Clustering

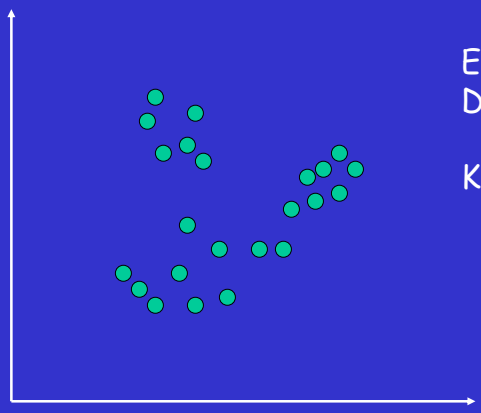
Bioinformatik 5. Semester

Werner Backfrieder

Einführung

- Basis mehrdimensionale Daten
- Gruppierung der Daten
  - Inhärente Merkmale
  - Abstände (Metrik)
- Hierarchische Methoden
  - Stammbäume
- Partitionierende Methoden
  - NN, k-means
- Kein „bester" Algorithmus (Trial & Error)

Problemstellung



Einteilung der Datenpunkte in Gruppen

Kriterium: Abstand

Backfrieder-Hagenberg

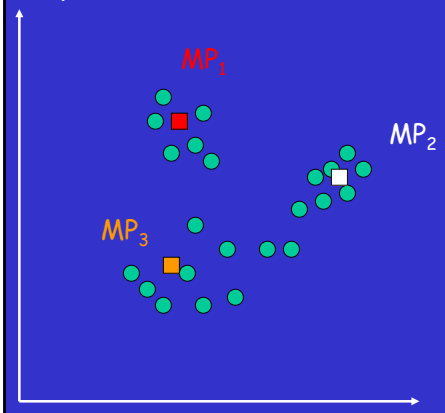
Nächster-Nachbar (NN) Clustering

- Anzahl der Cluster vorgegeben
- Mittelpunkte definieren
  - # Mittelpunkte = # Cluster
- Punkt -> Cluster definiert durch nächsten Mittelpunkt
- Vor- und Nachteile
- (+) Schnelles Verfahren
- (-) Starke Abhängigkeit von vordefinierten Mittelpunkten
- (-) Starre Klassifizierung

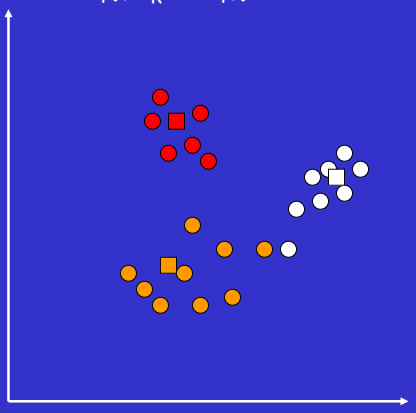
Backfrieder-Hagenberg

NN-Clustering

Definieren von 3 Mittel-punkten MP_i



Klassifizierung: $\min_i(|x_k - MP_i|)$



Backfrieder-Hagenberg

k-Means Clustering

- Definition # Cluster
- Bedingungen:
  - Jeder Cluster enthält mindestens 1 Element
  - Jedes Element muß einem Cluster zugeordnet sein
- Cluster durch Zentroiden beschrieben

Backfrieder-Hagenberg

Klassifizierung

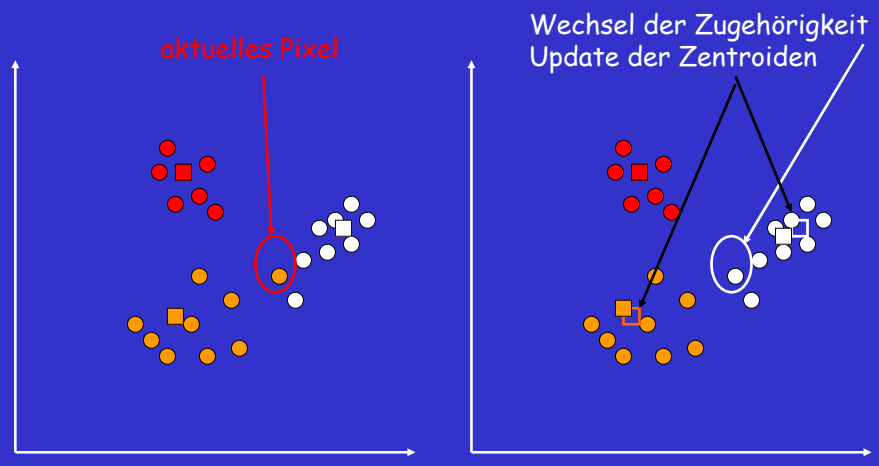
- Dichte eines Clusters
  - Summe der Abstände aller Punkte des Clusters vom Zentroiden
- Gesamtdichte
  - Summe über alle Cluster

k-means -> Gesamtdichte minimal

Algorithmus

- Auswahl von k Zentroiden
- A-priori Klassifizierung in k-Cluster (NN)
- Neuberechnung der Zentroiden
- Testen aller Objekte ob anderer Zentroid näher ist.
  - Wechsel des Clusters, Update der beiden Centroiden
- Abbruchkriterium für minimale Änderungen

Cluster-Update



Backfrieder-Hagenberg

Diskussion

- Schneller, stabiler Algorithmus
- Gutes Konvergenzverhalten
- Effizienter Update der Zentroiden
- Keine „leeren" Cluster können entstehen
- Probleme nur wenn Cluster a-priori leer ist!

Backfrieder-Hagenberg