

5. Übung „Knowledge Discovery“

Wintersemester 2008/2009

1 DBSCAN

Das dichte-basierte Clustermodell von DBSCAN definiert Kernpunkt und Randpunkte eines Clusters. Randpunkte sind Punkte, die zu einem Cluster gehören, weil sie dichte-erreichbar von Kernpunkten sind, aber selbst keine Kernpunkte sind. Wie der Name schon sagt, sind das Punkte, die am Rand eines Clusters liegen.

1. Kann es Randpunkte geben, die gleichzeitig zu verschiedenen Clustern gehören?

Ja, es kann laut Definition Randpunkte geben, die zu mehr als einem Cluster gehören.

2. Wie geht DBSCAN mit solchen Randpunkten um, d. h. welchem Cluster werden diese Punkte zugeordnet?

DBSCAN ordnet Randpunkte immer dem ersten gefundenen Cluster zu.

Nennen Sie eine bessere Lösung zur Zuordnung dieser Randpunkte. Begründen Sie Ihre Entscheidung.

z. B.

- Randpunkt dem Cluster zuordnen, dessen Abstand zum nächsten Punkt im Cluster am geringsten ist.
- Randpunkt mehrfach zuordnen.

3. Es sei eine Menge von Punkten gegeben. Es sollen Cluster mittels DBSCAN mit den Parametern $minPts = 4$ und $\varepsilon = 10$ gebildet werden. Die folgende Tabelle gibt die Abstände zwischen den Punkten an:

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11
P1	0	54,1	5,0	8,2	13,0	62,5	69,1	59,4	65,3	49,7	55,9
P2	54,1	0	49,1	46,4	43,9	10,4	15,0	12,4	13,0	7,3	11,7
P3	5,0	49,1	0	4,1	9,5	57,5	64,1	54,5	60,4	44,7	50,9
P4	8,2	46,4	4,1	0	5,4	55,2	61,4	52,5	57,3	42,4	49,0
P5	13,0	43,9	9,5	5,4	0	53,2	58,8	51,0	54,1	40,6	47,7
P6	62,5	10,4	57,5	55,2	53,2	0	9,1	6,3	14,1	12,8	9,0
P7	69,1	15,0	64,1	61,4	58,8	9,1	0	15,3	9,1	20,2	18,0
P8	59,4	12,4	54,5	52,5	51,0	6,3	15,3	0	20,0	10,8	3,6
P9	65,3	13,0	60,4	57,3	54,1	14,1	9,1	20,0	0	20,1	21,5
P10	49,7	7,3	44,7	42,4	40,6	12,8	20,2	10,8	20,1	0	8,1
P11	55,9	11,7	50,9	49,0	47,7	9,0	18,0	3,6	21,5	8,1	0

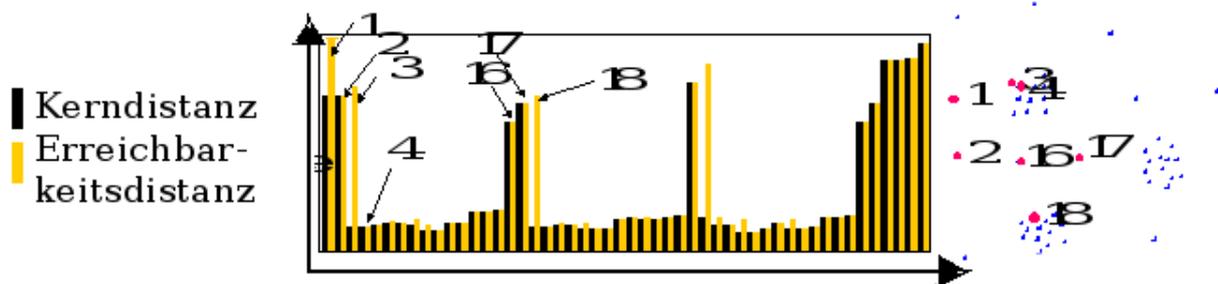
Bilden Sie die Cluster. Numerieren Sie diese Cluster und tragen Sie in eine Tabelle für jeden Punkt die Punkt-Art (Kernobjekt, Randobjekt, Rauschen) sowie die Nummer des Clusters, dem der Punkt angehört, ein.

Punkt Nr.	Art	Cluster
P1	Rand	2
P2	Rauschen	
P3	Kern	2
P4	Kern	2
P5	Rand	2
P6	Kern	1
P7	Rand	1
P8	Rand	1
P9	Rauschen	
P10	Rand	1
P11	Kern	1

In welchem Wertebereich müsste ε liegen (bei $MinPts = 4$), damit alle Punkte als „Rauschen“ identifiziert werden? Begründen Sie ihre Antwort!

Es müsste ε zwischen 0 und 8,2 (exklusive der Obergrenze von 8,2) liegen. Denn kein Objekt darf Kernobjekt sein, d. h. ε muss so gewählt sein, dass das viertnächste Objekt nicht mehr in der ε -Umgebung liegt. Das höchstmögliche ε muss daher kleiner sein als der kleinste Abstand zum viertnächsten Objekt. Dieser kleinste Abstand ist bei Punkt P4 zu finden (8,2).

2 OPTICS



Es sei OPTICS auf eine Datenbank mit den Parametern ε und $MinPts$ angewendet worden. Geben Sie ein Verfahren an, mit dem man aus dem Resultat des OPTICS-Laufes (Clusterordnung, Erreichbarkeitsdiagramm und Kerndistanzdiagramm) das DBSCAN-Clustering für ein gegebenes $\varepsilon' \leq \varepsilon$ extrahieren kann. Benutzen Sie möglichst intuitiven Pseudocode.

```

ExtractDBSCANClustering (clusterOrderedObjs,  $\varepsilon'$ , minPts)
// Precondition:  $\varepsilon' \leq$  generating dist  $\varepsilon$  for clusterOrderedObjs
clusterId := NOISE
FOR  $i$  FROM 1 TO clusterOrderedObjs.size DO
  object := clusterOrderedObjs.get( $i$ )
  IF object.reachabilityDistance >  $\varepsilon'$  THEN
    // UNDEFINED >  $\varepsilon$ 
    IF object.coreDistance  $\leq \varepsilon'$  THEN
      clusterId := nextId(clusterId)
      object.clusterId := clusterId
    ELSE
      object.clusterId := NOISE
  ELSE
    // object.reachabilityDistance  $\leq \varepsilon'$ 
    object.clusterId := clusterId
END // ExtractDBSCANClustering
  
```

Kann aus dem OPTICS-Ergebnis eine eindeutige Clusterzugehörigkeit abgeleitet werden, die DBSCAN bzgl. des gegebenen $\varepsilon' \leq \varepsilon$ erzeugen würde? Mit anderen Worten, stimmt das Ergebnis ihres Verfahrens exakt mit dem Ergebnis eines DBSCAN-Laufes bzgl. ε' überein? Begründen Sie Ihre Antwort!

Ja, das Ergebnis stimmt überein.

3 Hierarchische Clusterverfahren

1. Entwickeln Sie einen Algorithmus zum Update einer Distanz-Matrix D , die die Abstände zwischen Clustern beim Single-Link-Verfahren speichert.

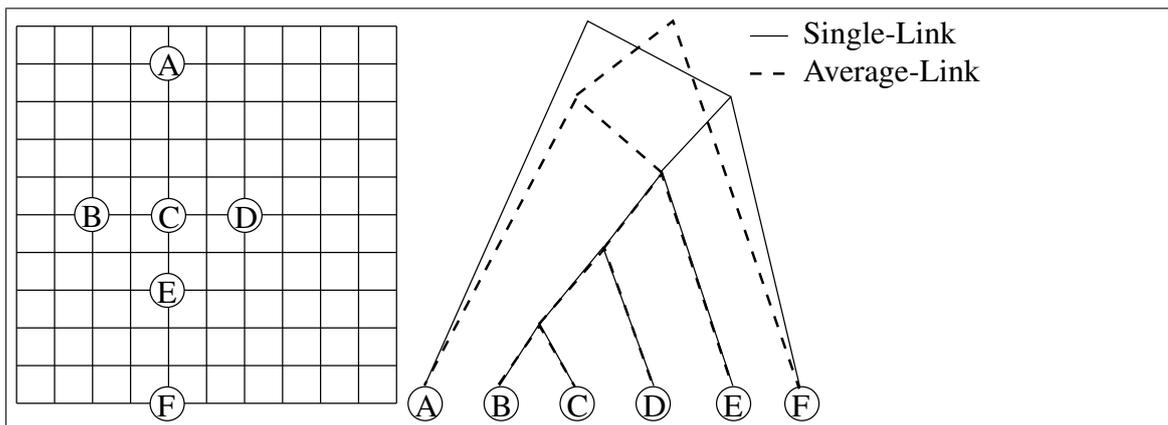
- $C^1 = \{\{X_1\}, \dots, \{X_m\}\}; k = 1$
- $D^1(i, j) = \text{dist}(\{X_i\}, \{X_j\}) = \text{dist}(X_i, X_j)$
- while $|C^k| \geq 2$ loop
 - $(q, s) = \text{argmin } D^k(i, j)$
 - $C^{k+1} = (C^k \setminus \{C_q^k, C_s^k\}) \cup \{C_q^k \cup C_s^k\}$
 - $D^{k+1}(i, j) = D^k(i, j)$ for $i \neq q, j \neq q$
 - $D^{k+1}(q, j) = f(D^k(q, j), D^k(s, j))$
 - $D^{k+1}(i, q) = f(D^k(i, q), D^k(i, s))$
 - delete $D^k(i, s), D^k(s, j)$
 - inc(k)
- result: C^1, \dots, C^k

Dabei ist $f = \min$ für Single-Link und $f = \max$ für Complete-Link.

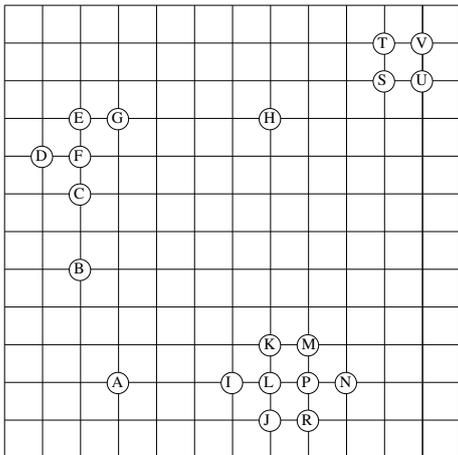
2. Wie muß die Funktion zum Aktualisieren der zusammgelegten Matrixelemente aussehen, wenn statt des Single-Link-Verfahrens das Average-Link-Verfahren verwendet werden soll?

$$f(D^k(q, j), D^k(s, j), |C_q^k|, |C_s^k|) = \frac{D^k(q, j)|C_q^k| + D^k(s, j)|C_s^k|}{|C_q^k| + |C_s^k|}$$

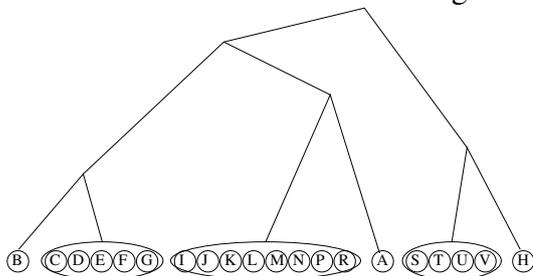
3. Überlegen Sie sich ein (möglichst kleines) Beispiel, in dem die Single-Link-Methode ein anderes Dendrogramm liefert als die Average-Link-Methode.



4. Berechnen Sie für den folgenden Datensatz jeweils ein Dendrogramm für den Single-Link-Ansatz und für den Average-Link-Ansatz. Nutzen Sie als Distanzfunktion zwischen Punkten die Manhattan-Metrik.



Skizze eines vereinfachten Dendrogramms (kann durch beide Verfahren erzeugt werden):



Genau genommen muß am Dendrogramm der vertikale Abstand der Knoten von den Blättern der Distanz zwischen den Clustern entsprechen. Dabei kann es vorkommen, daß Knoten nicht nur binär sind. Trägt man diese Distanzen exakt ein, unterscheiden sich die (vereinfachten) Dendrogramme der beiden Ansätze.