Een simulatieonderzoek naar het gedrag van enige maatgrootheden in de clusteranalyse

M.C. Deurloo'), D.J. Kuik "), J.A. Theune ")
Vrije Universiteit, Amsterdam.

1. Inleiding

Bij clusteranalyse wordt de vraag gesteld hoe een verzameling objecten (of punten) op zodanige wijze in groepen (clusters) is in te delen, dat een in de verzameling aanwezige structuur naar voren komt en als basis kan dienen voor verdere interpretatie.

Het vage karakter van de probleemstelling heeft er toe geleid dat talloze heuristische procedures, gedefinieerd door een of ander algoritme en zonder verdere formele achtergrond, ter oplossing zijn voorgesteld.

Een belangrijke beperking in de waarde van clusteranalytische methoden is in veler ogen het ontbreken van standaardtechnieken die na kunnen gaan hoe adequaat een verkregen groeperingsresultaat is. Met name is er gebrek aan middelen om de relevantie van gevonden clusters te beoordelen.

Het onderzoek waarvan hier verslag wordt gedaan is opgezet om bouwstenen aan te dragen voor grootheden die althans een hypothese van structuurloosheid in de verzameling objecten kunnen toetsen. Daartoe wordt een reeks eenvoudige maten geïntroduceerd waarvan verwacht mag worden dat ze uiteenlopende aspecten van structuur kwantificeren. Er is getracht om rekening te houden met hetgeen in de literatuur over toetsingsgrootheden is gepubliceerd en daarnaast de maten zo veel mogelijk te laten aansluiten bij de gegevens die gebruikt worden in de uitvoering van het algoritme.

2. Groeperingsmethoden

De meeste groeperingsmethoden gaan uit van het begrip similariteit of van het begrip dissimilariteit. D.w.z. voor elk tweetal

') subfaculteit Sociale Geografie en Planologie
") vakgroep Medische Statistiek

objecten v_i en v_j is een getal s_{ij} (resp. d_{ij}) gedefinieerd dat de mate van overeenkomst (resp. de afstand) aangeeft.

De vele, sterk uiteenlopende methoden kunnen op verschillende wijzen worden ingedeeld. Een bekende hoofdindeling is in partitiemethoden, hiërarchische methoden en overige.

Bij partitiemethoden wordt, uitgaande van een begingroepering, gezocht naar een indeling van de objecten in een vast aantal groepen waarbij een gekozen doelfunctie een uiterste waarde bereikt. De gevonden groepering is echter afhankelijk van de begingroepering.

Hiërarchische methoden zijn methoden die op een stapsgewijze benadering berusten. Bij elke stap - ook niveau genaamd - wordt een volgende indeling verkregen, uitgaande van de vorige. Dit kan op twee wijzen geschieden, waardoor een onderverdeling ontstaat in agglomeratie- en splitsingsmethoden. Bij agglomeratiemethoden worden op elk niveau (bij elke stap) twee of meer clusters samengevoegd tot één nieuw cluster, terwijl bij splitsingsmethoden op elk niveau een opsplitsing van één of meer clusters plaatsvindt.

De overige methoden berusten in het algemeen op denkwijzen die ontleend zijn aan het zich voorstellen van de wijze waarop in puntenwolken in het platte vlak clustervorming tot uiting komt.

Onder de vele agglomeratiemethoden zijn de single link en de complete link procedure de meest eenvoudige. Bij single link wordt voor elk paar clusters het minimum bepaald van alle afstanden van objecten uit het ene cluster met objecten uit het andere, terwijl bij complete link het maximum van deze afstanden bepaald wordt. In beide gevallen worden die twee clusters samengevoegd waarvoor deze bepaalde afstand minimaal is. Single link leidt vaak tot langgerekte, ketenvormige groepen, terwijl bij complete link compacte clusters gevormd worden die intern sterk homogeen zijn. Beide methoden zijn monotoon invariant. Dit betekent dat de opbouw van de hiërarchie dezelfde blijft bij een monotone transformatie van de afstanden. Zij hangt dus slechts af van de rangorde der afstanden. Wat betreft hoeveelheid rekenwerk steken zowel single link als complete link gunstig af bij agglomeratiemethoden die niet monotoon invariant zijn.

In dit onderzoek zijn om bovenstaande redenen de beschouwingen beperkt tot de single link en de complete link methode. Dit mede gezien het feit dat deze methoden in gebruik ruim verbreid zijn en zij

binnen het scala der agglomeratiemethoden tegenpolen zijn.

3. Toetsingsproblematiek

Als er een grootheid is gedefinieerd die een aspect van structuur kwantificeert - verder ook aangeduid als structuurmaat - en de kansverdeling ervan onder een gestelde nulhypothese is bekend, is het mogelijk tot toetsing over te gaan.

Het is echter bekend, dat verschillende groeperingsalgoritmen op hetzelfde waarnemingsmateriaal sterk uiteenlopende groepsindelingen kunnen geven. Dit heeft als consequentie dat structuurmaten voor verschillende groeperingsmethoden een sterk verschillend gedrag zullen vertonen. Men kan dus de kansverdeling van een structuurmaat onder de gestelde nulhypothese slechts zien in het licht van een gegeven algoritme.

Om een toestand van "structuurloosheid" te beschrijven is het meest gebruikte uitgangspunt, aan te nemen dat de objecten dezelfde verwachtingswaarde hebben en dat de in het waarnemingsmateriaal gevonden afstanden door toeval aanwezig zijn.

Zo bestaat de mogelijkheid aan te nemen dat de metingen aan de objecten afkomstig zijn uit een bekende multivariate simultane verdeling. Dit wordt onder meer gebruikt in een simulatiestudie van Rand (1971) die een vijfvariate normale verdeling gekozen heeft en in een studie van Kent Kuiper en Fisher (1975) die met normale verdelingen en (op beperkte schaal) met een Cauchy verdeling experimenteren.

Een andere mogelijkheid is door Hartigan aan Ling (1973) voorgesteld. In dit model zijn de afstanden trekkingen uit een multivariate normale verdeling met d_{ij} en d_{kl} onafhankelijk voor $i \neq k$ en $j \neq l$ en met correlatie ρ tussen d_{ij} en d_{kj} voor $i \neq k$. In de zo opgebouwde afstandenmatrix zou dan tot op zekere hoogte aan de driehoeksongelijkheid zijn voldaan.

Een bezwaar van beide mogelijkheden is dat ze voor theoretische beschouwingen te ingewikkeld zijn. Bij de huidige stand van zaken zijn er slechts resultaten te boeken d.m.v. simulatiestudies. Vanwege dit bezwaar kiest Ling, die een theoretische behandeling voorstaat, daarom noodgedwongen een ander kansmodel. Hij veronderstelt eenvoudig dat elke afstandenmatrix een even grote mate van waarschijnlijkheid bezit. Met behulp van de theorie van "random graphs" leidt hij een

distributietheorie af voor de single link methode. Een bezwaar hiervan is dat in praktijksituaties in het algemeen wel relaties tussen afstanden bestaan zelfs al zijn de objecten geheel onafhankelijk verdeeld. Deze constatering heeft ons er toe gebracht een kansmodel te kiezen dat ontleend is aan het voorstel van Hartigan.

4. Structuurmaten

4.1. In de literatuur gegeven maten

Speciaal voor gebruik binnen agglomeratiemethoden zijn door Ling twee begrippen gedefinieerd:

a) de isolatieindex van een cluster. Hieronder wordt verstaan het verschil (in rangnummer) van de afstand waarbij het cluster opgaat in een groter cluster en de afstand waarbij het is gevormd.

b) de compactheid van een cluster. Dit is een maat voor de relatie tussen (het rangnummer van) de afstand waarbij het cluster wordt gevormd en de relatieve omvang van het cluster.

Estabrook (1966) en Wirth c.s. (1966) definieren voor gebruik bij de single link methode:

c) de samenhangendheid van een cluster door

$$S = \frac{t-(m-1)}{\frac{1}{2}m(m-1)-(m-1)}$$

waarin t: aantal afstanden tussen punten in het cluster, die hoogstens gelijk zijn aan de afstand waarbij het cluster is gevormd

m: aantal punten in het cluster.

d) de "moat" van een cluster. Dit is een maat die overeenkomt met de isolatieindex van Ling.

4.2. Nieuwe maten

4.2.1 Maten gebaseerd op clusteromvangen

Voor single link en complete link kan men ook op elk niveau de omvangen der clusters als structuurmaten gebruiken, evenals functies hiervan. De omvangen der clusters, geordend naar afnemende grootte, worden aangegeven met n1, n2, enz., terwijl de omvang van het zojuist gevormde cluster ook aangeduid wordt met n_C.

Daarnaast zijn de volgende functies van de omvangen in ogenschouw genomen:

ng: het geometrisch gemiddelde

nv: de variantie

ns: de scheefheid

Om ze als structuurmaat handzamer te kunnen gebruiken zijn de grootheden ng, nv en ns uniform gemaakt, d.w.z. ze zijn gedeeld door het theoretisch maximum dat ze op het betreffende niveau aan kunnen nemen.
Opmerking: De keuze van het geometrisch gemiddelde komt voort uit het feit dat het arithmetisch gemiddelde der clusteromvangen op zeker niveau k constant is, en wel: $\bar{n} = \frac{n}{n-k}$.

4.2.2 Maten gebaseerd op de afstandenmatrix

Zij V de verzameling van objecten v_i en n het aantal elementen van V (d.w.z. $n = || V ||$). Beschouw verder de matrix A van rangnummers van afstanden. Voor elk cluster C gelden de volgende definities:

$I_{C,d} = \{(v_i,v_j) | a_{ij} < d;\ v_i \varepsilon C,\ v_j \varepsilon C,\ i<j\}$

$U_{C,d} = \{(v_i,v_j) | a_{ij} < d;\ v_i \varepsilon C,\ v_j \notin C\}$

waarin d een willekeurig getal is (de z.g. "drempelwaarde").

Men kan zich (v_i,v_j) voorstellen als de verbindingslijn van v_i en v_j. $I_{C,d}$ is dan te interpreteren als de verzameling van alle verbindingslijnen tussen punten van C met lengte kleiner dan d; terwijl $U_{C,d}$ op te vatten is als de verzameling van alle lijnen die punten van C verbinden met punten buiten C en met eveneens een lengte kleiner dan d.

Bij elk cluster C zijn twee afstanden op natuurlijke wijze in de loop van een clusteranalyse aan te geven, t.w.

s_C: de afstand waarbij C uit twee clusters wordt gevormd

r_C: de afstand waarbij C in een groter cluster opgaat

Zodoende komen vier "natuurlijke" aantallen naar voren:

$i_C = ||I_{C,s_C}|| + 1$ $\qquad$ $i'_C = ||I_{C,r_C}||$

$u_C = ||U_{C,s_C}||$ $\qquad$ $u'_C = ||U_{C,r_C}|| + 1$

Opmerkingen:

1) Het toevoegen van +1 aan twee van deze aantallen vloeit voort uit het feit dat voor elk cluster op het betreffende moment juist één afstand bestaat die gelijk is aan de drempelwaarde. D.w.z. dat er op het ogenblik dat het cluster verschijnt juist één verbindingslijn met lengte s_C is tussen twee punten van het cluster, terwijl er op het moment dat het cluster verdwijnt juist één verbindingslijn met lengte r_C is die een punt van het cluster verbindt met een punt buiten het cluster.

2) Uit de definities volgt, dat voor single link geldt: $u_C=0$, $u'_C=1$ en voor complete link: $i_C=i'_C=\frac{1}{2}n_C(n_C-1)$.

Met behulp van deze aantallen zijn op meerdere wijzen structuurmaten te vormen. Zo zijn twee van de in de literatuur gegeven maten er in uit te drukken, n.l.

1) de isolatieindex van Ling: $\iota = (i'_C+1)-i_C$

en 2) de samenhangendheid van Estabrook: $S = \dfrac{i_C-(n_C-1)}{\frac{1}{2}n_C(n_C-1)-(n_C-1)}$

Deze maten zijn echter voor complete link nutteloos daar dan $\iota=1$ en $S=1$.

Door ons zijn daarom de volgende generalisaties ingevoerd:

$q1 = (i'_C+u'_C) - (i_C+u_C)$

$$q2 = \frac{i_C}{\frac{1}{2}n_C(n_C-1)} - \frac{u_C}{n_C(n-n_C)}$$

Daarnaast is als maat voor compactheid gedefinieerd:

$$q3 = \frac{i'_C}{u'_C}$$

Om een goed inzicht in het gedrag van q2 en q3 te krijgen, zijn tevens de volgende, van hen afgeleide maten beschouwd:

$q2a = q2 * n_C$

$q3a = q3 / n_C$

$$q3b = q3 * \frac{2(n-n_C)}{n_C-1}$$

5. Vraagstelling en uitvoering van het onderzoek

In dit onderzoek is van de volgende vraagstelling uitgegaan: Wat is het gedrag van de grootheden q1, q2, q3, q2a, q3a, q3b, ng, nv, ns, n_c, n1, n2, ... in een clusteranalyse volgens zowel de single link methode als volgens de complete link methode onder de aanname dat de punten dezelfde verwachtingswaarde hebben, terwijl bij steekproeftrekking de afstandenmatrix ontstaan geacht mag worden op de wijze zoals aangegeven in het model van Hartigan?

Om deze vraag voor n=20 te beantwoorden zijn er 1000 simulaties verricht. Elke simulatie is op de navolgende wijze uitgevoerd en uitgewerkt. Er zijn 20 onafhankelijke trekkingen P_i gedaan uit de standaardnormale verdeling. Ter verkrijging van de afstanden d_{ij} $(i<j)$ is voor elke afstand één trekking u_{ij} gedaan uit een normale verdeling met gemiddelde μ_{ij} en variantie σ^2. Dan is d_{ij} = abs $(u_{ij}+P_i * P_j)$ gesteld. Daarmee is de afstandenmatrix opgebouwd naar het model van Hartigan. Daar wij de kansverdelingen der structuurmaten willen bepalen onder de veronderstelling dat de verzameling punten geen structuur bezit, is $\mu_{ij}=0$ gekozen. Verder is $\sigma^2=\frac{1}{2}$ genomen. In de ontstane matrix zijn de afstanden vervangen door hun rangnummers. Op deze - gerangordende - matrix A is zowel een single link als een complete link clusteranalyse uitgevoerd. In beide analyses is voor elke structuurmaat de waarde berekend op alle relevante niveau's. Voor n=20 zijn dat maximaal de niveau's 1 t/m 18. Hierbij zij opgemerkt dat onder meer niveau 1 irrelevant is voor de structuurmaten die gebaseerd zijn op de omvangen der clusters; immers op niveau 1 is in elke simulatie slechts 1 cluster aanwezig die dan omvang twee heeft. Om vergelijkbare redenen zijn er meer niveau's voor bepaalde maten irrelevant.

Tevens is voor een aantal maten het maximum bepaald dat deze over de 18 niveau's aannemen. De minima blijven voor de meeste structuurmaten over alle simulaties constant.

Om na te gaan in hoeverre er (lineaire) samenhangen bestaan tussen de verschillende structuurmaten, die tijdens een clusteranalyse berekend worden, zijn de lineaire correlatiecoëfficiënten (volgens Pearson) bepaald voor alle mogelijke combinaties van maten en niveau's.

6. Globale resultaten en discussie

Bij onderlinge vergelijking der structuurmaten blijken er een aantal voor beide clusteranalyses een grote mate van gescheiden informatie te dragen. Deze groep bestaat uit q1, q2, q2a, q3b, n1, n2 en n3. (De relevante informatie in n3 zal overigens gezien de beperkte steekproefomvang (n=20) niet groot meer zijn.) De overige grootheden voegen weinig nieuwe informatie meer toe.

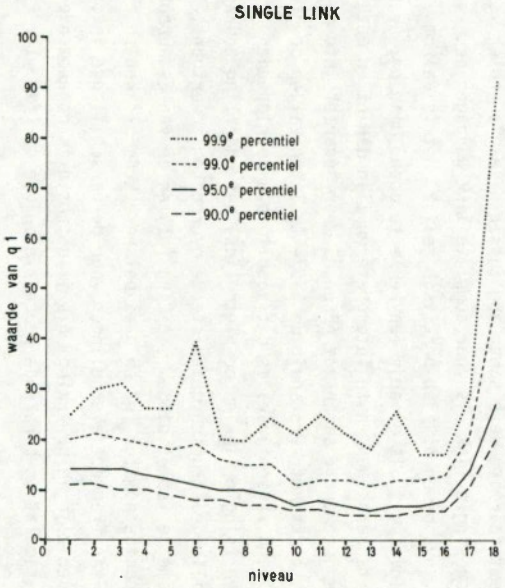
De maxima van de grootheden q1 en q3b blijken nog extra informatie omtrent groepsstructuur te bevatten, d.w.z. informatie die niet tot uitdrukking komt in de structuurmaten zoals die op de verschillende niveau's berekend worden. Daar de maxima op een gehele analyse betrekking hebben, zijn ze echter niet volledig vergelijkbaar met de ingevoerde structuurmaten.

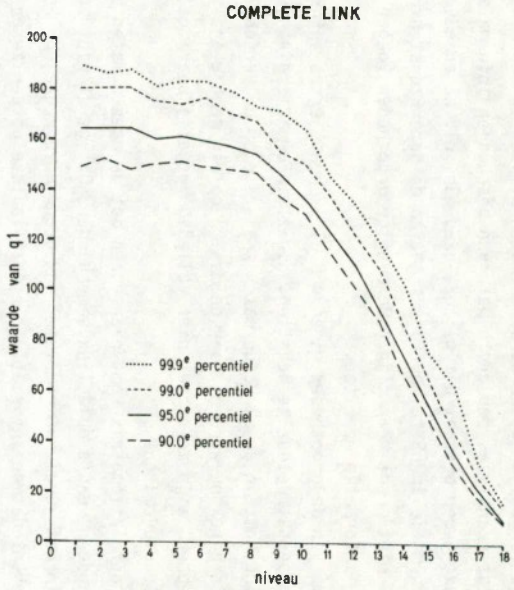
Uit de correlatiematrix komt naar voren dat de meeste maten op de verschillende niveau's duidelijk samenhangen. Dit geldt niet voor q1 en q2 en in het geval van complete link evenmin voor q2a, q3 en q3b. Dit betekent dat de informatie die deze laatste op een zeker niveau geven, in redelijke mate als onafhankelijk kan worden beschouwd van de informatie die zij op andere niveau's verstrekken. Er zij hier opgemerkt dat de praktische bruikbaarheid van een structuurmaat afneemt naarmate deze samenhang tussen niveau's groter is.

Zowel voor single link als voor complete link gedragen de grootheden ng, nv, ns en n1 zich praktisch gelijk. Het is daarom voor de praktijk zinvol uit dit groepje een keuze te doen. Daar ng, nv en ns moeilijker te berekenen en te interpreteren zijn dan n1 en deze laatste, zoals opgemerkt, samen met n2 en n3 meer "aparte" informatie verstrekt, lijkt het verstandig de keuze te laten vallen op n1.

Een opvallend verschijnsel is dat voor twee uiteenlopende analysemethoden (single link en complete link) het onderling gedrag van de verschillende structuurmaten sterke overeenkomst vertoont, waarbij met name de isolatieindex van Ling (q1) en de van Estabrook afgeleide compactheidsmaten (q2 en q2a) belangwekkend blijken te zijn.

Tabellen van frequentieverdelingen voor de verschillende interessante grootheden zijn in bewerking. Als illustratie is opgenomen een aantal uit de simulaties berekende percentielen van q1 (fig. 1).





Figuur 1. Percentielen van q1.

In verder onderzoek zou het gedrag van de ontwikkelde structuurmaten bij verschillende steekproefgrootten bestudeerd kunnen worden. Ook kunnen alternatieve hypothesen opgesteld worden en kunnen de effecten van verschillende modellen ter verkrijging van de afstandenmatrix met elkaar worden vergeleken.

Tenslotte bestaat de mogelijkheid naast single link en complete link andere (hiërarchische) methoden toe te passen.

Referenties

Estabrook, G.F., "A mathematical model in graph theory for biological classifications", Journal of Theoretical Biology, 12, 297-310 (1966).

Kent Kuiper, F. and Fisher, L., "A Monte Carlo Comparison of six clustering procedures", Biometrics 31, 777-783 (1975).

Ling, R.F., "A probability theory of cluster analysis", Journal of the American Statistical Association 68, 159-164 (1973).

Rand, W.M., "Objective criteria for the evaluation of clustering methods", Journal of the American Statistical Association 66, 846-850 (1971).

Wirth, M., Estabrook, G.F. and Rogers, D.J., "A graph theory model for systematic biology", Systematic Zoology 15, 59-69 (1966).