A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise

Ester, Kriegel, Sander, Xu 1996

Clustering in Spatial Databases

- Weltraum- / Geodatenbanken
- Weltraum- / Geowissenschaften
- Extrem großes Datenaufkommen
- Wissensentdeckung in Datenbanken

(knowledge discovery in databases)

Probleme

- Fehlendes Wissen über die Domäne
- Gruppenfindung unabhängig von Form (konvex, konkav, ...)
- Effizienz bei großen Datenbanken

Existierende Lösungen

- Es existieren zwei Verfahren:
 - Hierarchisch
 - Partitionierend
- Jedoch nur CLARANS für KDD geeignet

Lösung

- DBSCAN funktioniert dichtebasiert
- Ausgelegt für große Datenbanken
- Nur zwei Parameter: ε und minPts
- $O(n \cdot \log(n))$ (Durchschnitt)

DBSCAN-Algorithmus

```
DBSCAN(D, eps, MinPts)
  C = 0
   for each unvisited point P in dataset D
      mark P as visited
      N = getNeighbors (P, eps)
     if sizeof(N) < MinPts
         mark P as NOISE
      else
         C = next cluster
         expandCluster(P, N, C, eps, MinPts)
expandCluster(P, N, C, eps, MinPts)
   add P to cluster C
   for each point P' in N
     if P' is not visited
         mark P' as visited
        N' = getNeighbors(P', eps)
         if sizeof(N') >= MinPts
            N = N joined with N'
     if P' is not yet member of any cluster
         add P' to cluster C
```

Ergebnisse

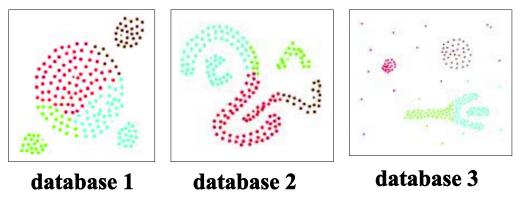


figure 5: Clusterings discovered by CLARANS

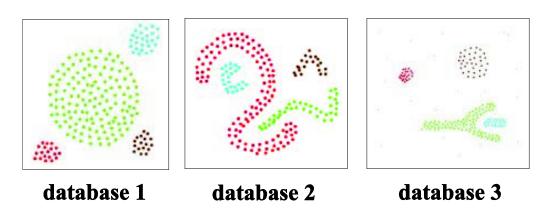


figure 6: Clusterings discovered by DBSCAN

Ergebnisse

Table 1: run time in seconds

number of points	1252	2503	3910	5213	6256
DBSCAN	3.1	6.7	11.3	16.0	17.8
CLAR- ANS	758	3026	6845	11745	18029
number of points	7820	8937	10426	12512	
DBSCAN	24.5	28.2	32.7	41.7	
CLAR- ANS	29826	39265	60540	80638	

Checklist

- Stimmt das Resultat?
- Erkenntnisgewinn?
- Neue Ideen?
- Problem wichtig?
- Ergebnis relevant?

Kritik

- Mathematische Formeln
- Pseudocode
- Schwächen fehlen