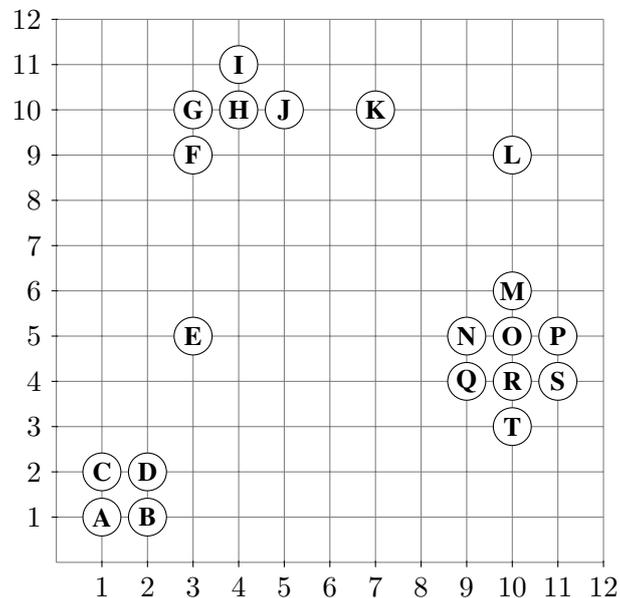


Knowledge Discovery in Databases  
 SS 2014

Übungsblatt 5: Clusteranalyse – Single-Link und OPTICS

**Aufgabe 5-1 Single-Link Hierarchical Clustering**

Gegeben sei der folgende Datensatz:



Als Distanzfunktion zwischen den Punkten verwenden Sie die Manhattan-Distanz ( $L_1$ -Norm):

$$L_1(x, y) = |x_1 - y_1| + |x_2 - y_2|$$

Berechnen Sie zwei Dendrogramme für diesen Datensatz. Als Distanzfunktion zwischen Mengen von Objekten verwenden Sie

- (a) den Single-Link Ansatz,
- (b) den Average-Link Ansatz.

Tipp: Innere Knoten müssen nicht binär sein, d.h. sie können mehr als zwei Söhne haben.

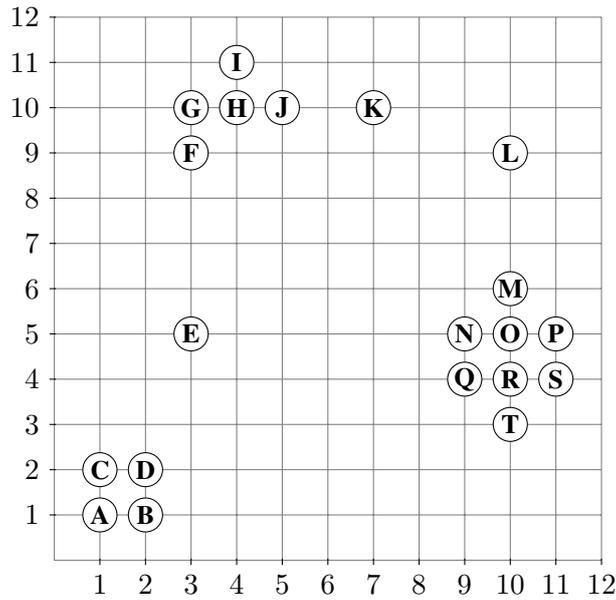
**Aufgabe 5-2 Zusammenhang DBSCAN/OPTICS und Single-Link**

Was ist der Zusammenhang von DBSCAN bei  $minPts = 2$  zu single-linkage Clustering?

Warum läuft DBSCAN in  $\mathcal{O}(n^2)$  Zeitkomplexität (mit Index typischerweise sogar  $\mathcal{O}(n \log n)$ ), während hierarchische Clusteranalyse auf Distanzmatrizen als  $\mathcal{O}(n^3)$  angegeben wird, und SLINK in  $\mathcal{O}(n^2)$  läuft?

Warum ist das kein Widerspruch?

### Aufgabe 5-3 OPTICS



Als Distanzfunktion verwenden Sie die Manhattan-Distanz  $L_1(a, b) := |a_1 - b_1| + |a_2 - b_2|$ .

Konstruieren Sie ein Erreichbarkeitsdiagramm mit dem Algorithmus OPTICS (siehe beiliegenden Pseudocode) für folgende Parameter:

- $\epsilon = 5$  and  $minPts = 2$
- $\epsilon = 5$  and  $minPts = 4$
- $\epsilon = 2$  and  $minPts = 4$
- $\epsilon = \infty$  and  $minPts = 4$

#### Pseudocode OPTICS

```

seedlist =  $\emptyset$  // implemented as a heap
for  $i = 0$  to  $n-1$  do
    if( $seedlist = \emptyset$ ) then  $seedlist = \{(random\_not\_handled\_point, \infty)\}$ 
    ( $x, x.reach$ ) = get_and_remove_point_with_min_reach( $seedlist$ )
     $x.pos = i$ 
     $x.handled = TRUE$ 
     $neighbors = rangeQuery(x, \epsilon)$ 
     $x.core = nnDist(x, neighbors, MinPts)$ 
    if( $x.core < \infty$ )
        for each  $y \in neighbors$  with not( $y.handled$ )
            if ( $y \notin seedlist$ )  $seedlist = seedlist \cup \{(y, reach-dist(y,x))\}$ 
            else
                 $curr\_reach = lookup(seedlist, y)$ 
                update( $y, \min(curr\_reach, reach-dist(y,x))$ )
        endfor
    endfor
endfor

```