Exercise Sheet 5

(until 17.05.2019)

1. Use DBSCAN to find clusters from nearby large cities around the world (metropolitan areas). A city with at least 50,000 inhabitants is considered large. The *e*-neighborhood of a city contains all adjacent cities with a Euclidean distance of 0.15 or less in latitude and longitude. A city is considered a core object of a conurbation if at least 8 cities are located in its *e*-neighborhood. For clustering, use the *maps::world.cities* dataset. Answer the following questions:
    a. How many clusters, core objects, border objects and noise objects are found by DBSCAN?
    b. How many cities does the largest cluster contain and in which country are the cities of the largest cluster located?
    c. Which three countries have the most cities in clusters?
    d. Are the Indian cities Rajendranagar and Rajpur (directly) density-reachable or density-connected?
    e. Are Essen and Castrop-Rauxel (directly) density-reachable or density-connected?
    f. Which cities are density-reachable from Bochum, but not directly density-reachable?

2. Given again be the dataset from task 2 of task sheet 3. This time use DBSCAN with *minPts = 6* for clustering. First determine a suitable value for *e*. Display the clustering in a scatter plot. Highlight cluster assignments and noise points in color. Compare and discuss the clustering of DBSCAN with the clustering of k-Means.
3. Given again, be the dataset from task 2 of task sheet 3. Use OPTICS to create a density reachability diagram for *minPts =6*. Extract a clustering for each *reachability-dist = {1,1.5,...,5}* and display the result in a scatter plot, respectively. Highlight cluster assignments and noise points in color. Evaluate the change of the clustering result with increasing threshold for *reachability-dist* regarding the number of clusters as well as the number of core, border, and noise points.
4. Using the example of the silhouette coefficient, discuss the strengths and weaknesses of internal quality measures. Why are they only conditionally suitable for the comparison between clusterings of different algorithms (e.g. K-Means and DBSCAN)? In which cases should they still be used?

Dataset for task 2 and 3: http://isgwww.cs.uni-magdeburg.de/cv/lehre/VisAnalytics/material/exercise/datasets/clustering-student-mat.csv