Ludwig-Maximilians-Universität München
Institut für Informatik
Prof. Dr. Thomas Seidl
Julian Busch, Evgeniy Faerman,
Florian Richter, Klaus Schmid

Knowledge Discovery in Databases
SS 2016

Exercise 8: Clustering, Outlier Detection

Regarding tutorials on 15.06.-17.06.2016.

Exercise 8-1 DBSCAN

Let $\mathcal{C} = \{C_1, C_2\}$ be a clustering result of DBSCAN. Prove or disprove:

(a) a is a core object in both C_1 and in $C_2 \implies C_1 = C_2$.

(b) a is a border object in both C_1 and in $C_2 \implies |C_1 \cap C_2| > min\_pts$

(c) $a \in C_1$ is a core object $\implies \bigcup_{x \in N_\epsilon(a)} N_\epsilon(x) = C_1$.

(d) $a \in C_1$ is a border object $\implies$ there is at least one core object $b \in N_\epsilon(a)$.

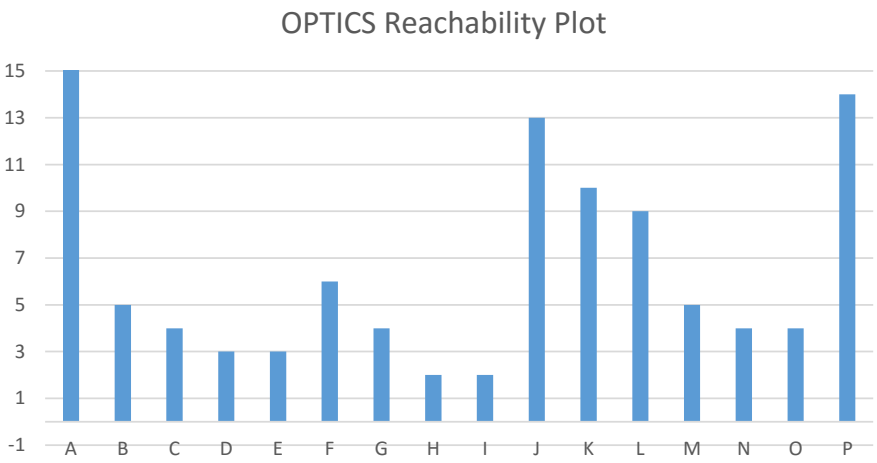
(e) Let $C' = \{x \in C_1 | N_\epsilon(x) < min\_pts\}$ be the set of all border objects in C_1
$\implies$ there exists a density-connected pair $(x, y) \in C'$.

(f) C_1 has more border objects than core objects.

(g) C_1 has more core objects than border objects.

Exercise 8-2 OPTICS

Given the following OPTICS reachability plot as a result, answer the following questions:



(a) Which clustering can be obtained with $\epsilon = 11$?

(b) Again using $\epsilon = 11$, which objects are outlier?

(c) What about $\epsilon = 4$?

(d) For which ϵ is P not an outlier?

(e) Which ϵ has to be used to form the cluster $\{GHI\}$?

(f) Is it possible to choose an ϵ, such that there is a one-element-cluster?

(g) How many different cluster partitions can be found in this dataset?

Exercise 8-3 Evaluation of Clustering Results

Given a dataset DB, a clustering $\mathcal{C} = \{C_1, \ldots, C_k\}$ and a ground truth $\mathcal{G} = \{G_1, \ldots, G_l\}$ consisting of a set of classes, we consider all pairs of objects $(o_i, o_j) \in DB \times DB$ and construct a confusion matrix as follows:

		clustering result	
		same cluster	**different clusters**
ground truth	**same class**	True Positives (TP)	False Negatives (FN)
	different classes	False Positives (FP)	True Negatives (TN)

For instance, the number of true positives corresponds to the number of pairs, which appear in a common cluster and belong to the same class:

$$TP = |\{(o_i, o_j) \in DB \times DB \mid o_i \neq o_j \wedge \exists C \in \mathcal{C} : o_i, o_j \in C \wedge \exists G \in \mathcal{G} : o_i, o_j \in G\}|$$

The remaining entries are defined analogously. Based on the confusion matrix, we can define the following quality measures:

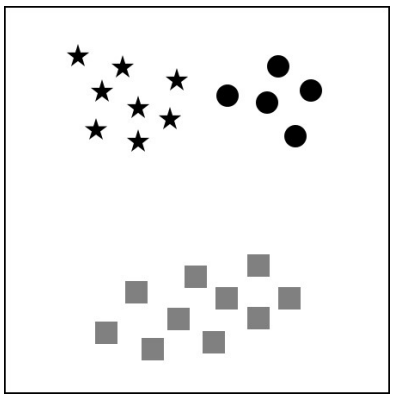
$$RI = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1 = \frac{2 \cdot Recall \cdot Precision}{Recall + Precision}$$

Consider the following dataset, where the classes are represented by different colors and the clusters are indicated by the object shapes:

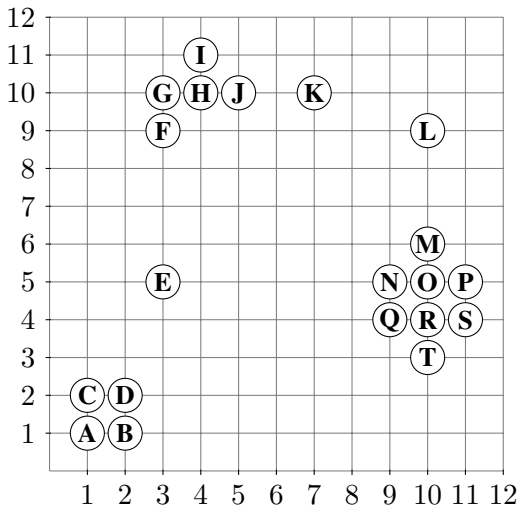


(a) Construct the confusion matrix and compute the Rand Index, Precision, Recall and F1-Measure.

(b) How would you evaluate the quality of the clustering result as an 'expert'? Does your assessment correspond with the external quality measures? In general, can you identify some potential problems or drawbacks of external cluster evaluation measures?

Exercise 8-4 Outlier Scores

Given the following 2 dimensional data set:



As distance function, use Manhattan distance $L_1(a, b) := |a_1 - b_1| + |a_2 - b_2|$.

Compute the following (without including the query point when determining the kNN):

- LOF using $k = 2$ for the points E, K and O.
- LOF using $k = 4$ for the points E, K and O.
- kNN distance using $k = 2$ for all points.
- kNN distance using $k = 4$ for all points.
- aggregated kNN distances for $k = 2$ and $k = 2$ for all points
  (aggregated kNN distance = sum of the distances to all the kNN!)