Data Mining and Visualisation - CA2 - Task 7

Jack Brown 201056294

Figure 1 displays the graph output for the k-Means algorithm on the merged dataset at varying values of k. For all the outputs, results sometimes vary due to the random initialisation of representatives required by the algorithm, which can occasionally result in poor evaluation measures due to the algorithm getting stuck at local minima of the objective function (total sum of distances between objects and representatives) which does not represent the global minimum.

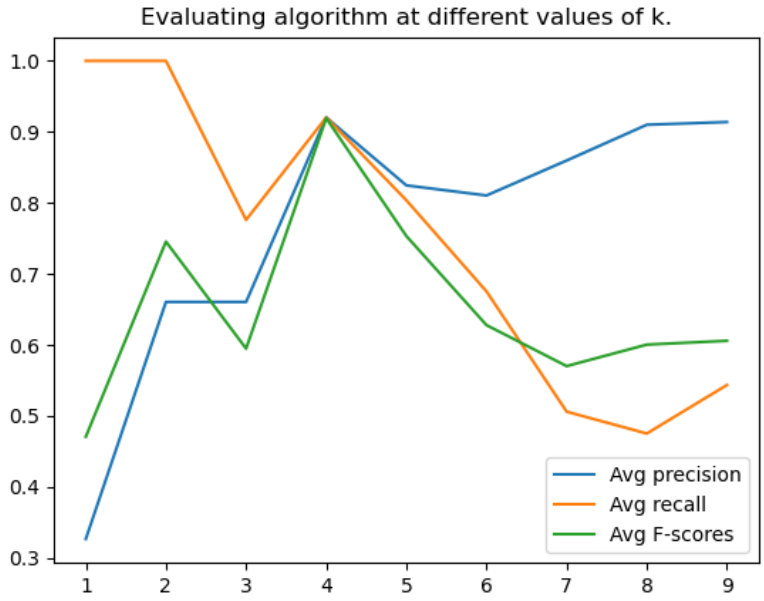

Fig 1 | Representative output for k-Means algorithm on merged dataset

This output represents a relatively good result as measured by the evaluation measures, with average precision, recall and f-scores at the expected k = 4 mark to be all around 0.92.

Figure 2 displays the graph output for the k-Means algorithm on the merged dataset at varying values of k, except the vectors of the dataset have been normalised to unit vector length.

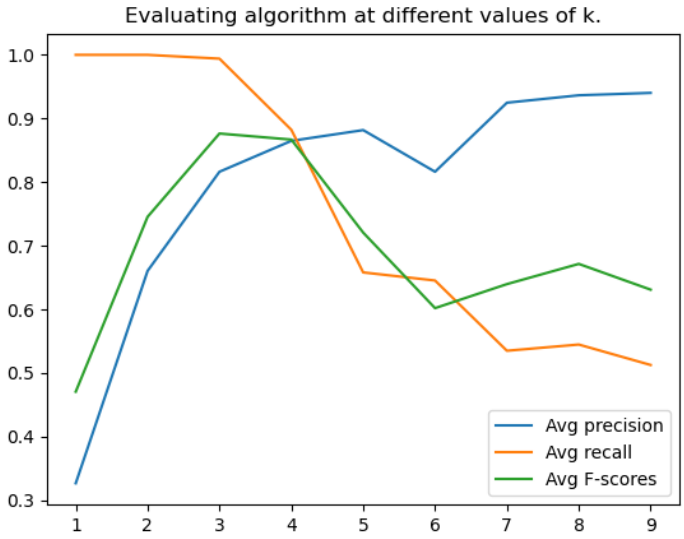

Fig 2 | Representative output for k-Means algorithm on merged dataset, where vectors in the dataset have been normalised to unit length

The output here is similar to the output for the unnormalized dataset, with a slightly lower best performance at k=4, where the measures intersect at roughly 0.87.

Figure 3 displays the graph output for the k-Medoids algorithm on the merged dataset at varying values of k.

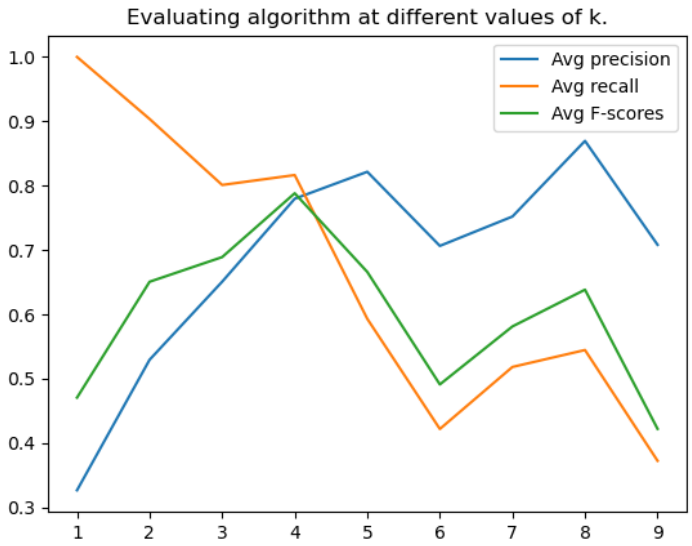

Fig 3 | Representative output for k-Medoids algorithm on merged dataset

Generally the k-Medoids algorithm did not perform as well as the k-Means algorithm, achieving a lower best performance, again at k = 4, with evaluation measures intersecting at approximately 0.80.

Finally, **Fig 4** displays the graph output for the k-Medoids algorithm on the merged dataset at varying values of k, except the vectors of the dataset have been normalised to unit vector length.

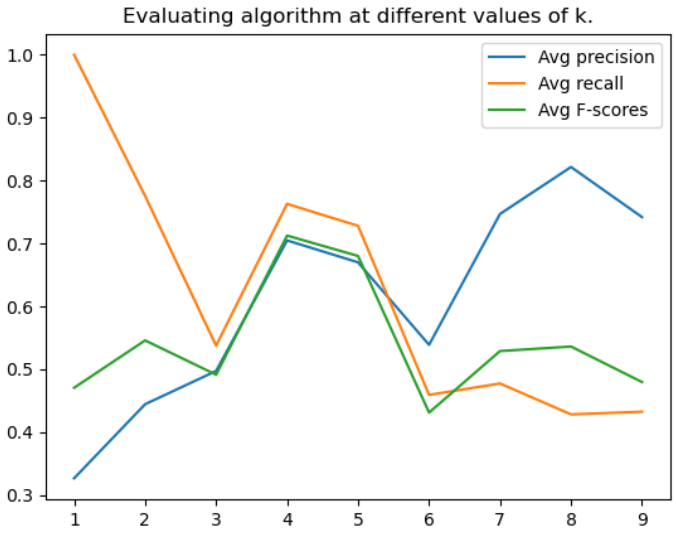

Fig 4 | Representative output for k-Medoids algorithm on merged dataset, where vectors in the dataset have been normalised to unit length

Again the k-Medoids algorithm did not perform as well on the normalised dataset as the k-Means algorithm, achieving a peak performance at k = 4 where evaluation measures are around 0.70.

Given this analysis of the evaluation measures, it appears the implementation of the k-Means algorithm on the unnormalised dataset performed best.