Practical 4: Clustering

Prof. M- Tahar Kechadi

School of Computer Science & Informatics
University College Dublin

The aim of this practical is to use RapidMiner to generate clusters from data sets using clustering algorithms. The data sets to be used can be found on Blackboard.

The deliverables for this practical are:

- A (results) document (e.g. a Word document) containing the clusters (Text View, Centroid Table for Cluster Model and Data View, Plot View (Pie or Scatter) for ExampleSet) generated by RapidMiner, in a section per question. Any discussion should also be placed in this document.
- For each question, the RapidMiner process file should be exported and submitted.

All files generated should be placed in a zip file with the name <studentname> _<studentnumber>_ comp40370 practical4.zip, and submitted via blackboard.

Question 1 Creating clusters with K-Means (I)

Using the P4_Q2.xls dataset, generate a Rapid Miner process that does the following:

1. Using all of the attributes, run the k-means algorithm with k=3 (default parameters i.e. select *add cluster attribute*, max runs = 10, *max optimization* step = 100).

2. Switch to Plot View of ExampleSet, Select Scatter for Plotter, X for x-Axis, Y for y-Axis and cluster for Colour Column. Giving the discussion on the clustering results.

   The results' document should be submitted, along with the RapidMiner process xml file (File - Export Process menu option), as explained at the start of this document.

Question 2 Creating clusters with K-Means (II)

Using the cereals.xls data set, generate a process that does the following:

1. Using all variable attributes except *NAME, MANUF, TYPE* and *RATING*, run the k-means algorithm with k= 5 (select *add cluster attribute*, max runs = 5, *max optimization step* = 100) to identify clusters within the data.

2. Change *max optimization step* to 1000 and rerun the process. Are these results similar to those obtained in Question 2.1? Explain your answer.

3. Rerun the k-means algorithm with *k = 3*.

4. Which clustering solution is better? why?

   The results' document should be submitted, along with the RapidMiner process xml file (File - Export Process menu option), as explained at the start of this document.

Question 3 Comparing K-Means with DBSCAN

Using the local.xls data set, generate processes that do the following:

1. Using all the variable attributes except *ID*, run the k-means algorithm with k = 7 (select *add cluster attribute*, max *runs* = 5, *max optimization step* = 100) to identify clusters within the data.

2. Switch to Plot View of ExampleSet, Select Scatter for Plotter, X for x-Axis, Y for y-Axis and cluster for Color Column. Discuss the clustering results.

3. Using all the variable attributes except *ID*, normalise the X and Y attributes to the range [0.0,1.0], run the DBSCAN algorithm with epsilon = 0.04, min points = 4 (*select add cluster attribute*, *measure types* = NumericalMesures, *numerical measure* = EuclideanDistance) to identify clusters within the data.

4. Switch to Plot View of ExampleSet, Select Scatter for Plotter, X for x-Axis, Y for y-Axis and cluster for Color Column. Discuss the clustering results and compare them to the results of Question 3.2.

5. Rerun the DBSCAN process (Question 3.3) with *epsilon* = 0.08. Discuss the clustering results.

   The results' document should be submitted, along with the RapidMiner process xml file (File - *Export Process* menu option), as explained at the beginning of this document.