# Geautomatiseerde clustertechnieken

Y. Berkouwer

## 1. Inleiding

Het is jammer maar waar, dat clusteranalyse in Nederland op nog niet al te grote schaal gebruikt wordt. Redenen kunnen zijn:

(i) Er zouden "betere" technieken beschikbaar zijn, zoals faktoranalyse en schaalanalyse.

(ii) Het feit, dat voor een goed begrip van veel clustertechnieken middelbare school wiskunde voldoende is, doet velen twijfelen aan de waarde ervan.

(iii) Tot voor kort was er in Nederland nauwelijks enige programmatuur van omvang voor aanwezig.

Over punt (i) is veel gepraat, m.n. over de voor- en nadelen van faktoranalyse t.o.v. clusteranalyse. De meest gebruikte argumenten zijn enerzijds dat faktoranalyse een krachtiger methode is waarmee datastrukturen op meer verantwoorde wijze geanalyseerd kunnen worden, anderzijds dat clusteranalyse veel minder assumpties doet en op relatief kleine datasets al met sukses toegepast kan worden. Wat betreft problemen over het interpreteren van de resultaten zullen de beide technieken elkaar in elk geval weinig ontlopen.

Veel vragen blijven bestaan doordat in Nederland weinig gedaan is aan het evalueren van clustertechnieken en de daarbij gebruikte maten tussen objekten. Het is daarom nuttig dat onlangs binnen de "Contactgroep statistiese programmatuur" van de V.V.S. een aktieve groep is gevormd die zich met clusteranalyse bezighoudt. Een goede bijdrage is kort geleden ook geleverd door A. Verbeek in een artikel in Mens & Maatschappij (1).

Waarschijnlijk zal een algemenere evaluatie van clustertechnieken pas goed op gang komen wanneer eerst op grotere schaal geëxperimenteerd zal worden. Dit artikel probeert hiertoe een aantal mogelijkheden op een rijtje te zetten. Besproken zullen worden de belangrijkste technieken en bijbehorende programmatuur, die in Nederland beschikbaar zijn.

Aanvullende informatie over belangrijke, algemeen bruikbare en uitgeteste programmatuur is ten allen tijde zeer welkom.

## 2. Beschikbare technieken

Van de technieken zullen slechts zeer in het kort de hoofdlijnen beschreven worden. Tevens zullen alleen die technieken besproken worden, die disjunkte clusters opleveren.

I. Hiërarchiese clusteranalyse (Lit.:vrijwel iedere publ. over cl.an.)

A. Samenvoegende technieken (Agglomerative-)

Hierbij worden alle objekten (N stuks) aanvankelijk beschouwd als clusters bestaande uit één element. Tussen de objekten wordt een maat gedefinieerd (associatief of dissociatief) en een maat tussen clusters. Allereerst worden nu de twee meest gelijkende (resp. dichtstbijzijnde) clusters gefuseerd. Er zijn nu nog N-1 clusters over. Deze stap wordt herhaald totdat er nog slechts één cluster over is, bestaande uit alle objekten.

De definitie van de maat tussen twee clusters is het essentiële punt in het proces:

- In de beginsituatie is deze gelijk aan de tevoren gedefinieerde maat tussen de twee objekten waaruit de clusters bestaan.
- Bij iedere fusie, van zeg cluster P en cluster Q, wordt de maat, M, tussen het nieuwe cluster, P + Q, en de overige clusters, bijv. R, als volgt berekend:

$$M(R,P+Q)=a*M(R,P)+b*M(R,Q)+c*M(P,Q)+d*abs(M(R,P)-M(R,Q))$$

waarbij de keuze van a, b, c en d steeds een andere methode oplevert:

(i) Single Linkage: (Johnson)
a=b=0.5 , c=0 , d=+0.5 (bij een associatieve maat)
d=-0.5 (bij een dissociatieve maat)

(ii) Complete Linkage: (Johnson)
a=b=0.5 , c=0 , d=-0.5 (bij een associatieve maat)
d=+0.5 (bij een dissociatieve maat)

(iii) Average Linkage:
a=p/(p+q) , b=q/(p+q) , c=d=0

(iv) Centroïd:

a=p/(p+q) , b=q/(p+q) , c=-a*b , d=0

(v) Median:

a=b=0.5 , c=-0.25 , d=0

(vi) Ward's method (Error Sum of Squares):

a=(p+r)/(p+q+r) , b=(q+r)/(p+q+r) , c=-r/(p+q+r) , d=0

(vii) Lance-Williams Flexible Beta Method:

a=b=(1-BETA)/2 , c=BETA , G=0 (BETA kleiner dan 1)

(viii) McQuitty's Similarity Analysis:

a=b=0.5 , c=d=0

(ix) Methode van Elshout:

a=p/(p+q) , b=Sgn(M(P,Q))*q/(p+q) , c=d=0

Over deze methode bestaat niet veel literatuur. De methode is slechts zinvol in kombinatie met de produkt-moment korrelatie als maat. De standaard-methode bestaat hier uit het steeds fuseren van die twee clusters die in absolute zin de grootste maat hebben (N.B. De maat tussen twee clusters komt hier neer op de gemiddelde pmc tussen de objekten waaruit ze bestaan). Er bestaat ook een variant: het steeds fuseren van reciproce paren van clusters ( dat zijn clusters die elk met elkaar gemiddeld hoger korreleren dan met enig ander cluster).

In het bovenstaande geven p, q en r de aantallen objekten weer van de clusters P, Q en R.

B. Opsplitsende technieken (Divisive-)

Hierbij wordt uitgegaan van één cluster, bestaande uit alle objekten. Dit cluster wordt stap voor stap opgesplitst in steeds kleinere, "zo homogeen mogelijke" deelclusters.

1. Monothetiese divisie:
   De objekten worden onderverdeeld in de deelclusters o.g.v. hun skores op één variabele. Hiervoor wordt steeds die variabele gekozen, die de meest gunstige splitsing teweeg brengt in "homogene" groepen die onderling veel verschillen.
2. Polythetiese divisie.
   Indeling vindt plaats o.g.v. skores op meerdere variabelen.

II. Iteratieve clusteranalyse (Lit.:(2), (7) en (8))

Bij deze methode wordt van te voren opgegeven in hoeveel clusters (zeg K) de objekten geklassificeerd moeten worden. Verder moet een associatieve of dissociatieve maat tussen objekten gedefinieerd worden en eventueel een kriterium dat geoptimaliseerd dient te worden (bijv. de Error Sum of Squares).

Allereerst wordt een beginklassifikatie verkregen. Dit kan zijn:
a. een door de gebruiker gedefinieerde indeling, b. het resultaat van een voorafgaande (bijv. hiërarchiese) analyse, c. een random-indeling.
Vervolgens wordt een iteratief proces in werking gesteld om de gevonden klassifikatie te verbeteren. Hiervoor zijn twee methoden:

1. De objekten worden stuk voor stuk bekeken en (her-)ingedeeld bij het cluster met (bij een dissociatieve maat) het dichtstbijzijnde zwaartepunt. Indien alle objekten op deze manier behandeld zijn wordt gekeken of er een verbetering is opgetreden in termen van het tevoren gedefinieerde kriterium. Zo ja, dan worden opnieuw alle objekten bekeken. Zo nee, dan is het proces ten einde.
2. De objekten worden stuk voor stuk bekeken en (her-)ingedeeld bij een ander cluster indien daarmee een verbetering optreedt in termen van het kriterium. Dit wordt herhaald totdat geen enkel objekt meer een verbetering oplevert.

III. Mode Analysis (Lit.: (2) en (9))

De methode vormt natuurlijke clusters o.g.v. de dichtheid van objekten. Met het begrip "dichtheid van een objekt" wordt bedoeld de mate van aanwezigheid van andere objekten rondom dit objekt.
Allereerst wordt tussen de objekten een afstandsmaat berekend (of een associatieve maat; het proces verloopt verder analoog). Aan elk objekt, I, wordt nu een getal, A(I), toegevoegd zodanig dat A(I) kleiner is naarmate er meer objekten in de buurt van I liggen. (A(I) is dus een soort omgekeerde maatstaf voor de dichtheid). A(I) kan bijvoorbeeld zijn de gemiddelde afstand tussen objekt I en zijn K dichtstbijzijnde buren. De objekten worden nu gerangschikt naar de grootte van hun A(I). In deze volgorde worden de objekten behandeld beginnend bij het objekt met de kleinste A(I).
Hierbij zijn twee begrippen van belang:
(i) Een objekt heet "dense" als het reeds behandeld is.
(ii) Tijdens het proces wordt een drempelwaarde R bijgehouden.

Deze R is steeds gelijk aan de A(I) van het objekt dat op dat moment behandeld wordt. In de loop van het proces stijgt R dus monotoon.
De behandeling van het eerste objekt (met de kleinste A(I)) bestaat uit het initiëren van een nieuw cluster waarin dit objekt geplaatst wordt.
Bij iedere behandeling van een volgend objekt, zeg P, wordt R verhoogd tot A(P). Afhankelijk van de situatie kunnen de volgende akties ondernomen worden:

1. P verschilt meer dan R van alle dense objekten.
   Dan wordt door P een nieuw cluster geïnitieerd. Het aantal clusters wordt dus één groter.
2. P verschilt minder dan R van één of meer dense objekten die alle tot hetzelfde cluster behoren.
   In dit geval wordt P aan dat cluster toegevoegd.
3. P verschilt minder dan R van een aantal dense objekten, die tot meer dan één cluster behoren.
   Is dit het geval, dan worden al deze clusters gefuseerd en wordt P hieraan toegevoegd.
4. Steeds wanneer een nieuw punt behandeld wordt, wordt het kleinste verschil, D, berekend tussen dense objekten, die tot verschillende clusters behoren. Als nu tijdens het proces deze D kleiner is dan R dan worden de bijbehorende clusters gefuseerd.

Dit proces gaat door totdat aan een aantal eindkondities is voldaan (bijv. een grens voor het percentage dense objekten).

IV. Clusteranalyse volgens Boon van Ostade (Lit.:(4))

De methode gaat altijd uit van gedichotomiseerde data en gebruikt als maat altijd de produkt-moment-korrelatie. Meestal worden variabelen geclusterd. De clusters van variabelen worden stuk voor stuk in hun geheel opgespoord; m.a.w. eerst wordt een homogene deelverzameling van de te clusteren variabelen gezocht; dit wordt cluster 1.Daarna wordt uit de resterende variabelen cluster 2 gezocht, etc.
Een cluster wordt als volgt gevormd:
Als eerste element wordt de variabele gekozen die, in absolute zin, het hoogst korreleert met de totale somskore (Hiervoor kan men kiezen de som van alle variabelen of de som van alle overige variabelen). Vervolgens wordt steeds die variabele toegevoegd, die, in absolute zin, het hoogst korreleert met de somskore van de variabelen die reeds in het cluster zijn opgenomen.

Tevens wordt, indien die korrelatie negatief is, de betrokken variabelen omgekodeerd, d.w.z. de nullen worden enen en omgekeerd.
Indien de korrelatie van de clustersomskore met de nieuwe variabele te laag wordt, gaat het proces over op een volgend cluster.

3. Programmatuur

| naam en auteur | methoden;indeling: zie hierboven | data | berekende maten |
|---|---|---|---|
| CLUSTAN-IB D. Wishart | - Hiër.: Sam.: 1-8<br>Opspl.: 1-2<br>- Iter.: 1(Geen "kriterium"; stopt als geen herindeling meer plaatsvindt)<br>- Mode Analysis<br>Kan geen missing values behandelen. | Ruwe data of ass-/diss matrix | Keuze uit 38 stuks. Mogelijkheid om zelf een maat te programmeren |
| CLUSTAN-IC D. Wishart | Alsbij IB.<br>Tevens: plotten van dendrogrammen en scattergrammen Labelling | Alsbij IB | Alsbij IB |
| CLUSTER (1) J. Bethlehem | Hiër.: Sam.: 1-6 en 9<br>Tevens dendrogrammen | ass-/diss matrix | - |
| DISTANC J. Bethlehem/R.Kaas | berekent slechts maten | Ruwe data | Keuze uit 14 maten |
| CLUSTER (2) A.H.J. Schmidt P.F.v. Jaarsveld | Hiër.: Elshout.<br>Beide varianten<br>Tevens dendrogrammen | ass.-matrix | - |

| naam | taal en systeem | dokumentatie | kontaktadres | kode in het Steinmetz-archief |
|---|---|---|---|---|
| CLUSTAN-IB | FTN CDC | PB: Y.Berkouwer | Y.Berkouwer<br>CDA-Utr.<br>Varkenmarkt 2<br>T.030-331211-343 | 8 |
| CLUSTAN-IC | FORTRAN IV Hext IBM | PB: D.Wishart | P.J.S. Boon<br>URC<br>Toernooiveld<br>Nijmegen<br>T.080-558833-2271 | 8 |
| CLUSTER (1) | ALGOL CDC | PB: J.Bethlehem | J.Rijvoordt-MC<br>2de Boerhavestr.49<br>Amsterdam<br>T.020-947272-08 | - |
| DISTANC | ALGOL CDC | PB: R.Kaas en J.Bethlehem | J.Rijvoordt<br>(zie boven) | - |
| CLUSTER (2) | FTN CDC | PB: P.E.H.Uphoff | Y.Berkouwer<br>(zie boven) | - |

| naam en auteur | methoden;indeling: zie hierboven | data | berekende maten |
|---|---|---|---|
| HCAJOHN C.v.d.Wijgaart | Hiër.: Johnson | ass./diss. matrix | - |
| HCACOR C.v.d.Wijgaart | Hiër.: Average Linkage | korr.- matrix | - |
| MARIMAX J.Elshout | Hiër.: Elshout m.b.v.Reciproce paren | korr.- matrix | - |
| HCADIST C.v.d.Wijgaart | Hiër.: Centroïd | Ruwe data | Eucl.afstand |
| HICLA CRI-RUL | Hiër.: Johnson | ass./diss. matrix | - |
| Gron.-psych-09 B.J.Kouwer | Hiër.: Johnson Tevens dendrogrammen | Ruwe data | |
| HICLU T.Kwaaitaal | Hiër.: Johnson Tevens:Kendall's-tau-test | diss-matrix | Ultrametric distances |
| MIKCA D.J.McRae | Iter.:1-2 Vier optimaliserings-kriteria(w.o. Error Sum of Squares).Genereert drie redelijke begin-klassifikaties en kiest de beste. | Ruwe data | 1.Eucl.afstand 2.Gewogen Eucl. afstand (korrektie voor verschillen in varianties) 3.Mahalanobis afstand. korrektie voor kovarianties en verschillen in varianties. |
| OSTER H.Borgers | Cl.An. volgens Boon van Ostade | Ruwe data (binair) | p.m.c. |
| ICLA CRI-RUL | Cl.An. volgens Boon van Ostade | Ruwe data (binair) | p.m.c. |

| naam | taal en systeem | dokumentatie | kontaktadres | kode in het Stein-metz-archief |
|---|---|---|---|---|
| HCAJOHN | ALGOL CDC | PB: C.v.d.Wijgaart | TC-Roeterstr.15 A'dam.T.020-5222702 | - |
| HCACOR | ALGOL CDC | PB: C.v.d.Wijgaart | T.C. (zie boven) | 81 |
| MARIMAX | ALGOL CDC | PB: J.Elshout | T.C. (zie boven) | 82 |
| HCADIST | ALGOL CDC | PB: C.v.d.Wijgaart | T.C. (zie boven) | 94 |
| HICLA | FTN IBM | PB:?(wel aanwezig) | TC Remmetzwaal-CRI RUL.:T.01710-14833 | 284 |
| Gron-psych.-09 | ALGOL CDC | PB:?(wel aanwezig) | H.Camstra-RUG. Oude Bot.str.23 T.050-115385 | 433 |
| HICLU | FORTRAN IV G/H IBM | PB: T.Bezembinder T.Kwaaitaal | M. Thissen Psych.Lab. KUN Erasmuslaan 16 T.080-512009 | 22 |
| MIKCA | FTN CDC | PB: D.J.McRae | Listor-groep R.C.- RUG. W&N-gebouw De Paddepoel-Gron. | - |
| OSTER 2/3 | FORTRAN IV G/H FTN IBM/CDC | PB: H.Borgers | Thissen (zie boven) | - |
| ICLA | FTN IBM | PB: CRI-RUL | T.C.Remmelzwaal (zie boven) | 283 |

Verder kunnen nog een aantal aanverwandte programma's genoemd worden, zoals:

- GROUPAN: Roll-Call-Analysis (Inf.: TC-Amsterdam; St.Archief:87)
- LPA2: Latent Profile Analysis (Inf.: H.Camstra-Gron.; St.Archief:445)
- AID2: Automatic Interaction Detection (Inf.: TC-Amsterdam)

en vele anderen.

4. Aantekeningen bij het schema.

Behalve dat het lang niet volledig is, zou een schema als hierboven er over een half jaar heel anders uitzien:

- Door een sterk groeiend interuniversitair samenwerkingsverband (waarvan dit blad een gevolg is) zullen de vele overlappingen die erin voorkomen hopelijk afnemen. De noodzaak hiertoe blijkt bijvoorbeeld uit het feit dat er momenteel in Nederland minstens zes verschillende programma's operationeel zijn om clusteranalyse volgens Johnson uit te voeren!
- De ontwikkeling van grotere pakketten, waarin veel uiteenlopende clusteralgorithmen ondergebracht zijn, gaat zeer snel:
  (i) Het toch al grote pakket CLUSTAN wordt nog steeds uitgebreid.
  (ii) In Groningen wordt een pakket ontwikkeld waar MIKCA slechts een onderdeel van is.
  (iii) In Nijmegen is het pakket ASTER ontwikkeld dat veel meer omvat dan OSTER.
  (iv) In Amsterdam (TC.) zijn de methoden van Elshout, Johnson en Boon van Ostade opgenomen in het STAP-projekt en zullen dus binnenkort binnen een SPSS-raam operationeel zijn.

Vrijwel alle genoemde technieken kunnen door de volgende vier programma's uitgevoerd worden:

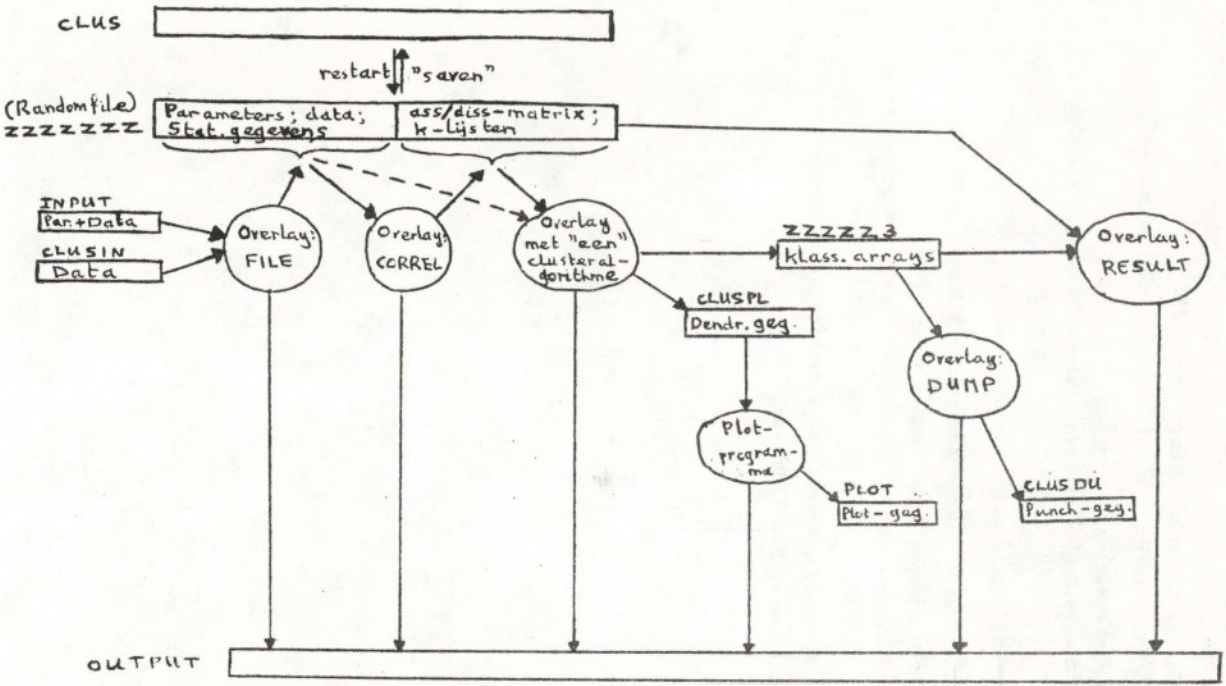
1. CLUSTAN
2. MIKCA
3. CLUSTER (2)
4. OSTER

ad 1. Voor de CDC-versie zijn de plotroutines echter nog niet operationeel. Daarom moeten de dendrogrammen nog gedeeltelijk met de hand gemaakt worden. Ook kan voor geprinte dendrogrammen van zeven hiërarchiese methoden het programma CLUSTER (1) gebruikt worden. Van de door CLUSTAN gebruikte files zal hierna ter informatie nog een overzicht gegeven worden.

ad 2. Dit programma wordt, voor zover bekend, nog niet uitgebreid gebruikt (slechts bij Med. Anatomie-Gron. en bij Soc. Geogr.-VUA). Het voert echter een zeer gedegen iteratieve analyse uit en beschikt over de Mahalanobis-afstand.

ad 3. Clusteranalyse volgens Elshout wordt door MARIMAX (T.C.) en CLUSTER(1-MC) ook verricht maar CLUSTER(2-CDA) voert beide varianten uit en print ook een dendrogram.

ad 4. OSTER is, in een aangepaste versie, ook in Amsterdam en Utrecht aanwezig, maar met het beschikbaar komen van vele andere technieken is deze methode enigszins op de achtergrond geraakt. Mogelijk komt dit door het feit dat slechts gedichotomiseerde data verwerkt kunnen worden en door het nadeel dat eenmaal gevormde clusters gedurende het gehele verdere proces niet meer gewijzigd kunnen worden.




Overzicht van door CLUSTAN-IB gebruikte files.

Literatuur

(1) A.Verbeek:"Clusteranalyse" - Mens en Maatschappij 1976-3 pag. 230-272.

(2) B.Everitt:"Cluster Analysis" - Heinemann, '74.

(3) J.G.Bethlehem:"Handleiding voor hiërarchiese clusteranalyse", Mathematisch Centrum, '76.

(4) A.H.Boon van Ostade:"De iteratieve clusteranalyse", Dissertatie, KUN, '69.

(5) H.H.Bock:"Automatische Klassifikation", Vandenhoeck & Ruprecht,'74.

(6) B.S.Duran & P.L.Odell:"Cluster Analysis - A Survey", Springer, '74.

(7) J.A.Hartigan:"Clustering Algorithms", Wiley '75.

(8) M.G.Kendall:"Cluster Analysis", Scientific Control Systems Ltd., '68.

(9) D.Wishart:"Fortran II programs for 8 methods of Cluster Analysis", Kans.Geol.Comp.Contr. No.38, Kansas, '69.