| 4 | Rainy | Mild | High | Weak | Yes |
| 5 | Rainy | Cool | Normal | Weak | No |
| 6 | Rainy | Cool | Normal | Strong | Yes |
| 7 | Overcast | Cool | Normal | Weak | Yes |
| 8 | Sunny | Mild | High | Weak | Yes |
| 9 | Sunny | Cool | Normal | Weak | Yes |
| 10 | Rainy | Mild | Normal | Strong | Yes |
| 11 | Sunny | Mild | Normal | Strong | Yes |
| 12 | Overcast | Mild | High | Strong | Yes |
| 13 | Overcast | Hot | Normal | Weak | Yes |
| 14 | Rainy | Mild | High | Strong | No |

**ĐÁP ÁN**

**Bước 1: Tính khối lượng thông tin cần thiết cho 14 dòng dữ liệu và 2 lớp quyết định “Play”**

Ta có:

S = 14 vì tổng cộng có 14 dòng dữ liệu

m = 2 vì chỉ có 2 lớp “Yes” và “No”

C1 = “Yes”, C2 = “No”

S1 = 10 vì có 10 “Yes”, S2 = 4 vì có 4 “No”

I(S1,S2) = I(10,4) = – (10/14) log2(10/14) – (4/14) log2(4/14) = 0.863

**Bước 2: Tính entropy của từng thuộc tính Outlook, Temperature, Humidity, Wind**

Vớipilà lớp “Yes”, ni là lớp “No”:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Outlook | pi | ni | pi + ni | I(pi,ni) |
| Sunny | 3 | 2 | 5 | I(3,2) = 0.971 |
| Overcast | 4 | 0 | 4 | I(4,0) = 0 |
| Rainy | 3 | 2 | 5 | I(3,2) = 0.971 |

Ta có: E(Outlook) = (5/14) I(3,2) + (4/14) I(4,0) + (5/14) I(3,2) = 0.694

Do đó: Gain(Outlook) = I(10,4) – E(Outlook) = 0.169

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Temperature | pi | ni | pi + ni | I(pi,ni) |
| Hot | 2 | 2 | 4 | I(2,2) = 1 |
| Mild | 5 | 1 | 6 | I(5,1) = 0.65 |
| Cool | 3 | 1 | 4 | I(3,1) = 0.811 |

Ta có: E(Temperature) = (4/14) I(2,2) + (6/14) I(5,1) + (4/14) I(3,1) = 0.796

Do đó: Gain(Temperature) = I(10,4) – E(Temperature) = 0.067

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Humidity | pi | ni | pi + ni | I(pi,ni) |
| High | 4 | 3 | 7 | I(4,3) = 0.985 |
| Normal | 6 | 1 | 7 | I(6,1) = 0.592 |

Ta có: E(Humidity) = (7/14) I(4,3) + (7/14) I(6,1) = 0.789

Do đó: Gain(Humidity) = I(10,4) – E(Humidity) = 0.074

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Wind | pi | ni | pi + ni | I(pi,ni) |
| Weak | 6 | 2 | 8 | I(6,2) = 0.811 |
| Strong | 4 | 2 | 6 | I(4,2) = 0.918 |

Ta có: E(Wind) = (8/14) I(6,2) + (6/14) I(4,2) = 0.857

Do đó: Gain(Wind) = I(10,4) – E(Wind) = 0.006

* Kết luận 1: Gain(Oulook) là lớn nhất trong số các gain của các thuộc tính. Vì vậy, chọn Outlook là thuộc tính phân nhánh đầu tiên. Outlook phânh nhánh thành Sunny, Overcast và Rainy.

**Bước 3: Lập bảng con cho từng giá trị của Outlook và tính entropy của 3 thuộc tính còn lại**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Day** | **Outlook** | **Temperature** | **Humidity** | **Wind** | **Play?** |
| 1 | Sunny | Hot | High | Weak | No |
| 2 | Sunny | Hot | High | Strong | No |
| 8 | Sunny | Mild | High | Weak | Yes |
| 9 | Sunny | Cool | Normal | Weak | Yes |
| 11 | Sunny | Mild | Normal | Strong | Yes |

Với Outlook = Sunny: I(3,2) = 0.971

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Temperature | pi | ni | pi + ni | I(pi,ni) |
| Hot | 0 | 2 | 2 | I(0,2) = 0 |
| Mild | 2 | 0 | 2 | I(2,0) = 0 |
| Cool | 1 | 0 | 1 | I(1,0) = 0 |

Ta có: E(Temperature) = (2/5) I(0,2) + (2/5) I(2,0) + (1/5) I(1,0) = 0

Do đó: Gain(Temperature) = I(3,2) – E(Temperature) = 0.971

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Humidity | pi | ni | pi + ni | I(pi,ni) |
| High | 1 | 2 | 3 | I(1,2) = 0.918 |
| Normal | 2 | 0 | 2 | I(2,0) = 0 |

Ta có: E(Humidity) = (3/5) I(1,2) + (2/5) I(2,0) = 0.551

Do đó: Gain(Humidity) = I(3,2) – E(Humidity) = 0.421

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Wind | pi | ni | pi + ni | I(pi,ni) |
| Weak | 2 | 1 | 3 | I(2,1) = 0.918 |
| Strong | 1 | 1 | 2 | I(1,1) = 1 |

Ta có: E(Wind) = (3/5) I(2,1) + (2/5) I(1,1) = 0.951

Do đó: Gain(Wind) = I(3,2) – E(Wind) = 0.02

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Day** | **Outlook** | **Temperature** | **Humidity** | **Wind** | **Play?** |
| 3 | Overcast | Hot | High | Weak | Yes |
| 7 | Overcast | Cool | Normal | Weak | Yes |
| 12 | Overcast | Mild | High | Strong | Yes |
| 13 | Overcast | Hot | Normal | Weak | Yes |

* Kết luận 2: Gain(Temperature) là lớn nhất trong số các gain của các thuộc tính. Vì vậy, chọn Temperature là thuộc tính phân nhánh của Sunny.

Với Outlook = Overcast: I(4,0) = 0

* Kết luận 3:

Với Outlook = Rainy: I(3,2) = 0.971

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Temperature | pi | ni | pi + ni | I(pi,ni) |
| Mild | 2 | 1 | 3 | I(2,1) = 0.918 |
| Cool | 1 | 1 | 2 | I(1,1) = 1 |

Ta có: E(Temperature) = (3/5) I(2,1) + (2/5) I(1,1) = 0.951

Do đó: Gain(Temperature) = I(3,2) – E(Temperature) = 0.02

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Humidity | pi | ni | pi + ni | I(pi,ni) |
| High | 1 | 1 | 2 | I(1,1) = 1 |
| Normal | 2 | 1 | 3 | I(2,1) = 0.981 |

Ta có: E(Humidity) = (2/5) I(1,1) + (3/5) I(2,1) = 0.951

Do đó: Gain(Humidity) = I(3,2) – E(Humidity) = 0.02

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Wind | pi | ni | pi + ni | I(pi,ni) |
| Weak | 1 | 1 | 2 | I(1,1) = 1 |
| Strong | 2 | 1 | 3 | I(2,1) = 0.981 |

Ta có: E(Wind) = (2/5) I(1,1) + (3/5) I(2,1) = 0.951

Do đó: Gain(Wind) = I(3,2) – E(Wind) = 0.02

* Kết luận 4: Do Gain(Temperature) = Gain(Humidity) = Gain(Wind) = 0.02, ta lựa chọn ngẫu nhiên Humidity là thuộc tính phân nhánh của Rainy.

|  |  |
| --- | --- |
| **TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**  **ĐẠI HỌC QUỐC GIA HỒ CHÍ MINH** | **IE403.E21.CN1.CNTT**  **Khai thác dữ liệu truyền thông xã hội** |

**BÀI TẬP SỬA ĐỀ THI CŨ**

**Bài 1:** Tính phụ thuộc thuộc tính của “Kết quả” vào 2 thuộc tính “Độ tuổi,Mức độ đau ngực”. (1.5đ)

|  |  |  |  |
| --- | --- | --- | --- |
| Ký hiệu | ĐT | ĐN | KQ |
| 1 | Trên 60 | Vừa | Có |
| 2 | 41 – 60 | Nhẹ | Không |
| 3 | Trên 60 | Nặng | Có |
| 4 | Dưới 40 | Nhẹ | Không |
| 5 | Dưới 40 | Không đau | Có |
| 6 | Trên 60 | Vừa | Có |
| 7 | 41 – 60 | Nhẹ | Không |
| 8 | 41 – 60 | Vừa | Có |
| 9 | Trên 60 | Nặng | Có |
| 10 | 41 – 60 | Vừa | Không |

Đặt D = {ĐT, ĐN}, C = {KQ}

Với C = {KQ}, ta có:

X1 = {1,3,5,6,8,9}

X2 = {2,4,7,10}

Ta tính được:

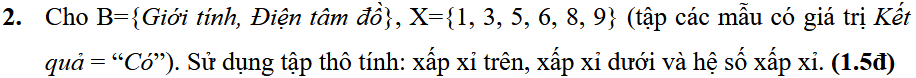
IND(ĐT, ĐN) = {{1,6}, {2,7}, {3,9}, {4}, {5}, {8,10}}

D(X1) = {1, 3, 5, 6, 9}

D(X2) = {2,4,7}

Do vậy:

K = (|D(X1)| + |D(X2)|) / |O| = 8/10 = 0.4



|  |  |  |  |
| --- | --- | --- | --- |
| Ký hiệu | GT | ECG | KQ |
| 1 | Nam | Bất thường | Có |
| 2 | Nữ | Bình thường | Không |
| 3 | Nam | Bất thường | Có |
| 4 | Nữ | Bình thường | Không |
| 5 | Nam | Phì đại | Có |
| 6 | Nữ | Bất thường | Có |
| 7 | Nam | Bất thường | Không |
| 8 | Nữ | Phì đại | Có |
| 9 | Nam | Bất thường | Có |
| 10 | Nữ | Bình thường | Không |

IDN(GT, ECG) = {{1,3,7,9}, {2,4,10},{5},{6},{8}}

X = {1,3,5,6,8,9}

Ta có:

B(X1) = {5,6,8}

Upper(B,X) = {1,3,7,9,5,6,8}

Hệ số xấp xỉ = 3/7 = 0.4285



Độ tuổi:

Trên 60 (4 Có): Gini = 0

41 – 60 (1 có, 3 không): Gini = 1 – (1/4)^2 – (3/4)^2 = 0.45

Dưới 40 (1 có, 1 không): Gini = 1 – (1/2)^2 – (1/2)^2 = 0.5

* Gini (Độ tuổi) = 4/10 \* 0 + 4/10 \* 0.45 + 2/10 \* 0.5 = 0.28

Giới tính:

Nam (4 có, 1 không): Gini = 0.44

Nữ ( 2 có, 3 không) Gini = 0.48

* Gini (Giới tính) = 0.46

ECG:

Bất thường (4 có, 1 không): Gini = 0.44

Bình thường (3 không): Gini = 0

Phi đại (2 có): Gini = 0

* Gini (ECG) = 0.22

ĐN:

Đau vừa (3 có, 1 không): Gini = 0.45

Đau nhẹ ( 3 không): Gini = 0

Đau nặng ( 2 có): Gini = 0

Không đau (1 có): Gini = 0

* Gini (ĐN) = 0.18

**Kết luận:** Gini (ĐN) nhỏ nhất = 0.18, do đó sử dụng ĐN làm gốc cây.

A close up of text

AI-generated content may be incorrect.

P(Có) = (6+1)/(10+2) = 7/12

P(Không) = (4+1)/(10+2) = 5/12

P(Trên 60|Có) = (4+1)/(6+3) = 5/9

P(Trên 60|Không) = 0

P(Nữ|Có) = (2+1)/(6+2) = 3/8

P(Nữ|Không) = (3+1)/(4+2) = 4/6

P(Phi đại|Có) = (2+1)/(6+3) = 3/9

P(Phi đại|Không) = 0

P(Đau vừa|Có) = (3+1)/(6+4) = 4/10

P(Đau vừa|Không) = (1+1)/(4+4) = 2/8

P(Có|X) = 7/12 \* 5/9 \* 3/8 \* 3/9 \* 4/10 = 0.016204

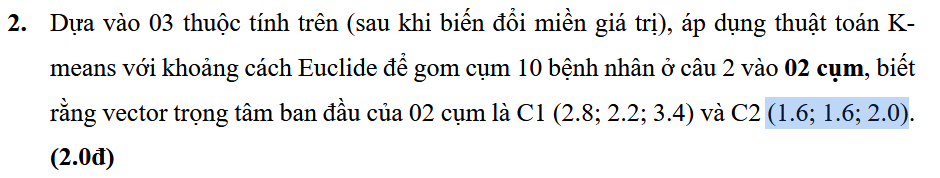
P(Không|X) = 0

* Vậy X thuộc lớp Có

A close up of text

AI-generated content may be incorrect.

|  |  |  |  |
| --- | --- | --- | --- |
| Ký hiệu | ĐT | ECG | ĐN |
| 1 | 3 | 2 | 3 |
| 2 | 2 | 1 | 2 |
| 3 | 3 | 2 | 4 |
| 4 | 1 | 1 | 2 |
| 5 | 1 | 3 | 1 |
| 6 | 3 | 2 | 3 |
| 7 | 2 | 2 | 2 |
| 8 | 2 | 3 | 3 |
| 9 | 3 | 2 | 4 |
| 10 | 2 | 1 | 3 |



C1 = (2.8; 2.2; 3.4)

C2 = (1.6; 1.6; 2.0)

|  |  |  |  |
| --- | --- | --- | --- |
| Ký hiệu | C1 | C2 | Cụm |
| 1 | 0.490 | 1.766 | C1 |
| 2 | 2.010 | 0.721 | C2 |
| 3 | 0.663 | 2.474 | C1 |
| 4 | 2.577 | 0.849 | C2 |
| 5 | 3.105 | 1.822 | C2 |
| 6 | 0.490 | 1.766 | C1 |
| 7 | 1.625 | 0.566 | C2 |
| 8 | 1.200 | 1.766 | C1 |
| 9 | 0.663 | 2.474 | C1 |
| 10 | 1.497 | 1.233 | C2 |