Sam Nuttall

201608203

Data Mining and Visualisation – Assignment 2 - Questions 5 and 6

5. Compute the confusion matrix, macro-averaged Precision, Recall, and F-score for the clustering shown in Figure 1. (20)

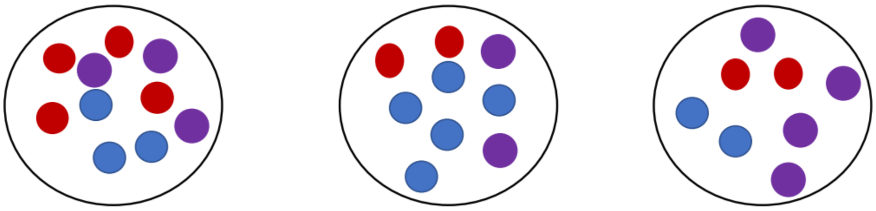

Figure 1: Outcome of a Clustering Algorithm

RED BLUE PURPLE

For the sake of the calculations, I have assigned predicted values of Red, Blue and Purple to the clusters left to right respectively. The labels are assigned by which colour appears most in the cluster.

CONFUSION MATRIX		Actual Values		
		Red	Blue	Purple
Predicted Values (Clusters in Fig 1)	Red (Left)	**4**	3	3
	Blue (Middle)	2	**5**	2
	Purple (Right)	2	2	**4**

Overall Accuracy = $\frac{TP+TN}{TP+TN+FP+FN}$ = $\frac{4+5+4}{4+2+2+3+5+2+3+2+4}$ = 0.481

Precision = $\frac{\boldsymbol{TP}}{\boldsymbol{TP+FP}}$

Precision for Left Cluster (Red) = $\frac{4}{4+3+3}$ = 0.4

Precision for Middle Cluster (Blue) = $\frac{5}{5+2+2}$ = 0.555

Precision for Right Cluster (Purple) = $\frac{4}{4+2+2}$ = 0.5

Recall = $\frac{\boldsymbol{TP}}{\boldsymbol{TP+FN}}$

Recall for Left Cluster (Red) = $\frac{4}{4+2+2}$ = 0.5

Recall for Middle Cluster (Blue) = $\frac{5}{5+3+2}$ = 0.5

Recall for Right Cluster (Purple) = $\frac{4}{4+3+2}$ = 0.444

F-Score = $\frac{\boldsymbol{2\ x\ Precision\ x\ Recall}}{\boldsymbol{Precision + Recall}}$

F-Score for Left Cluster (Red) = $\frac{2\ x\ 0.4\ x\ 0.5}{0.4\ +\ 0.5}$ = 0.444

F-Score for Middle Cluster (Blue) = $\frac{2\ x\ 0.555\ x\ 0.5}{0.555\ +\ 0.5}$ = 0.526

F-Score for Right Cluster (Purple) = $\frac{2\ x\ 0.5\ x\ 0.444}{0.5\ +\ 0.444}$ = 0.47

Macro-Averaged Precision = Average of all class precisions = $\frac{0.4+0.555+0.5}{3}$ = 0.485

Macro-Averaged Recall = Average of all class recalls = $\frac{0.5+0.5+0.444}{3}$ = 0.481

Macro-Averaged F-Score = Average of all class F-scores = $\frac{0.444+0.526+0.47}{3}$ = 0.48

	Left Cluster	Middle Cluster	Right Cluster	MACRO AVG
Precision	0.4	0.555	0.5	0.485
Recall	0.5	0.5	0.444	0.481
F-Score	0.444	0.526	0.47	0.48

6. For the same clusters as in Figure 1, compute B-CUBED Precision, Recall, and F-score. (10)

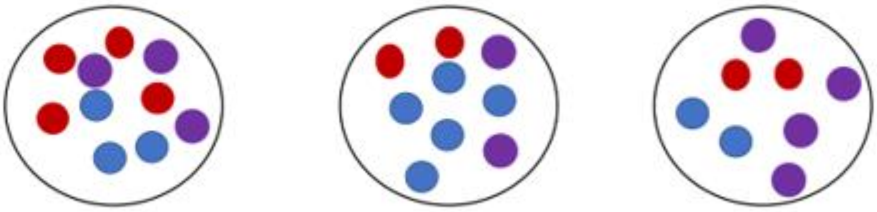

Figure 1: Outcome of a Clustering Algorithm

B-CUBED PRECISION

Precision for a point = $\frac{\text{No. of items in Cluster X belongs to with label X}}{\text{Total No. of items in Cluster X}}$

Cluster 1

Precision for each Red point = 4 / 10 = 0.4

Precision for each Blue point = 3 / 10 = 0.3

Precision for each Purple point = 3 / 10 = 0.3

Cluster 2

Precision for each Red point = 2 / 9 = 0.222

Precision for each Blue point = 5 / 9 = 0.555

Precision for each Purple point = 2 / 9 = 0.222

Cluster 3

Precision for each Red point = 2 / 8 = 0.25

Precision for each Blue point = 2 / 8 = 0.25

Precision for each Purple point = 4 / 8 = 0.5

B-CUBED Precision = Average precision of all points in the dataset =

(0.4 + 0.4 + 0.4 + 0.4 + 0.3 + 0.3 + 0.3 + 0.3 + 0.3 + 0.3 + 0.222 + 0.222 + 0.222 + 0.222 + 0.555 + 0.555 + 0.555 + 0.555 + 0.555 + 0.25 + 0.25 + 0.25 + 0.25 + 0.5 + 0.5 + 0.5 + 0.5) / 27 =

0.3727

B-CUBED RECALL

Recall for a point = $\frac{\text{No. of items in Cluster X belongs to with label X}}{\text{Total No. of items with label X}}$

Cluster 1

Recall for each Red point = 4 / 8 = 0.5

Recall for each Blue point = 3 / 10 = 0.3

Recall for each Purple point = 3 / 9 = 0.333

Cluster 2

Recall for each Red point = 2 / 8 = 0.25

Recall for each Blue point = 5 / 10 = 0.5

Recall for each Purple point = 2 / 9 = 0.222

Cluster 3

Recall for each Red point = 2 / 8 = 0.25

Recall for each Blue point = 2 / 10 = 0.2

Recall for each Purple point = 4 / 9 = 0.444

B-CUBED Recall = Average recall of all points in the dataset =

(0.5 + 0.5 + 0.5 + 0.5 + 0.3 + 0.3 + 0.3 + 0.333 + 0.333 + 0.333 + 0.25 + 0.25 + 0.5 + 0.5 + 0.5 + 0.5 + 0.5 + 0.222 + 0.222 + 0.25 + 0.25 + 0.2 + 0.2 + 0.444 + 0.444 + 0.444 + 0.444) / 27 =

0.371

B-CUBED F-SCORE

F-Score for a point = $\frac{2\ x\ Precision\ for\ point\ x\ Recall\ for\ point}{Precision\ for\ point+Recall\ for\ point}$

Cluster 1

F-Score for each Red point = $\frac{2\ x\ 0.4\ x\ 0.5}{0.4+0.5}$ = 0.444

F-Score for each Blue point = $\frac{2\ x\ 0.3\ x\ 0.3}{0.3+0.3}$ = 0.3

F-Score for each Purple point = $\frac{2\ x\ 0.3\ x\ 0.333}{0.3+0.333}$ = 0.316

Cluster 2

F-Score for each Red point = $\frac{2\ x\ 0.222\ x\ 0.25}{0.222+0.25}$ = 0.235

F-Score for each Blue point = $\frac{2\ x\ 0.555\ x\ 0.5}{0.555+0.5}$ = 0.526

F-Score for each Purple point = $\frac{2\ x\ 0.222\ x\ 0.222}{0.222+0.222}$ = 0.222

Cluster 3

F-Score for each Red point = $\frac{2\ x\ 0.25\ x\ 0.25}{0.25+0.25}$ = 0.25

F-Score for each Blue point = $\frac{2\ x\ 0.25\ x\ 0.2}{0.25+0.2}$ = 0.222

F-Score for each Purple point = $\frac{2\ x\ 0.5\ x\ 0.444}{0.5+0.444}$ = 0.47

B-CUBED F-Score = Average F-Score of all points in the dataset =

(0.444 + 0.444 + 0.444 + 0.444 + 0.3 + 0.3 + 0.3 + 0.316 + 0.316 + 0.316 + 0.235 + 0.235 + 0.526 + 0.526 + 0.526 + 0.526 + 0.526 + 0.222 + 0.222 + 0.25 + 0.25 + 0.222 + 0.222 + 0.47 + 0.47 + 0.47 + 0.47) / 27 =

0.370

B-CUBED PRECISION = 0.3727

B-CUBED RECALL = 0.371

B-CUBED F-SCORE = 0.370