Annalytical characterization

The collected data for target clas: graduate and constructing class: undergrate are given in table I and table II, respectively.

Table I: The collected data for target class: graduate

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Gender | Major | Birth-country | Age-rage | GPA | Count |
| M | Science | Canada | 21…25 | Very good | 16 |
| F | Science | USA | 26…30 | Excellent | 22 |
| M | Engineering | USA | 26…30 | Excellent | 18 |
| F | Science | USA | 26…30 | Excellent | 25 |
| M | Science | USA | 26…30 | Excellent | 21 |
| F | Engineering | Canada | 21…25 | Excellent | 18 |

Table II: The collected data for target class: undergraduate

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Gender | Major | Birth-country | Age-rage | GPA | Count |
| M | Science | USA | <=20 | Very good | 18 |
| F | Business | Canada | <=20 | Fair | 20 |
| M | Business | Canada | <=20 | Fair | 22 |
| F | Science | Canada | 21…25 | Fair | 24 |
| M | Engineering | USA | 21…25 | Very good | 22 |
| F | Engineering | Canada | <=20 | Fair | 24 |

The atributes in the candidate relation are evaluated using information gain. Let C1 and C2 correspond to class “graduate” and “undergraduate”, respectively. Suppose that C1=120, C1=130, where C1 and C2, respectively. Expected information *I* (S1, S2) is

*I* (S1, S2) = -

with S1= C1=120, S2= C2=130. We may compute the entropy (expected information) for each attribute.

1. For Gender

Gender = “M” (attribute 1), we have

S11= 55, S21 = 62, *I* (S11, S21) = - 0.9974

Gender = “ F” (attribute 2), we have

S12= 65, S22 = 68, *I* (S12, S22) = - 0.9996

The expected information for the attribute “gender” is

E (gender) =  *I* (S11, S21) +  *I* (S12, S22) = 0.9986

The information gain for this partition is

Gain (Gender) = *I* (S1, S2) - E (gender) = 0.0023

Similarly, we can compute the information gain for the remaining attributes

1. For Birth-country

Birth-country = “Cannada” (atribute 1), we have

S11= 34, S21 = 90, *I* (S11, S21) = - 0.8474

Birth-country = “USA” (attribute 2), we have

S12= 86, S22 = 40, *I* (S12, S22) = - 0.9016

The expected information for the attribute “Birth-country” is

E (Birth-country) =  *I* (S11, S21) +  *I* (S12, S22) = 0.8747

The information gain for this partition is

Gain (Birth-country) = *I* (S1, S2) - E (Birth-country) = 0.1241

1. For GPA

GPA = “Very good” (atribute 1), we have

S11=16, S21 = 40, *I* (S11, S21) = - 0.8631

GPA = “Excellent” (attribute 2), we have

S12= 104, S22 = 0, *I* (S12, S22) = -

GPA = “Fair” (attribute 3), we have

S12= 0, S22 = 90, *I* (S12, S22) = -

The expected information for the attribute “GPA” is

E (GPA) =  *I* (S11, S21) +  *I* (S12, S22) +  *I* (S13, S23) =

The information gain for this partition is

Gain (GPA) = *I* (S1, S2) - E (GPA) = 0.4493

1. For Age-range

Age-range = “21…25” (atribute 1), we have

S11=34, S21 = 46, *I* (S11, S21) = - 0.9837

Age-range = “26…30” (attribute 2), we have

S12= 86, S22 = 0, *I* (S12, S22) = -

Age-range = “<=20” (attribute 3), we have

S12= 0, S22 = 84, *I* (S12, S22) = -

The expected information for the attribute “Age-range” is

E (Age-range) =  *I* (S11, S21) +  *I* (S12, S22) +  *I* (S13, S23) =

The information gain for this partition is

Gain (Age-range) = *I* (S1, S2) - E (Age-range) = 0.5971