COMP527 Data Mining

Assignment 2: Data Clustering

Implementing the k-means clustering algorithm

Maciej Lechowski

Student ID: 201059287

m.j.lechowski@liverpool.ac.uk

1 Description

Instructions on how to run the program and obtain different results are described in the README.md file.

Source code as well as formatted README file are available at author's github page:

https://github.com/lchsk/k-means

2 Results

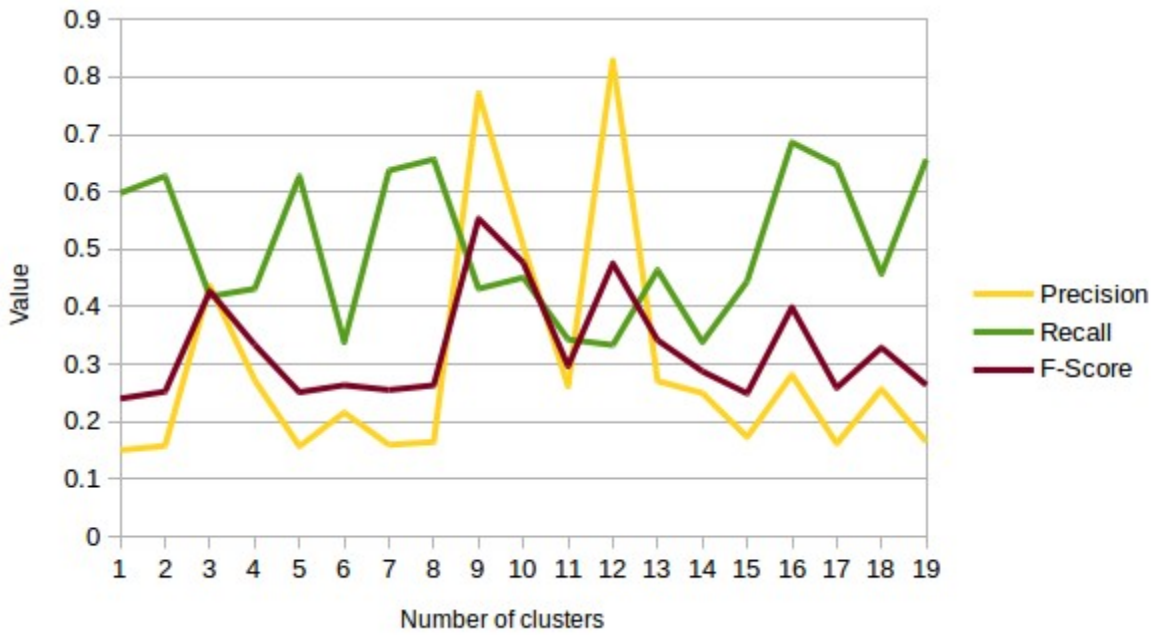

Figure 3.1. This graph was generated using cluster means which are not instances of the cluster. Number of iterations used to generate this graph was 5.

Macro-averaged precision has three peaks: when the number of clusters is 3, 9 or 12. Macro-averaged recall and macro-averaged F-Score vary throughout the experiment with F-Score being in range 0.25 - 0.55 and F-Score 0.35 - 0.7. The conclusion of using this method is that the results seem to be very unstable, changing significantly even when the number of clusters used changes little.

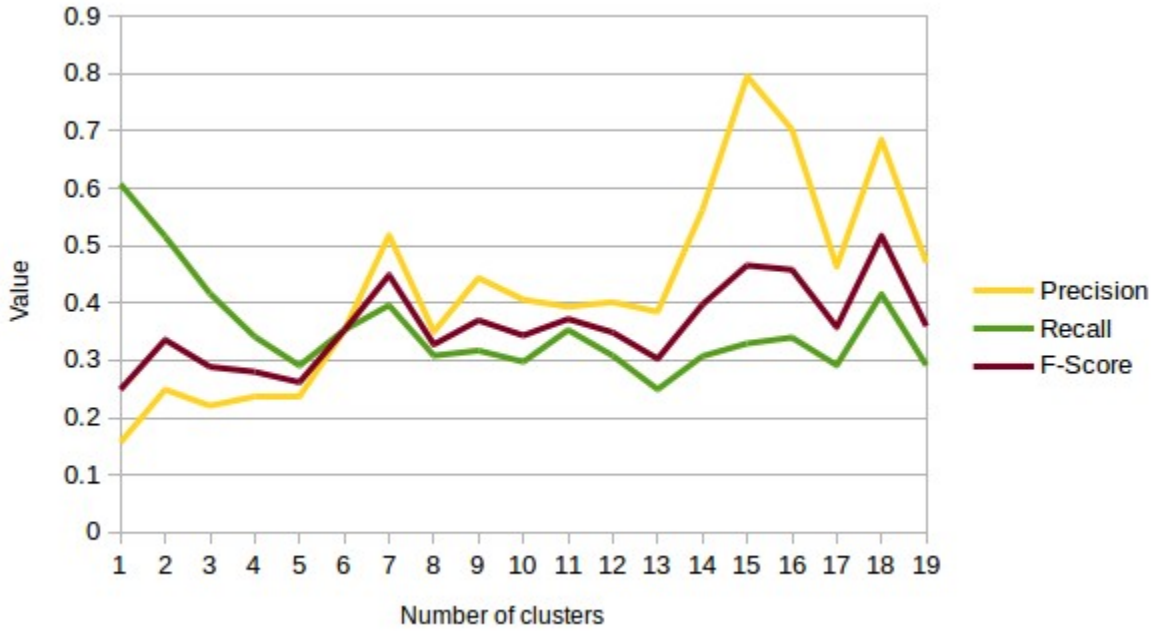

Figure 3.2. This graph was generated using cluster means which are instances of the cluster. Number of iterations used to generate this graph was 3. Two conclusions can be reached from this experiment. First, that it takes significantly more time to obtain results when cluster centres are instances of the clusters. That is because of additional computation involved in calculating distances between all instances of each cluster, which proves to be a heavy task. Second, looking at the graph, results in this case seem more stable with precision and F-Score growing as the number of clusters is increased. Precision reaches its highest value with 15 clusters and recall with 18.