Sucheera Nakkhum 633020449-1 SIDS

Attribute selection with Information gain

age	p_i	n_i	$I(p_i, n_i)$
<=30	2	3	0.971
31...40	4	0	0
>40	3	2	0.971

class P : buys_computer = "yes"
class N : buys_computer = "no"

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

- Expected information (entropy) needed to classify a tuple in D:

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

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

class age
1) <=30 : Y:2 / N:3 (5)
2) 31-40 : Y:4 / N:0 (4)
3) >40 : Y:3 / N:2 (5)

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

$= \frac{5}{14}\left(-\frac{2}{5}\log_2\left(\frac{2}{5}\right) - \frac{3}{5}\log_2\left(\frac{3}{5}\right)\right) + \frac{4}{14}\left(-\frac{4}{4}\log_2\left(\frac{4}{4}\right) - 0\right)$

$+ \frac{5}{14}\left(-\frac{3}{5}\log_2\left(\frac{3}{5}\right) - \frac{2}{5}\log_2\left(\frac{2}{5}\right)\right)$

$= 0.694$

$Info_{income}(D) = \frac{4}{14}I(2,2) + \frac{6}{14}I(4,2) + \frac{4}{14}I(3,1)$

class income
high : Y:2 / N:2 (4)
medium : Y:4 / N:2 (6)
low : Y:3 / N:1 (4)

$= \frac{4}{14}\left(-\frac{2}{4}\log_2\left(\frac{2}{4}\right) - \frac{2}{4}\log_2\left(\frac{2}{4}\right)\right) + \frac{6}{14}\left(-\frac{4}{6}\log\left(\frac{4}{6}\right) - \frac{2}{6}\log\left(\frac{2}{6}\right)\right)$

$+ \frac{4}{16}\left(-\frac{3}{4}\log_2\left(\frac{3}{4}\right) - \frac{1}{4}\log_2\left(\frac{1}{4}\right)\right)$

$= 0.911$

$Info_{student}(D) = \frac{7}{14}I(6,1) + \frac{7}{14}I(3,4)$

$= 0.789$

class student
Yes : Y:6 / N:1 (7)
No : Y:3 / N:4 (7)

$Info_{credit}(D) = \frac{8}{14}I(6,2) + \frac{6}{14}I(3,3)$

$= 0.892$

class credit
fair Y:6 N:2 (8)
excellent Y:3 / N:3 (6)

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

Similarly, we can get

$$Gain(income) = 0.029$$
$$Gain(student) = 0.151$$
$$Gain(credit\_rating) = 0.048$$

Gain(age) : Info(D) - $Info_{age}(D)$ = 0.940 - 0.694 = 0.246 — gainage มากที่สุด

Gain(income) : Info(D) - $Info_{income}(D)$: 0.940 - 0.911 = 0.029

Gain(student) : Info(D) - $Info_{student}(D)$ = 0.940 - 0.789 = 0.151

Gain(credits_rating) : Info(D) - $Info_{credit}(D)$ = 0.940 - 0.892 = 0.048

① <=30

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
<=30	medium	yes	excellent	yes

yes : 2

No : 3

income
- high : Y:0 / N:2 (2)
- medium : Y:1 / N:1 (2)
- low : Y:1 / N:0 (1)

student
- yes : Y:2 N:0 (2)
- no : Y:0 N:3 (3)

credits
- fair : Y:1 / N:2 (3)
- excellent : Y:1 / N:1 (2)

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

$Info_{income}(D)$: $\frac{2}{5}I(0,2) + \frac{2}{5}I(1,1) + \frac{1}{5}(1,0) = 0.4$

$Info_{student}(D)$: $\frac{2}{5}I(2,0) + \frac{3}{5}I(0,3) = 0$

$Info_{credit}(D)$: $\frac{3}{5}I(1,2) + \frac{2}{5}(1,1) \approx 0.951$

Gain (income) : Info(D) - $Info_{income}(D)$: 0.971 - 0.4 = 0.571

Gain (student) : Info(D) - $Info_{student}(D)$: 0.971 - 0 : 0.971

Gain (credits) : info(D) - $Info_{credits}(D)$ = 0.971 - 0.951 : 0.02

student มากที่สุด

2) 31...40

age	income	student	credit_rating	buys_computer
31...40	high	no	fair	yes
31...40	low	yes	excellent	yes
31...40	medium	no	excellent	yes
31...40	high	yes	fair	yes

yes 4
NO 0

Income
- high Y:2 N:0 (2)
- medium Y:1 N:0 (1)
- low Y:1 N:0 (1)

student
- yes Y:2 N:0 (2)
- NO Y:2 N:0 (2)

Credit
- fair Y2 N:0 (2)
- excellent Y:2 N:0 (2)

3) age > 40

age	income	student	credit_rating	buys_computer
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
>40	medium	yes	fair	yes
>40	medium	no	excellent	no

yes (3)
NO (2)

income
- high Y:0 N:0 (0)
- medium Y:2 N:1 (3)
- low Y:1 N:1 (2)

student
- yes Y:2/N:1 (3)
- no Y:1/N:1 (2)

credits
- fair : Y:3 / N:0 (3)
- excellent : Y:0 / N:2 (2)

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

$\text{Info}_{income}(D) = \frac{3}{5}I(2,1) + \frac{2}{5}I(1,1) = 0.951$

$\text{Info}_{student}(D) = \frac{3}{5}I(2,1) + \frac{2}{5}(1,1) = 0.951$

$\text{Info}_{credit}(D) = \frac{3}{5}I(3,0) + \frac{2}{5}(0,2) = 0$

Gain(income) = 0.971 - 0.951 = 0.02

Gain(student) = 0.971 - 0.951 = 0.02

Gain(credit_rating) = 0.971 - 0 = 0.971

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

age	buys_computer
<=30	no
<=30	no
<=30	no
<=30	yes
<=30	yes

Gain student : 0.791
เช็ค student เป็นเกณฑ์

age	buys_computer
31…40	yes
31…40	yes
31…40	yes
31…40	yes

buy

age	buys_computer
>40	yes
>40	yes
>40	no
>40	yes
>40	no

Gain Credit (0.971)

student	buys_computer
no	no
no	no
no	no
yes	yes
yes	yes

no
(not student)
→ not buy

yes
(I'm student)
→ buy

credit_rating	buys_computer
fair	yes
fair	yes
excellent	no
fair	yes
excellent	no

excellent → not buy

fair → buy

ชุติรา นาคคำ 6330204491