CLUSTERING

- Clustering = gruparea inregistarilor (cazurilor) in clase de obiecte similare
- Cluster = o colectie de obiecte care sunt similare intre ele si nesimilare cu obiecte din alte clase
- In clustering nu exista o variabila target care trebuie clasificata (spre deosebire de clasificare)
- Clusteringul nu cauta sa clasifice, sa estimeze sau sa faca o predictie cu privire la valoarea unei variabile target, ci cauta sa segmenteze intregul set de date in subgrupuri relativ omogene
- similaritatea in interiorul clusterilor trebuie maximizata, in timp ce similaritatea cu obiecte din alti clusteri trebuie minimizata

- Trebuie sa determinam:
  - Cum masuram similaritatea
  - Cum codificam variabilele categoriale
  - Cum normalizam sau standardizam variabilele numerice

- Pentru a masura similaritatea intre valori numerice putem folosi distanta euclidiana, cityblock, Minkovsky
- Pentru valori categoriale putem folosi functia “diferit de”
- Pentru normalizare se poate folosi normalizarea min-max, standardizarea z-score

Clustering ierarhic

- Se creeaza o structura de arbore
  - Metode divizive: prin partitionarea recursiva
    - La inceput, toate obictele apartin unui singur cluster
    - Cele mai disimilare obiecte sunt separate
  - Metode aglomerative: prin combinarea clusterilor existenti
    - La inceput, fiecare obiect reprezinta un cluster
    - Apoi, cei mai apropiati 2 clusteri vor fi combinati intr-un nou cluster
- Este usor sa calculam distanta dintre doua obiecte, dar cum calculam distanta dintre doua grupuri de obiecte?
  - Single linkage – ia in considerare distanta minima intre oricare doua obiecte din cei doi clusteri
  - Complete linkage - ia in considerare distanta maxima intre oricare doua obiecte din cei doi clusteri
  - Average linkage – se considera distanta medie

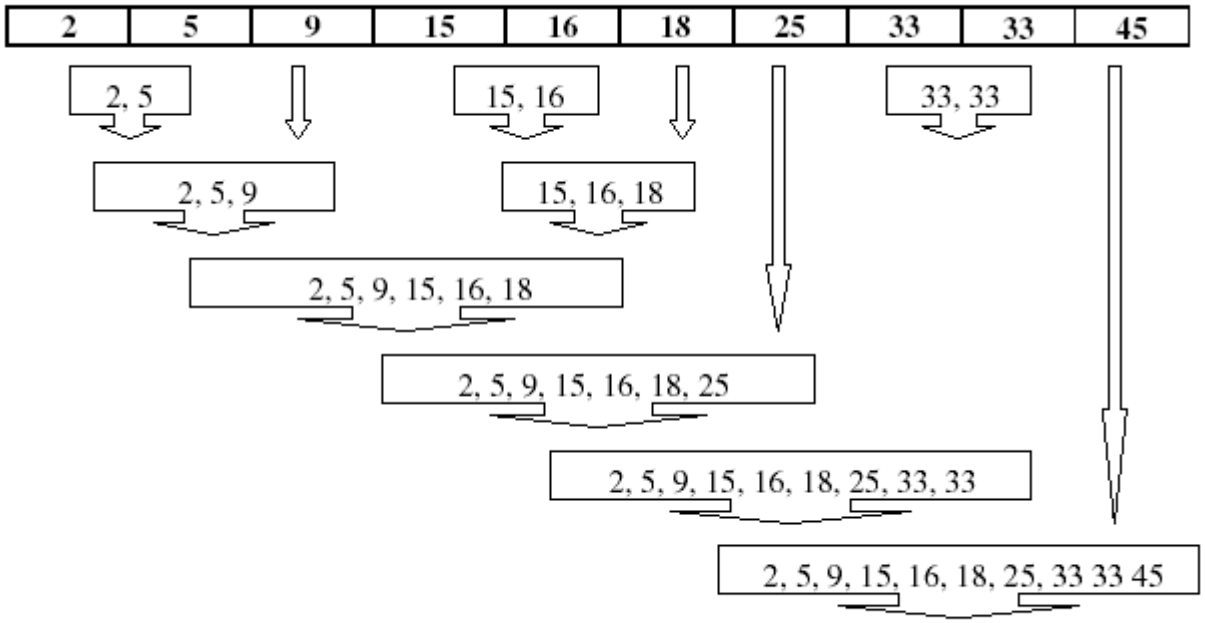


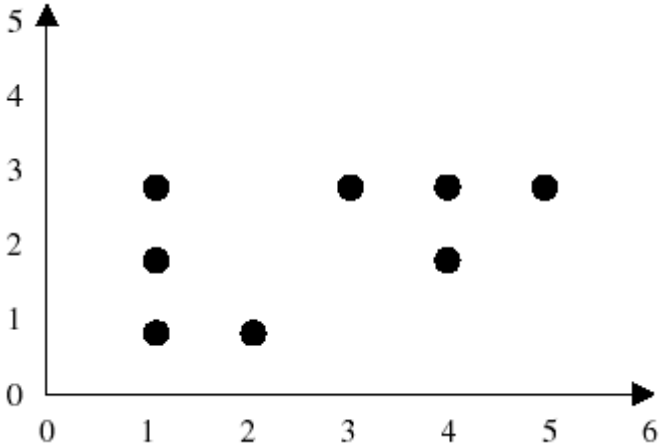
Figure 8.2 Single-linkage agglomerative clustering on the sample data set.

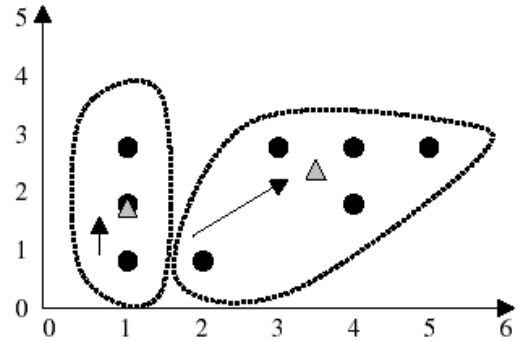
Referinta figura: D. Larose, *Discovering Knowledge in Data. An Introduction to Data Mining*, John Wiley & Sons, 2005.

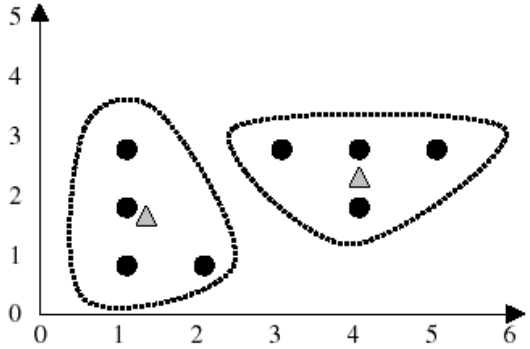
K-means clustering

1. Se stabileste valoarea lui k – numarul dorit de clusteri
2. Se aleg aleator k obiecte care sa reprezinte centrul initial al fiecaruia dintre cei k clusteri
3. Pentru fiecare obiect, se cauta cel mai apropiat centru si se adauga la clusterul respectiv
4. Pentru fiecare din cei k clusteri obtinuti se recalculeaza centrul
5. Se repeta pasii 3-5 pana la convergenta sau pana la indeplinirea unui criteriu de terminare

- Alogoritmul se poate termina
  - cand centrii clusterilor nu se mai modifica
  - Cand este indeplinit un anumit criteriu de convergenta
  - …







Referinta figuri: D. Larose, *Discovering Knowledge in Data. An Introduction to Data Mining*, John Wiley & Sons, 2005.