CMPT 459: Assignment 2 – DBSCAN

This report is based on a subset of the whole dataset - January dataset – as suggested in one of the Coursys news.

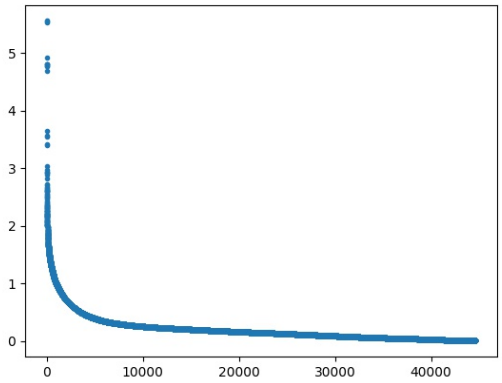
Preprocessing:

- Missing values where imputed using pandas library function *df.replace()* which imputes missing values based on surrounding data records.
- The data is normalized using *StandardScaler* from *sklearn.preprocessing*. *StandardScaler* is used because it uses z-score method which is not sensitive to outliers.
- The date and time column is dropped before normalization.

Implementation of DBSCAN:

- There are 3 functions that is used to implement the algorithm: *fit(), density_reachable(), core_object()*
- *fit()* is the main function of the DBSCAN implementation which calls *density_reachable()* on objects that are not assigned yet and if the object is a core object. *fit()* returns a list of cluster labels and save the labeled data in a csv file.
- Given a core object, *density_reachable()* assigns a cluster label to the core object, its neighbors in the epsilon neighborhood, and all other density reachable objects. The function is not recursively implemented because from a core object, there can be so many density reachable object that the stack memory will be full. Hence, an iterative approach is taken to implement this function.
- Given an object, *core_object()* return True or False if an object is a core object.

Heuristic approach to choose epsilon and MinPts: k-distance diagram



- For each object, k nearest neighbors is found and the furthest distance to the neighbors is recorded and then plotted. The default value of k is (2*dimensions of dataset -1), i.e. k=13. From the k-distance-diagram, a

threshold value of epsilon is chosen, E=0.6. A value of (2*dimensions of dataset) is set as the default value for MinPts, i.e. MinPts=14.

- Statistics:
    - Number of clusters: 20
    - Objects per cluster: {c0: 24869, c1: 16501, c2: 72, c3: 40, c4: 345, c5: 36, c6: 32, c7: 46, c8: 26, c9: 72, c10: 461, c11: 28, c12: 24, c13: 34, c14: 38, c15: 38, c16: 29, c17: 56, c18: 19, c19: 31, c20: 43}
    - Noise/Unassigned: 2597
- Quality of clustering is determined by the ratio of Intra cluster distance to Inter cluster distance.
    - Average Intra cluster distance: 0.106
    - Average Inter cluster distance: 0.905
    - Intra cluster to Inter cluster ratio: 0.117
    - The clustering quality is good because the ratio is low.