UNIVERSITY OF LIVERPOOL

Data mining and visualisation COMP527

ASSIGNMENT 2: Clustering Algorithms

This report belongs to **VIDISHA** and **STUDENT ID** is **201709173.**

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

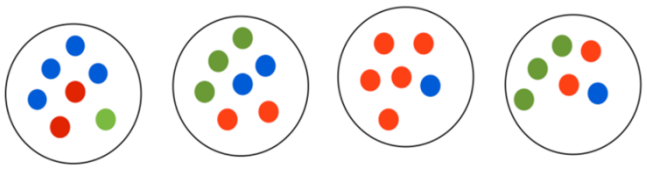

Figure 1: Outcome of a clustring algorithm

Figure 1: Caption

Solution:

Suppose, for the purposes of this discussion, that we have the true labels for every colour in the graph. (Just an assumption as we don't have actual labels)

The true cluster number for each associated data point is indicated in each item of a one-dimensional array that represents the true labels.

For an example- true labels = [0,0,1,1,2,2,3,3,…]

The length of the true labels should match the data points number.

For the given three cluster graphs based on the colour, let's consider zero cluster to all blue dots, one cluster to green dots and two clusters of all red dots.

Let's assign random three possibilities to each data point which will be known as predicted labels= 0,1, or 2

Similarly, assign random true labels to each data point = 0,1 or 2

Confusion matrix: The number of false positives, false negatives, true positives, and true negatives for each class are shown in a square matrix that lets us know the performance of the model and when the model is becoming confused.

[31 28 27]

[46 34 30]

[35 25 44]

Based on the true labels, each row in the matrix represents a true cluster.

Row 1: Supposedly blue dots, or true cluster 0.

Row 2: Supposedly green dots represent True Cluster 1.

Row 3: Supposedly red dots represent True cluster 2.

The predicted clusters:

Column 1: Cluster 0 prediction

Column 2: Cluster 1 prediction

Column 3: Cluster 2 prediction

The confusion matrix that is given reflects a three-cluster classification problem, where the predicted clusters are represented by each column and the actual clusters by each row. There are 31 correctly categorized points for cluster 0, however there are also some misclassified points, with 28 points expected to be in cluster 1 and 27 as cluster 2. 34 of the points in Cluster 1 are properly categorized, whereas a significant portion of the points are incorrectly identified as being in either of the other two clusters 46 as Cluster 0 and 30 as Cluster 2. With 44 accurate classifications, cluster 2 has the highest prediction accuracy. However, there is significant confusion within this cluster, since 25 points are projected to belong to cluster 1 and 35 to cluster 0.

This matrix shows that although the predictions are somewhat accurate, a sizable portion of the points are wrongly assigned to other clusters, indicating that the performance of the clustering algorithm might be enhanced.

Precision: It is the ratio of True positive to the total predicted positive.

Precision = TP/ TP + FP

The precision is 0.37 which means the model is roughly 37% accurate on average when it predicts a data point to be a member of a particular cluster.

Recall: It is the ratio of all observations made in the actual class to all correctly predicted positive observations.

Recall: TP/ TP + FN

It has been calculated as 0.36 which means 36% of the data points are correctly identified by the model in their appropriate clusters on average.

F- Score: It is the Precision and Recall weighted average. It maintains a balance between recall and precision. When there is an uneven distribution of classes, it is extremely helpful because it consider both positive and false scores.

F1 = 2 X (Precision X Recall) / (Precision + Recall)

I got it 0.36 which means the model is neither accurate nor robust in its recall, indicating that the balance between precision and recall tends towards the lower end.

Given that the precision, recall, and F-score values are all significantly below 0.5 the cut-off point for a balanced performance in a binary classification these results imply that the clustering algorithm's performance is not especially great. For an excellent model in a multi-class classification, the values should ideally be closer to 1.0. The

comparatively low scores for all metrics indicate that either the clusters are not well-separated to begin with or the algorithm may have trouble appropriately classifying points into the relevant clusters.

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

Solution:

Assume that the dots' colour indicates the actual class and that the clusters show the clusters that the algorithm predicted. Colours would need to be mapped to class labels, such as red for class 1, blue for class 2, and green for class 3.

B-CUBED Precision: The average "correctness" of the clustering process for every element is represented by this statistic. With a precision of 0.154, which is relatively poor, only 15.4% of the items in a predicted cluster are positioned correctly in relation to their true class on average. In simpler terms, there is a high probability that each given element is part of a cluster that includes objects belonging to distinct classes.

B-CUBED Recall: Recall measures how well the algorithm can put each member of a class in one group. Additionally, the recall of 0.123 is low, indicating that the clustering algorithm frequently fails to include every member of a class in a single cluster. Instead, it frequently distributes pieces of the same class among several clusters.

B-CUBED F-score: The F-score, which offers an individual metric of algorithmic quality for clustering, is the harmonic mean of precision and recall. The F-score of 0.137 indicates subpar performance and confirms limited recall and precision. This score indicates that the algorithm does not do well in recalling all true class members from the cluster or in keeping non-class members out of a cluster (precision).

In summary, these metrics imply that the method for clustering might not be working well for this specific dataset because the resulting clusters do not match the components' true class labels very well.