Mogelijkheden en keuzes bij het clusteren van onderwijsdata

Hanneke van der Hoef
Matthijs J Warrens

ORD Nijmegen
14 juni 2018

Mogelijkheden en keuzes bij het clusteren van onderwijsdata

Overzicht

- **Clusteranalyse**
- Keuzes
- Clustermethoden
- Aantal clusters

Clusteranalyse

Verzamelnaam voor een groot aantal methoden voor het vinden van groepen in data

Clustermethoden

- een tiental verschillende benaderingen
  verschillende definities van wat een cluster is
- vele honderden verschillende methoden

Grote diversiteit door

- ontwikkeling en toepassing in vele wetenschappelijke disciplines (biologie, informatica, geneeskunde, …)

Clusteranalyse

Verzamelnaam voor een groot aantal methoden voor het vinden van groepen in data

Groepen

- = clusters = profielen
- objecten, dieren, plaatjes
- individuen (leerlingen, patiënten)
- motivatieprofielen (Hanke Korpershoek)
- profielen van leerprestaties

Clusteranalyse

Verzamelnaam voor een groot aantal methoden voor het vinden van groepen in data

Data

- variabelen (metingen) nodig om objecten te onderscheiden
- typen data
  - interval data (bijv. toetsscores) **(focus hier)**
  - categorische data (bijv. goed/fout antwoorden)
  - symbolische data (bijv. verdelingen, intervallen)
  - gemengde data
- verschillende clustermethoden voor verschillende data

Voorbeeld clusteranalyse

Nederlandse Intelligentietest voor Onderwijsniveau (NIO)
(2 componenten, 6 deeltoetsen)

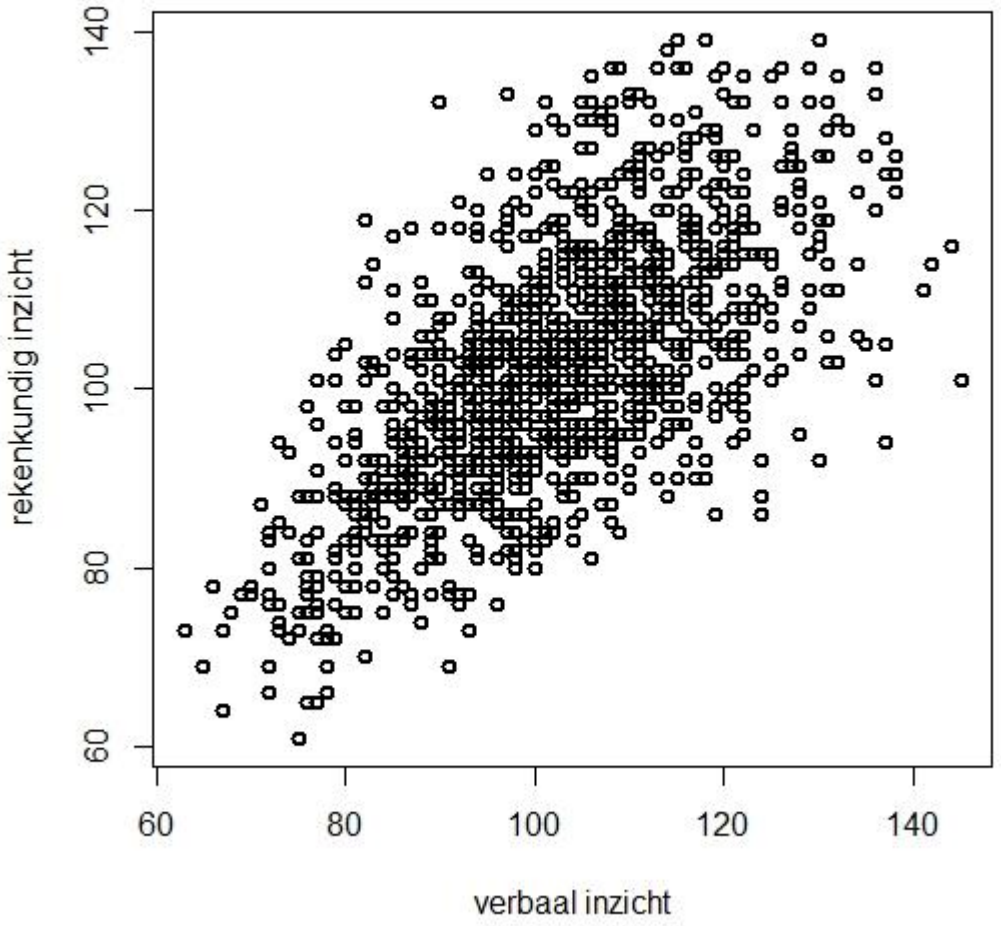
verbaal inzicht
- synoniemen
- analogieën
- categorieën

rekenkundig-ruimtelijk inzicht
- getallen
- rekenen
- uitslagen

Voorbeeld

NIO

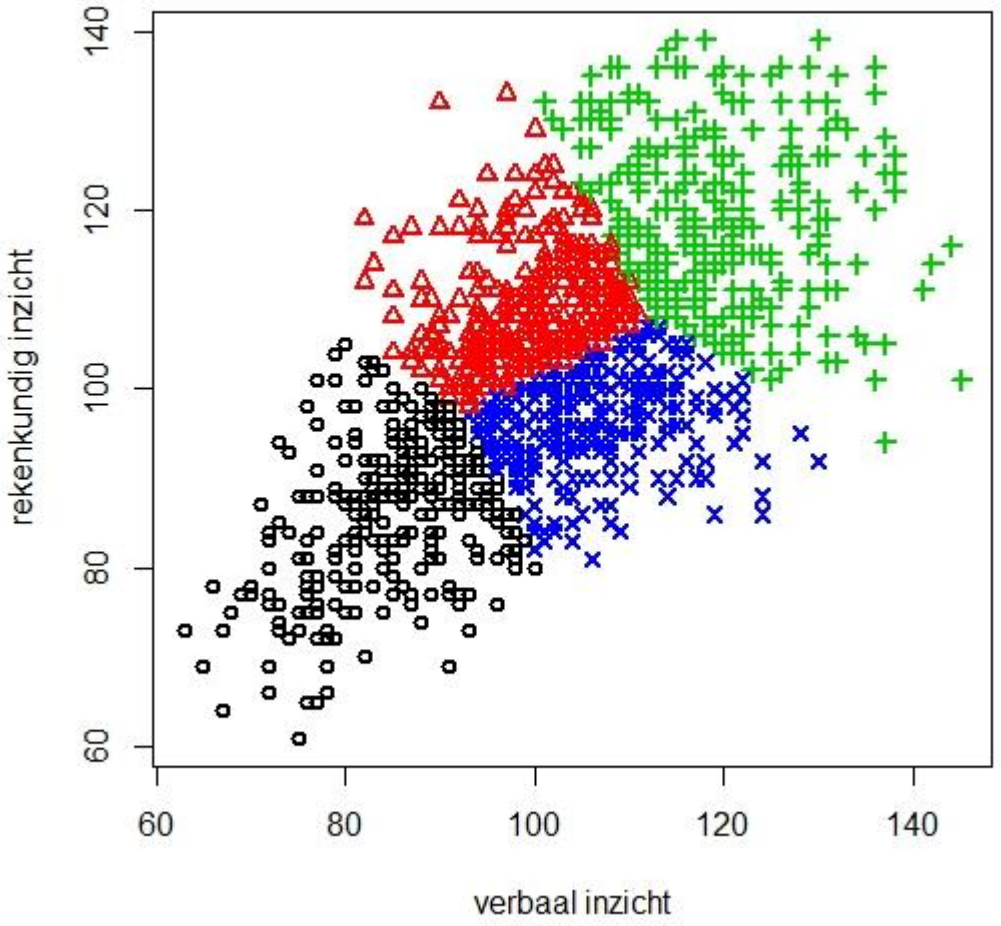
- spreidingsdiagram
- 1269 leerlingen in vo-2
- afname in 2000
- verbaal vs. rekenkundig inzicht



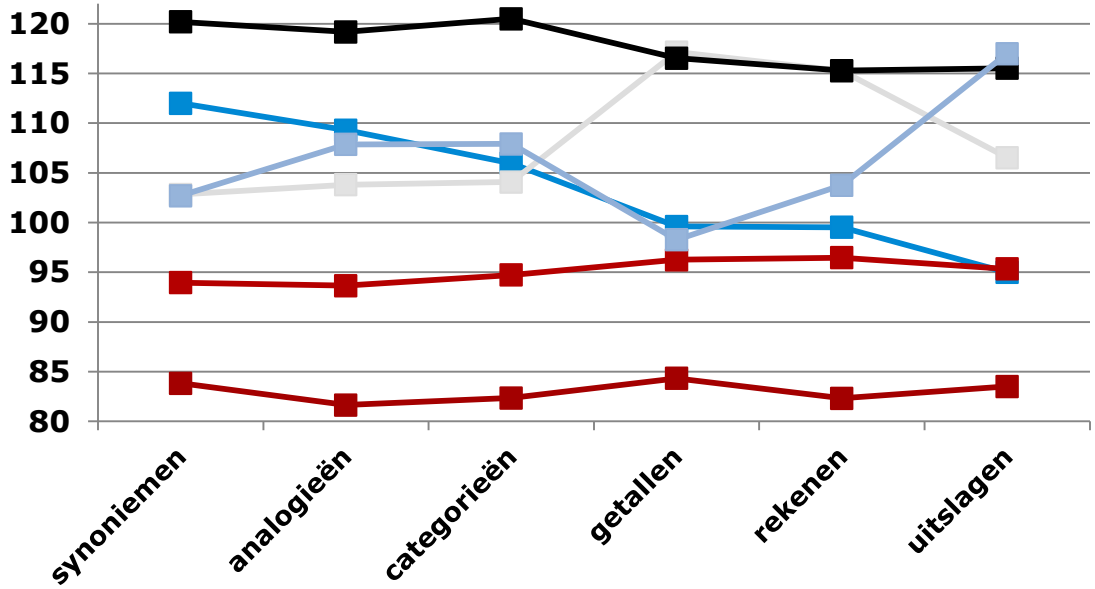
Voorbeeld

NIO

- spreidingsdiagram
- 1269 leerlingen in vo-2
- afname in 2000
- verbaal vs. rekenkundig inzicht
- k-means analyse
- 4 groepen



Profielen



Profielplots

- bij 4 of meer variabelen
- plot van gemiddelden voor iedere groep (hier 6 groepen)
- vergemakkelijkt interpretatie enigszins
- bevat relatief weinig informatie omtrent
  - aard/vorm van clusters
  - overlap tussen clusters

Mogelijkheden en keuzes bij het clusteren van onderwijsdata

Overzicht

- Clusteranalyse
- **Keuzes**
- Clustermethoden
- Aantal clusters

Moeilijkheid van clusteranalyse

Over het algemeen (o.h.a.)
is er geen ware of natuurlijke groepsindeling

Verschillende clustermethoden geven (o.h.a.) andere clusters

Verschillende keuzes geven (o.h.a.) andere clusters

Meerdere indelingen kunnen om verschillende redenen betekenisvol zijn

Wat is het doel van de clusteranalyse?

Keuzes bij clusteranalyse

Als we weten wat te clusteren, dan keuzes maken over

- variabelen (metingen)
  - alle variabelen of een selectie
  - scores op domeinen of deeldomeinen
- weging van variabelen (standaardisatie)
- clustermethode
  - benadering (partitie, hiërarchisch, mixtures)
  - afstandsmaat
- aantal clusters

Keuze variabelen

Profielen van leerprestaties

Cito Eindtoets Basisonderwijs

- hoofddomeinen (3): taal, rekenen, wereldoriëntatie
- deeldomeinen (16): 9x taal, 4x rekenen, 3x w.o.

We vinden andere clusters bij analyse van

- hoofd- of deeldomeinen (deeldomeinen interessanter?)
- alleen taaldomeinen of alle deeldomeinen

Weging variabelen

Profielen van leerprestaties

Cito Eindtoets Basisonderwijs

- hoofddomeinen (3): taal, rekenen, wereldoriëntatie
- deeldomeinen (16): 9x taal, 4x rekenen, 3x w.o.

Analyse van alleen deeldomeinen

- focus meer op taal (9x) dan rekenen (4x) of w.o. (3x)
- mogelijkheid: wegen van variabelen

Weging variabelen

Profielen van leerprestaties

Cito Eindtoets Basisonderwijs

- bereik hoofddomeinen (3): 0–135, 0–85, 0–90
- bereik deeldomeinen (16): 0–10, 0–15, 0–20, 0–25, 0–30

Domeinen met een groter bereik hebben een grotere invloed op groepsindeling

- mogelijkheid: standaardisatie van domeinen

Mogelijkheden en keuzes bij het clusteren van onderwijsdata

Overzicht

- Clusteranalyse
- Keuzes
- **Clustermethoden**
- Aantal clusters

Clustermethoden

Verschillende benaderingen

- partitie methoden
  - bijv. k-means, k-mediods
- hiërarchische methoden
  - bijv. single-, average-, complete-linkage, Ward's method
- density-based methoden
  - bijv. DBSCAN, OPTICS
- mixture modellen
  - bijv. Gaussian, t-verdeling

software: alles in R; in Mplus, Latent Gold mixtures

Keuze clustermethode

Voor alle methoden geldt

- individuen in hetzelfde cluster `lijken op elkaar'
- individuen in verschillende clusters `zijn verschillend'

→ helpt niet bij kiezen methode

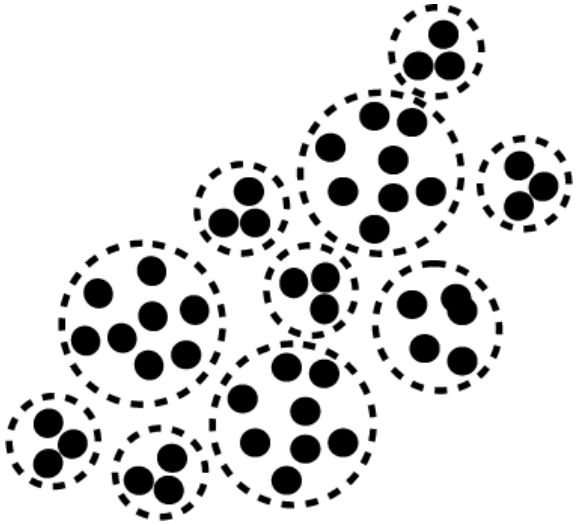
Bepaal definitie/vorm van een cluster

- kies een clustermethode welke in staat is clusters te vinden

Andere aspecten

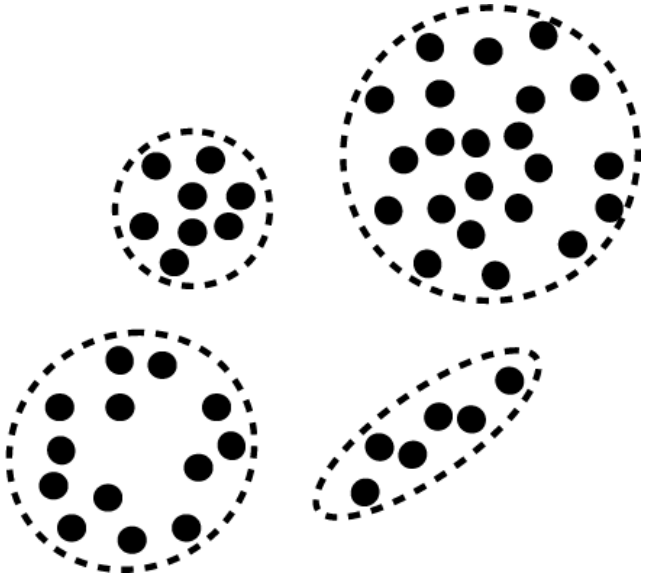
- scheiding tussen clusters
- omgang met extreme observaties (outliers)
- alle clusters even groot?

Definities/vormen clusters

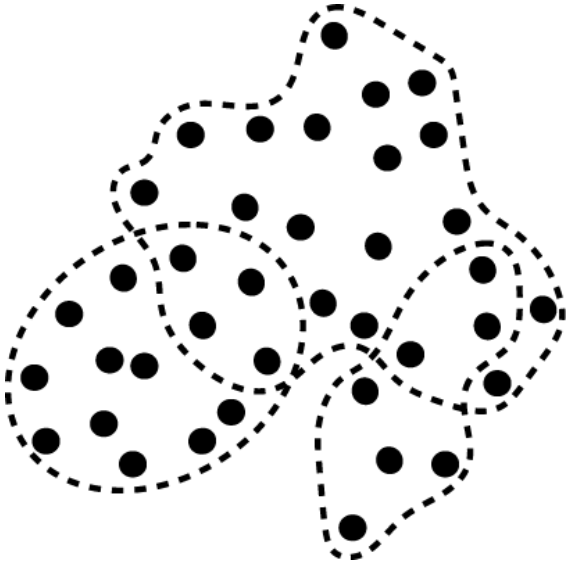


Compacte clusters
- individuen in een cluster hebben vergelijkbare scores

Clusters goed van elkaar gescheiden

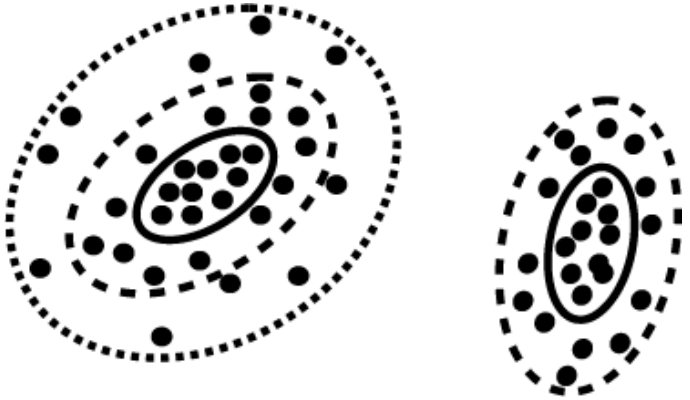


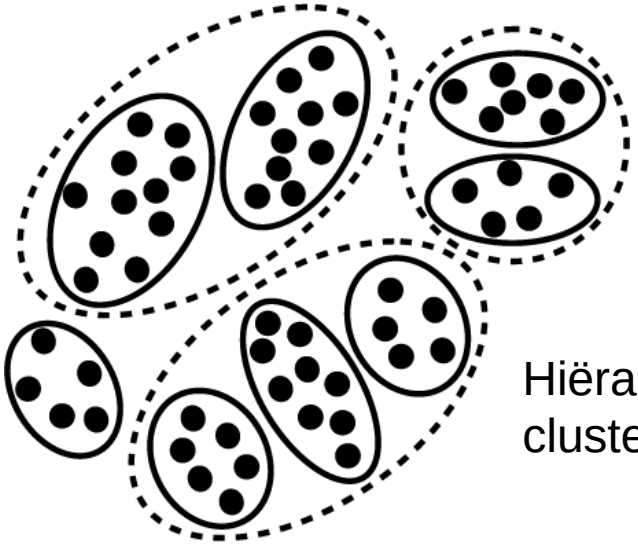
Definities/vormen clusters



Overlappende clusters
- individuen kunnen tot meerdere clusters behoren

Mixture modellen





Hiërarchie van clusters

Mogelijkheden en keuzes bij het clusteren van onderwijsdata

Overzicht

- Clusteranalyse
- Keuzes
- Clustermethoden
- **Aantal clusters**

Bepalen aantal clusters

Groot aantal statistieken beschikbaar

- Verklaarde variantie
- BIC, AIC, ASW (Average silhouette width)
  Dunn index, PG (Pearson Gamma), …

Statistieken bekijken allemaal verschillende aspecten

Niet duidelijk wat een goede statistiek is

- in het algemeen
- of voor onderwijsdata i.h.b.

Bepalen aantal clusters

Er is niet één beste groepsindeling

- indelingen kunnen om verschillende redenen betekenisvol zijn

In veel gevallen het meest informatief om een aantal groepsindelingen te bekijken

Validatie van een groepsindeling met andere informatie

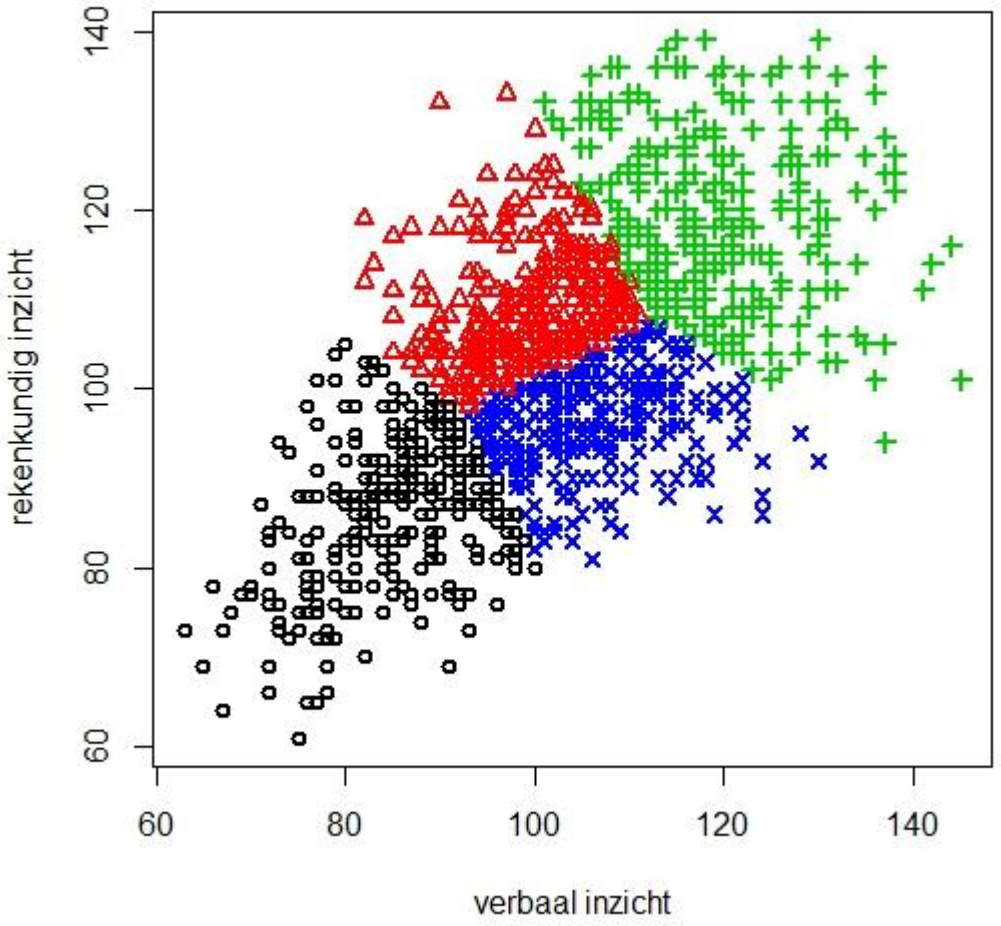
- predictieve validiteit

Onderwijsdata

Compacte clusters

- relatief veel clusters nodig
- individuen hebben anders zeer uiteenlopende scores

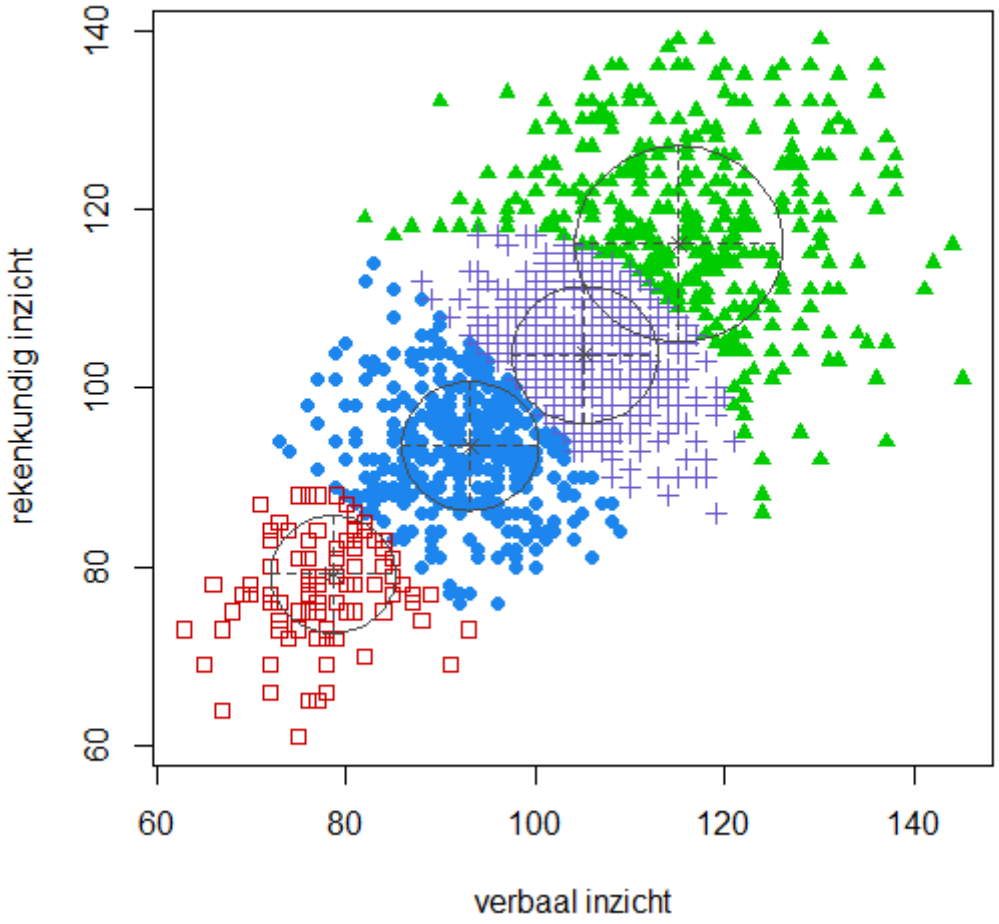
Goed van elkaar gescheiden clusters niet realistisch



Onderwijsdata

Mixture methoden minder geschikt?

- laagdimensionele data lijken niet uiteen te vallen in mixtures
- hiernaast: 1 bivariate verdeling (cluster)?



Tot slot

Meer onderzoek naar clustermethoden nodig

- in het algemeen
- en voor onderwijsdata i.h.b.

Voor meer informatie zie

- Hennig, C. (2015). What are the true clusters? Pattern Recognition Letters, 64, 53-62
- Hennig, C., Meila, M., Murtagh, F., & Rocci, R. (2015). Handbook of Cluster Analysis. New York: Chapman and Hall/CRC
- Jain, A.K. (2010). Data clustering: 50 years beyond k-means. Pattern Recognition Letters, 31, 651-666.