RAND INDEX

Given a set of n objects $S = \{O_1, O_2, O_3 \ldots O_n\}$

Suppose $U = \{u_1, u_2, u_3 \ldots u_R\}$ and $V = \{v_1, v_2, v_3 \ldots v_C\}$ represents two different partitions of the objects such that $\bigcup_{i=1}^{R} u_i = S = \bigcup_{j=1}^{C} v_j$ and $u_i \cap u_{i'} = \phi$

$= v_j \cap v_{j'} = \phi$

Suppose U is actual criteria.
and V is clustering result.

Let

a ⇒ The number of pairs of objects in the same class in U and in the same cluster in V.

b ⇒ The number of pairs of objects in the same class in U but not in the same cluster V.

c ⇒ The number of pairs of objects in the different class in U, but in same cluster in V.

d ⇒ The number of pairs of objects in the different class in U, and in the different cluster in V.

U/V	Same Cluster V	Different
Same Class	A (TP)	C (FN)
different	B (FP)	D (TN)

TP ⇒ Two similar documents (objects) assigned to same cluster.

TN ⇒ Two dissimilar documents (objects) assigned to different clusters.

FP ⇒ Two dissimilar documents (objects) assigned to same cluster.

FN ⇒ Two similar documents (objects) assigned to different cluster.

Example:

class / cluster	v_1	v_2	----	v_c	SUM
u_1	n_{11}	n_{12}	---	n_{1c}	$n_{1.}$
u_2	n_{21}	n_{22}	---	n_{2c}	$n_{2.}$
⋮	⋮	⋮		⋮	⋮
u_R	n_{R1}	R_{R2}	----	n_{Rc}	$n_{R.}$
SUM	$n_{.1}$	$n_{.2}$	---	$R_{.c}$	$n_{..} = n$

class / cluster	v_1	v_2	v_3	SUMS
u_1	1	1	0	2
u_2	1	2	1	4
u_3	0	0	4	4
SUMS	2	3	5	$n = 10$

For the given example a is

$$a = \sum_{i,j} \binom{n_{ij}}{2} = \binom{2}{2} + \binom{4}{2} = \underline{7}$$

$$b = \sum_{i} \binom{n_{i.}}{2} - \sum_{i,j} \binom{n_{ij}}{2} = \binom{2}{2} + \binom{4}{2} + \binom{4}{2} - 7 = \underline{6}$$

$$c = \sum_{j} \binom{n_{.j}}{2} - \sum_{i,j} \binom{n_{ij}}{2} = \binom{2}{2} + \binom{3}{2} + \binom{5}{2} - 7 = \underline{7}$$

$$d = \binom{n}{2} - (a+b+c)$$

$$= \binom{10}{2} - 7 - 6 - 7 = \underline{25}$$

class/cluster	same	diffend
same	a(TP) 7	c(FN) 7
different	b(FP) 6	d(TN) 25

$$\text{Rand Index} = \frac{a+d}{a+b+c+d}$$

$$P = \frac{A}{A+B}, \quad R = \frac{A}{A+C}$$

$$F\text{-measure} \quad \frac{2PR}{P+R}$$

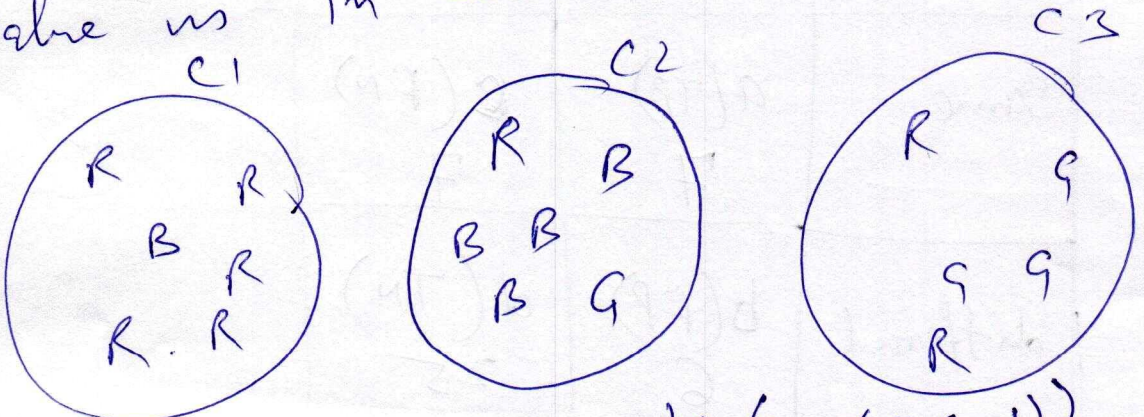
Purity $\Rightarrow$ maximum count of correctly assigned documents divided by total no. of documents.

$$\text{Purity}(\Omega, C) = \frac{1}{N} \sum_k \max_j |w_k \cap c_j|$$

$\Omega = \{w_1, w_2, w_3 \ldots w_k\}$

$C = \{c_1, c_2, c_3 \ldots c_j\}$.

value is in between 0 and 1.

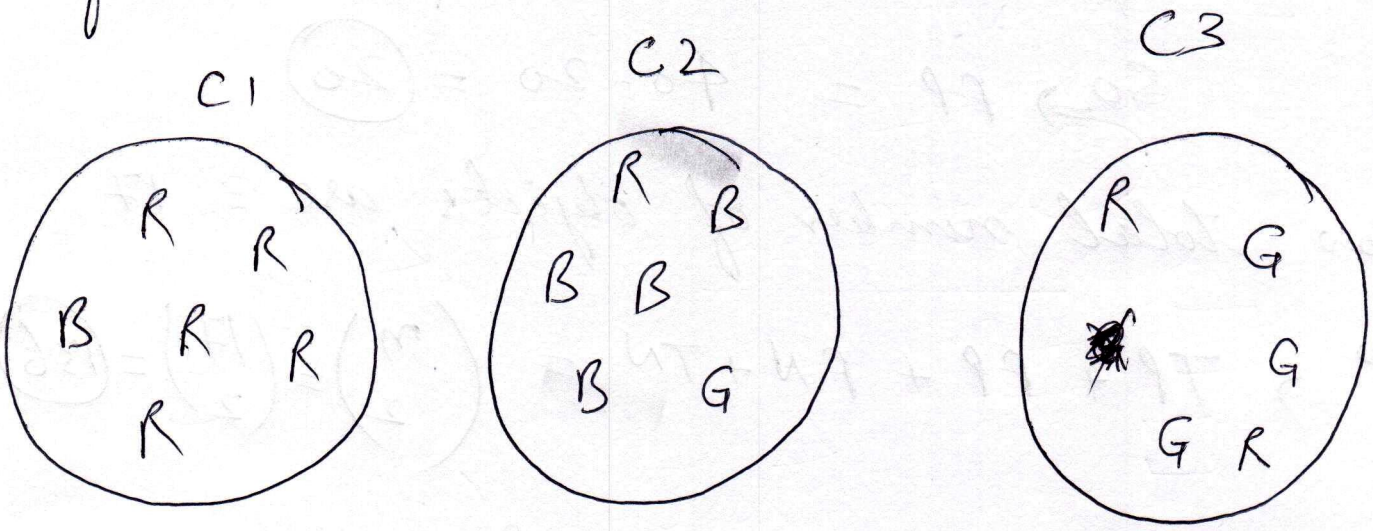


Cluster - III = $\frac{1}{5}(\max(2,3,0$
$= 3/5$

Cluster - I Purity = $\frac{1}{6}(\max(5,0,1)) = 5/6$

Cluster - II Purity = $\frac{1}{6}(\max(1,4,1)) = 4/6$.

Rand Index for the same example can be calculated as follows:



As per the definition of Rand index, we count the number of pairs of objects belong to same class and whether they are in the same cluster as well.

$\frac{TP+FP}{}$ The three cluster contains 6, 6 and 5 objects so total number of pairs of object that are in the same cluster are:

$$TP+FP = \binom{6}{2} + \binom{6}{2} + \binom{5}{2} = 40$$

out of these, the R pairs in cluster(C1), the B pairs in cluster(C2) and R and G pairs in cluster(C3) are true positive.

$$TP = \underbrace{\binom{5}{2}}_{(C1)} + \underbrace{\binom{4}{2}}_{(C2)} + \underbrace{\binom{3}{2} + \binom{2}{2}}_{(C3)}$$

$$= 10 + 6 + 3 + 1 = 20$$

Thus $TP + FP = 40$

$\Rightarrow TP = 20$

So $\Rightarrow FP = 40 - 20 = 20$

Now total number of objects are = 17

So $\Rightarrow TP + FP + FN + TN = \binom{n}{2} = \binom{17}{2} = 136$

FN $\Rightarrow$ (False negative) is when the number of pair of objects in the different class but in same cluster.

TN $\Rightarrow$ (True negative) is when the number of pair of objects in the different class and in different clusters as well.

So $TN = \binom{7}{2} + \binom{10}{2} + \binom{4}{2}$

$\binom{7}{2} \Rightarrow$ 4 'B' in C2 and 3 'G' in C3

$\Rightarrow 4 + 3 = 7$ (cluster 1)

$\binom{10}{2} \Rightarrow$ 5 'R' in C1 and (2 'R' and 3 'G' in C3)

$\Rightarrow 5 + 5 = 10$ (cluster 2)

$\binom{4}{2} \Rightarrow$ Cluster 2 contains 4 'B'.

$= 4$ (cluster 3)

$$TN = \binom{7}{2} + \binom{10}{2} + \binom{4}{2}$$

$$= \frac{7\times6}{2} + \frac{10\times9}{2} + \frac{4\times3}{2}$$

$$= 21 + 45 + 6 = 72$$

$$\boxed{TN = 72}$$

Now $TP + FP + FN + TN = 136$

$$FN = 136 - (TP + FP) - TN$$

$$= 136 - 40 - 72$$

$$\boxed{FN = 24}$$

So the confusion matrix is also known as Contingency table :

	TP (a) 20	FN (c) 24
	FP (b) 20	TN (d) 72

$$RA = \frac{a+d}{a+b+c+d} = \frac{92}{136} = 0.6764$$

$$P = \frac{a}{a+b}, \quad R = \frac{a}{a+c}, \quad F = \frac{2PR}{P+R}$$

So F-measure for the given example is

$$P = \frac{20}{40} = 0.5 \quad , \quad R = \frac{20}{44} = 0.45$$

$$F = \frac{2 \times 0.5 \times 0.45}{0.5 + 0.45}$$

$$\boxed{F = 0.47}$$