Ankit Raj 1906534.

Answer 2)(a) Compactness or cluster cohesion: Measure how close are the objects within the same cluster. A lower within-cluster variation is an indicator of a good compactness (i.e. a good clustering). The different indices for evaluating the compactness of clusters are based on distance measure such as the cluster wise within average/median distance between observations.

~~Sb~~ Separation: Measure how well-
~~Sb~~ Separated a cluster is ~~b~~ from other cluster. The indices used as separation measure include:

① distances between cluster centers
② the pairwise minimum distances between objects in different clusters.

Step1: choose the number of clusters k.

Step2: select K random points from the data as centroids

Step3: Assign all the points to the ~~closest~~ closest cluster centroid

Step4: Recompute the centroid of newly formed clusters.

Step5: Repeat 3 and 4.

Scanned by TapScanner

There are essentially three stopping criteria that can be adopted to stop the K-mean algorithm:

Centroids of newly formed clusters do not change points remain in the same cluster Maximum number of iterations are reached.

(b) Information Gain $= -\left[\frac{7}{16}\log_2\frac{7}{16} + \frac{9}{16}\log_2\frac{9}{16}\right]$

$= -\left[0.4375\log_2(0.4375) + \cancel{0.5625\log} + 0.5625\log_2(0.5625)\right]$

$= -\left[0.4375(-1.1926) + 0.5625(-0.83)\right]$

$= 0.5217 + 0.4668$

$= 0.9885$

~~Find~~ finding splitting Attribute

(i) A

	1	0
1	3	3
2	3	1
3	3	3

$E(A) = I(1) = \left[\frac{3}{6}\log_2\frac{3}{6} + \frac{3}{6}\log_2\frac{3}{6}\right]$

Scanned by TapScanner

$$= \left[2 \times \frac{3}{6} \times \log_2 \frac{3}{6} \right]$$

$$= +1$$

$$I(2) = -\left[\frac{3}{4} \log_2 \frac{3}{4} + \frac{1}{4} \log_2 \frac{1}{4} \right]$$

$$= -\left[0.75 \log_2 (0.75) + 0.25 \log_2 (0.25) \right]$$

$$= 0.75 (0.415) + 0.25 (2)$$

$$= 0.31125 + 0.5$$

$$= 0.81125$$

$$I(3) = -\left(\frac{3}{6} \log_2 \frac{3}{6} + \frac{3}{6} \log_2 \frac{3}{6} \right)$$

$$= +1$$

$$E(A) = I(1) \times P(1) + I(2) \times P(2) + I(3) \times P(3)$$

$$= \frac{6}{16} \times 1 + \frac{4}{16} \times 0.811 + \frac{1 \times 6}{16}$$

$$= 0.375 + 0.2027 + 0.375$$

$$= 0.9527$$

gain = Info. gain − Entropy

$$= 0.9005 - 0.9522$$

$$= 0.0358$$

Scanned by TapScanner

1900534

(ii) B

		0	1
E(B)	30	5	3
	25	6	2

$$I(30) = -\left[\frac{5}{8}\log_2\frac{5}{8} + \frac{3}{8}\log_2\frac{3}{8}\right]$$

$$= -(0.625 \times \log_2(0.625) + 0.375\log_2(0.375)$$

$$\simeq 0.625 \times 0.6781 + (0.375) \times (1.415)$$

$$= 0.4238 + 0.5306$$

$$= 0.9544$$

$$I(25) = -\left[\frac{3}{4}\log_2\frac{3}{4} + \frac{1}{4}\log_2\frac{1}{4}\right]$$

$$= 0.81125$$

$$E(B) = I(25) \times P(25) + I(30) \times P(30)$$

$$= 0.81125 \times \frac{1}{2} + 0.9544 \times \frac{1}{2}$$

$$= 0.8828$$

$$\text{Gain} = 0.9885 - 0.8825$$

$$= 0.106$$

Scanned by TapScanner

1906534

(iii) C

	1	0
25	2	4
30	1	2
75	4	3

$$I(25) = -\left[\frac{1}{3}\log_2\frac{1}{3} + \frac{2}{3}\log_2\frac{2}{3}\right]$$

$$= 0.33 \times 1.5995 + 0.66 \times 0.5995$$

$$= 0.9235$$

$$I(30) = -\left[\frac{1}{3}\log_2\frac{1}{3} + \frac{2}{3}\log_2\frac{2}{3}\right)$$

$$= 0.9235$$

$$I(75) = -\left(\frac{4}{7}\log_2\frac{4}{7} + \frac{3}{7}\log_2\frac{3}{7}\right)$$

$$= 0.9852$$

$$E(C) = I(25) \times P(25) + I(30) \times P(30) + I(75) \times P(25)$$

$$= 0.9235 \times \frac{6}{16} + 0.9235 \times \frac{3}{16} + 0.9852 \times \frac{2}{16}$$

$$= 0.346 + 0.1737 + 0.431$$

$$= 0.95$$

gain = ~~0.99~~ 0.9885 − 0.95

= 0.0385.

Scanned by TapScanner

1906534

classmate
Date
Page

gain (A) = 0.0358
gain (B) = 0.106
gain (c) = 0.038

The gain of B is the greatest therefore the first splitting attribute will be 'B'.

Scanned by TapScanner