Competitiesplitsing van volledige grafen

W.C. BEEK en A. VOLGENANT*

Faculteit der Economische Wetenschappen en Econometrie
Vakgroep Actuariaat, Kwantitatieve Methoden & Econometrie
Leerstoelgroep Besliskunde
Universiteit van Amsterdam

Wanneer een sportcompetitie op een eerlijke wijze moet worden verdeeld in twee groepen, is één van de eisen vaak dat de beide groepen even groot zijn. Dit artikel behandelt een dergelijk probleem, formeel omschreven als het splitsen van een complete gewogen graaf in twee even grote complete subgrafen, zo dat de som van de gewogen kanten in de subgrafen geminimaliseerd wordt.

Het probleem vertoont overeenkomsten met het maximale-snedeprobleem, waarvan gebruik kan worden gemaakt om de NP-compleetheid van het probleem te bewijzen. Daarna worden een nearest neighbour heuristiek en twee nieuwe penaltyheuristieken beschreven en onderzocht op snelheid en kwaliteit. Uit de rekenresultaten voor meer dan 60 testproblemen uit de literatuur met probleemgroottes tot 150 blijkt een van de penaltyheuristieken het best te zijn. Op een standaard personal computer kost het bepalen van een oplosssing in alle gevallen hoogstens 1.5 sec.

---

* Vakgroep AKE
Roetersstraat 11, 1018 WB Amsterdam
tel. 020 5254219 (4217)
E-mail: tonv@fee.uva.nl

1 Inleiding

Clusteren is het verdelen van items in groepen. Het verschil tussen een tweetal items uit twee groepen is groter dan het verschil tussen een tweetal uit één groep. Classificatie of partitionering, zoals clustering ook genoemd wordt, wordt veel gebruikt. Voorwerpen die slechts in detail van elkaar afwijken krijgen dezelfde naam, worden op dezelfde manier behandeld en worden geacht zich op dezelfde manier te gedragen. In de biologie bijvoorbeeld worden dieren ondergebracht in groepen, zodat we kunnen spreken over zoogdieren, reptielen, insekten en vogels.

Praktische toepassingen waarin partitioneringsproblemen voorkomen zijn er legio, bijvoorbeeld op het gebied van vrachtwagenroutes, telefoonnetwerken – zie Murphy & Ignizio (1980) –, luchtvaartbemanningsschema's – zie Lavoie, Minoux & Odier (1988) –, politieke kiesdistricten en informatievoorziening.

Balas & Ho (1980), Marsten (1974), Fisher & Kedia (1990) en vele anderen behandelen partitioneringsproblemen ten opzichte van diverse doelstellingsfuncties. Uitgaande van een variabel dan wel een vast aantal clusters wordt geminimaliseerd naar een afstand dan wel naar een maximale afstand binnen een cluster of naar een afstand tussen de clusters.

Wij beschouwen clustering onder voorwaarden die het aantal items in een cluster èn het aantal clusters aan banden legt, te weten het geval waarin het aantal items binnen clusters gelijk is. Christofides & Brooker (1976) hebben exacte oplossingen gegeven voor het twee-cluster probleem met maximaal 40 items op dunne grafen (*dun* wil zeggen dat er relatief weinig kanten zijn tussen de knopen van de graaf). Voor een variant met partitionering in meer groepen geven Holm & Sørensen (1993) exacte oplossingen voor instanties met maximaal 20 items en 7 groepen. De heuristieken die in het vervolg aan de orde komen, zijn geschikt voor veel grotere probleeminstanties.

Wanneer een sportcompetitie op een eerlijke wijze moet worden verdeeld in twee groepen, is één van de eerste eisen dat de beide groepen even groot zijn. We noemen het probleem daarom het 2-Groeps-Competitie-Partitionerings-Probleem (2GCPP). Het kan beschreven worden als:

Gegeven: Een even aantal clubs gevestigd in plaatsen met gegeven onderlinge afstanden.
Gevraagd: Twee even grote groepen waarbinnen een volledige competitie afgewerkt wordt en wel zo, dat de totaal te reizen afstand van alle clubs tezamen minimaal is.

In de literatuur heeft De Werra (1985, 1988) aandacht geschonken aan de mogelijkheden van grafen bij het oplossen van indelingsproblemen voor sportcompetities. Voor specifieke sportcompetities is er ook ruime aandacht; zonder volledigheid te willen suggereren, noemen we onderzoek van Schreuder (1992) voor de voetbalcompetitie in Nederland, van Willis &

Terrill (1993) voor de cricketcompetitie in Australië en van Russell & Leung (1994) voor de honkbalcompetitie in de Verenigde Staten.

In de volgende sectie komt de NP-compleetheid van het door ons beschouwde probleem aan de orde, in sectie 3 verschillende heuristieken, in sectie 4 2-optimaliteit, terwijl in sectie 5 de rekenresultaten gegeven worden. We besluiten met conclusies.

Tot slot van deze sectie een conventie voor de notatie van sommaties van de gewichtsfunctie w: voor een verzameling van kanten X staat W(X) voor $\sum_{e \in X} w(e)$ en $\overline{W}(X)$ voor $\sum_{e \notin X} w(e)$ in graaf G; W' en $\overline{W}$' staan voor de vergelijkbare grootheden in graaf G'.

2 NP-COMPLEETHEID

Een omschrijving van het 2-groeps-competitie-partitioneringsprobleem luidt:

Instantie: Complete graaf G = (V, E), gewichten $w(e) \in \mathbb{N}$ voor elke $e \in E$, $|V| = 2n$, positieve integer K.

Vraag: Is er een partitie van V in twee subsets V_1 en V_2 met $|V_1| = |V_2| = n$ zo, dat $V_1 \cap V_2 = \varnothing$ en de som van de gewichten van de kanten van E met beide eindpunten in dezelfde subset van V niet groter is dan K ?

Een ja-antwoord van het 2GCPP bestaat uit knoopverzamelingen V_1 en V_2 die elkaar niet overlappen, zó dat $|V_1| = |V_2| = n$.

Omdat de grafen, geïnduceerd op V_1 en V_2, beide compleet zijn, zijn er in totaal $2 * (\frac{1}{2}n(n-1)) = n^2-n$ kanten met even zoveel kantgewichten. Het 2GCPP is $\in$ NP, omdat het aantal rekenstappen om een ja-antwoord te controleren, dat wil zeggen om te controleren of de som van die gewichten niet groter is dan een bepaalde constante K, wordt begrensd door een polynomiale functie van n.

Om het bewijs van NP-compleetheid te completeren geven we een polynomiale transformatie van het maximale snede probleem (Π_{ms}) naar het 2GCPP. Het probleem (Π_{ms}) wordt als volgt omschreven:

Instantie: Graaf G = (V, E), gewichten $w(e) \in \mathbb{N}$ voor elke $e \in E$, positieve integer K.

Vraag: Is er een verdeling van V in niet overlappende verzamelingen V_1 en V_2 zó, dat de som van de gewichten van de kanten die één eindpunt in V_1 en één eindpunt in V_2 hebben, minstens gelijk is aan K ?

Wanneer V' wordt gedefinieerd als de knopenset van het 2GCPP, ziet een polynomiale transformatie van Π_{ms} naar π_2 (een instantie van het 2GCPP) er als volgt uit:

1) Voeg zonodig een knoop toe aan V zodat $|V| = 2n$, (1 stap);
2) Verdubbel alle knopen zodat $|V'| = 4n$; noem deze graaf G' = (V', E'), (2n stappen);
3) Maak G' = (V', E') compleet door maximaal $2n(4n-1)$ kanten toe te voegen en geef deze kanten gewicht 0, $(2n(4n-1)$ stappen);

4) Geef K' de waarde $W(E') - K$, (1 stap).

Deze transformatie van (Π_{ms}) naar *een* instantie van het 2GCPP (π_2), in $8n^2 + 2$ stappen, wordt dus begrensd door een polynomiale functie in n. Daarna bestaat π_2 uit: complete graaf $G'(V', E')$, $|V'| = 4n$, gewichten $w(e) \in \mathbb{N}$ voor elke $e \in E'$, positieve integer K'.

Rest nog te bewijzen dat π_2 een ja-antwoord heeft dan en slechts dan als Π_{ms} een ja-antwoord heeft. Zij E_1 en E_2 (V_1 en V_2) de verzamelingen kanten (knopen) na splitsing van $G(V, E)$ in $G_1(V_1, E_1)$ en $G_2(V_2, E_2)$.

Een ja-antwoord van π_2 impliceert een ja-antwoord van Π_{ms}:

Stel er bestaat een ja-antwoord voor π_2 zijnde V_1' en V_2' ($|V_1'| = |V_2'| = 2n$) met $W'(E_1 \cup E_2) \leq W(E') - K = W'(E)$. Doordat de kanten $e \in E' \setminus E$ gewicht nul hebben is de maximale snedeverzameling in graaf G' dezelfde als die in graaf G, die met kanten uit $E' \setminus E$ is aangevuld; dus:

$$\bar{W}(E_1 \cup E_2) = \bar{W}'(E_1 \cup E_2).$$

Aangezien $E' = \{\text{maximale snedeverzameling in } G'\} \cup E_1' \cup E_2'$ volgt dat de maximale snedeverzameling in G gelijk is aan $\{e \mid e \in E \cap (E' - E_1' - E_2')\}$ en ook dat

$$W'(E) = \bar{W}'(E_1 \cup E_2) + W'(E_1 \cup E_2);$$

dus, omdat $W'(E_1 \cup E_2) \leq W'(E) - K$ geldt: $\bar{W}(E_1 \cup E_2) \geq K$, met als conclusie dat een ja-antwoord voor π_2 een ja-antwoord impliceert voor Π_{ms}.

Een ja-antwoord van Π_{ms} impliceert een ja-antwoord van π_2:

Stel er bestaat een ja-antwoord voor Π_{mc} in de vorm van V_1 en V_2 met $\bar{W}(E_1 \cup E_2) \geq K$. Door eventueel een extra knoop aan V_1 toe te voegen (zodat $|V_1 \cup V_2| = 2n$) en vervolgens elke knoop in V_1 en V_2 te verdubbelen vormen we $\tilde{V}_1$ en $\tilde{V}_2$ ($|\tilde{V}_1 \cup \tilde{V}_2| = 4n$).

Als $|\tilde{V}_1| = |\tilde{V}_2| = 2n$ hebben we ook een ja-antwoord voor π_2.

Als $|\tilde{V}_1| \neq |\tilde{V}_2|$, stel zonder verlies der algemeenheid dat $|\tilde{V}_1| > |\tilde{V}_2|$, dan kunnen we, zonder dat $\bar{W}(E_1 \cup E_2)$ of $\bar{W}'(E_1 \cup E_2)$ verandert, $|\tilde{V}_1|$ gelijk maken aan $|\tilde{V}_2|$ door dummyknopen uit $\tilde{V}_1$ over te hevelen naar $\tilde{V}_2$. Aan deze knopen liggen namelijk alleen kanten met gewicht nul. In het uiterste geval is $|\tilde{V}_1| = 4n$, namelijk als alle gewichten, ook in G, nul zijn. Dan moeten er 2n knopen worden overgeheveld wat mogelijk is, omdat er 2n dummyknopen aangemaakt zijn door 2n knopen te verdubbelen. Op deze wijze verkrijgen we $|\tilde{V}_1| = |\tilde{V}_2| = 2n$.

Omdat $\bar{W}(E_1 \cup E_2)$ en dus $\bar{W}'(E_1 \cup E_2)$ door deze overheveling niet veranderen, geldt ook hier: $W'(E_1 \cup E_2) \leq W'(E) - K$ en $V_1' = \tilde{V}_1$ en $V_2' = \tilde{V}_2$.

Dus een ja-antwoord voor Π_{mc} impliceert ook een ja-antwoord voor π_2, waarmee bewezen is dat het 2GCPP NP-compleet is.

3 HEURISTIEKEN

Om een dataset op te delen in een willekeurig aantal groepen, is een niet-hiërarchische clusteringmethode nodig. Hiervan zijn er vele beschreven in de literatuur. De centrale idee in de meeste van deze methodes is het kiezen van een initiële verdeling van de data-items om vervolgens door een verwisseling van items tot een betere oplossing te komen. De initiële verdeling komt tot stand vanuit een verzameling beginpunten, vanaf nu *seedpoints* genoemd, waar de clusters omheen worden gevormd. Er zijn verschillende manieren om de seedpoints te bepalen.

De meeste clusteringmethodes houden geen rekening met de grootte van de clusters. Daarom zullen bestaande heuristieken moeten worden aangepast voor het 2GCPP. Deze aanpassingen zullen zo vroeg mogelijk in een methode worden ingebouwd teneinde geen tijd te verspillen met eerst het berekenen van een kwalitatief goede oplossing zonder gelijke groepen en vervolgens het gelijk maken van de groepen.

3.1 Seedpoints

We beschouwen drie manieren om de benodigde twee seedpoints te bepalen:

methode 1: Anderberg (1973)

neem de eerste twee items uit de dataset; deze methode is simpel en snel, maar niet erg doordacht.

methode 2:

kies de twee items die het verst van elkaar af liggen; daardoor wordt in de meeste gevallen tegemoet gekomen aan de wens, dat de seedpoints in een optimale verdeling niet bij elkaar in één groep zitten. De rekentijd hiervan is $O(n^2)$.

methode 3: Anderberg (1973)

deze methode, een uitwerking van de methode van Astrahan, kiest de seedpoints als volgt:

a. Bereken de 'dichtheid' van ieder punt door het aantal punten te tellen dat binnen een vaste straal d van dit punt ligt;
b. Orden de punten naar afnemende dichtheid;
c. Kies twee seedpoints op basis van de afnemende dichtheid zo, dat de punten met de hoogste dichtheid als seedpoints gekozen worden, maar dat de afstand tussen de seedpoints groter is dan een gegeven vaste waarde.

Duidelijk is dat de seedpoints zo centraal mogelijk in de clusters moeten liggen. In een optimale verdeling liggen de clustergemiddelden $(\bar{x}_1, \bar{y}_1)$ en $(\bar{x}_2, \bar{y}_2)$ het meest centraal. Het is dus belangrijk om de initiële seedpoints dicht in de buurt van die gemiddelden te kiezen. We weten van die clustergemiddelden dat ze op één lijn liggen met het algemene gemiddelde $((\bar{X}, \bar{Y}))$ en beide even ver van $(\bar{X}, \bar{Y})$, omdat beide clusters even groot zijn.

Deze argumenten gecombineerd met de methode van Astrahan leveren de volgende procedure op:

1. Bereken $(\bar{X}, \bar{Y})$, het algemene datagemiddelde, O(n);
2. Bereken de dichtheid van alle punten door het aantal punten te tellen dat binnen een straal d van zo'n punt ligt, $O(n^2)$;
3. Kies als seedpoint 1 het punt met de grootste dichtheid, O(n);
4. Maak een puntspiegeling van dit punt in $(\bar{X}, \bar{Y})$; noem dit Z;
5. Kies als seedpoint 2 het punt binnen een straal d om Z met de hoogste dichtheid.

De keuze van d in deze methodes vereist een zeker inzicht in de datastructuur. Bij een te grote waarde van d zullen alle dichtheden enorm groot zijn en zal er weinig onderscheidend vermogen van uitgaan. In het andere uiterste geval is de dichtheid 1 voor ieder punt. Ergens daartussen bevindt zich een beste waarde die de seedpoints het dichtst bij de clustergemiddelden plaatst. Empirisch blijkt $d = \frac{1}{2} *$ (de gemiddelde afstand tussen alle punten) goed te voldoen. Die benadering wordt beter naarmate de coördinaten van de punten een meer uniform verdeeld karakter hebben. De totale rekentijd van deze methode om de seedpoints te bepalen is $O(n^2)$.

Voor datasets bestaande uit meerdere, ver van elkaar verwijderde puntenwolken kan het beter zijn als initiële seedpoints twee punten te kiezen in twee verschillende puntenwolken.

3.2 Nearest neighbour

Als eerste beschouwen we een voor de hand liggende heuristiek, namelijk de nearest neighbour heuristiek, kortweg NN-heuristiek. Een data-item wordt toegewezen aan het cluster, wiens seedpoint (verkregen met een willekeurige methode) het dichtst in de buurt van het data-item ligt. De scheiding tussen clusters bestaat dan uit rechte lijnen, omdat de verzameling punten op gelijke afstand van twee seedpoints meetkundig gezien gegeven wordt door een rechte lijn loodrecht op de verbindingslijn van de seedpoints.

Nadat alle items zijn toegewezen, worden de nieuwe seedpoints berekend als de gemiddelden van de zojuist gevormde clusters. Vervolgens worden alle items opnieuw toegewezen aan de nieuwe seedpoints. Deze stap wordt herhaald tot de clusters niet meer veranderen. Anderberg (1973) heeft bewezen dat clusteren op deze manier convergeert.

Om te voorzien in clusters met een gelijk aantal items, is de volgende aanpassing toegepast. Wanneer één van beide clusters vol is, wordt de reguliere toewijzing gestaakt. Van de resterende items wordt één voor één bekeken hoe het gewicht van het volle cluster verandert als een resterend item wordt verwisseld met de items uit het volle cluster. Als het gewicht afneemt, worden de twee items met elkaar verwisseld en gaat het item uit de volle

groep over naar de andere groep. Op deze wijze bestaan de twee groepen uiteindelijk uit een gelijk aantal items. De rekentijd van deze methode bedraagt $O(n^2)$ per iteratie.

In tegenstelling tot de gewone nearest neighbour heuristiek convergeert de aangepaste versie niet altijd. Datasets waarbij zich dit voordoet hebben bijvoorbeeld een verdeling die bestaat uit 1 of 2 lijnen waarop de meeste data-items zich bevinden of vertonen een 'printplaat' structuur: datasets die bestaan uit zich herhalende gestructureerde patronen.

Uit de rekenresultaten blijkt dat de samenstelling van de groepen een cyclus doorloopt van 2, 3 of 4 iteraties. Daarvan is de tweevoudige cyclus ondervangen door ook op de overeenkomst van de nieuwe seedpoints met de voorgangers van de oude te controleren.

Het is duidelijk dat deze methode geometrisch georiënteerd is en bij uitstek geschikt om euclidische problemen op te lossen. Of de heuristiek ook geschikt is voor andere typen problemen, hebben we niet nagegaan.

3.3 Penaltyheuristieken

In deze nieuwe clusteringmethode wordt vanaf de start gelet op het verkrijgen van even grote groepen. De toewijzing van items aan de clusters vindt plaats op basis van penalties.

Eerst beschouwen we penalties op basis van de groepsafstand. De volgende definities zijn dan van belang:

- de *groepsafstand* van cluster i is de som van alle afstanden tussen alle items in cluster i.
- de *penalty* van een item dat nog niet toegewezen is:
  a. als beide clusters even groot zijn:
     de absolute waarde van het verschil tussen zijn bijdrage aan groepsafstand 1 en groepsafstand 2 als het in cluster 1 respectievelijk cluster 2 zou worden geplaatst;
  b. als cluster i kleiner is dan cluster k:
     (de bijdrage aan groepsafstand k) minus (de bijdrage aan groepsafstand i).

De heuristiek, kortheidshalve aangeduid als PG-heuristiek, luidt als volgt:

1. Start met 2 clusters gevormd door seedpoints 1 en 2 die verkregen zijn met een in sectie 3.1 beschreven methode.
2. Bereken de penaltywaarde voor alle items die nog niet zijn toegewezen en bepaal de grootste penaltywaarde; O(n).
3. Als cluster i kleiner is dan cluster k:
   voeg het item met de grootste penaltywaarde toe aan cluster i;
   als beide clusters even groot zijn:
   voeg het item met de grootste penaltywaarde toe aan het cluster i waar het de kleinste bijdrage levert aan groepsafstand i;

als er gelijke penaltywaarden zijn:

voeg van die items met gelijke penaltywaarden het item toe aan cluster i dat als eerste is gecontroleerd.

4. Herhaal de stappen 2 en 3 totdat alle items zijn toegewezen; $O(n^2)$.
5. Bereken de nieuwe clustergemiddelden en gebruik deze als de seedpoints 1 en 2; O(n).
6. Herhaal de stappen 2 t /m 5 totdat de clusters niet meer veranderen.

We merken op dat deze methode rekenintensief lijkt, maar dat de rekentijd van een iteratie slechts $O(n^2)$ is. Zo is in stap 3 na elke toewijzing groepsafstand 1 en groepsafstand 2 bij te houden door de bijdrage op te tellen bij de laatst berekende groepsafstand. Het berekenen van de bijdrage aan de groepsafstand is O(n), maar dit heeft al plaatsgevonden bij stap 2.

Vervolgens beschouwen we penaltywaarden te berekenen op basis van het verschil in afstand tot de seedpoints; deze waarden zijn met minder rekenwerk te berekenen. Deze variant is te beschouwen als een nearest neighbour penaltyheuristiek, af te korten als PN-heuristiek. Het voordeel van deze variant ten opzichte van de PG-heuristiek is de gereduceerde rekentijd per iteratie; immers het uitrekenen van de afstand tot een seedpoint is een O(n) sneller dan het berekenen van de bijdrage aan de groepsafstand, wat zelf O(n) is.

3.4 Convergentie of stopcriterium

Volgens de theorie moeten de iteraties van de beschouwde heuristieken worden herhaald totdat de groepsindeling niet meer verandert. Uit praktisch oogpunt worden de seedpoints van verschillende iteraties met elkaar vergeleken en worden de heuristieken afgebroken zodra de seedpoints niet meer veranderen.

Proefondervindelijk is gebleken dat de NN-heuristiek niet altijd convergeert, zodat voor die heuristiek het stop-criterium is aangepast. Omdat de samenstelling van de groepen vaak een cyclus doorloopt van een aantal iteraties, wordt zowel op de overeenkomst van het nieuwe seedpoint met het oude gecontroleerd, als op de overeenkomst met de voorganger van het oude seedpoint. Convergenties kunnen vaker worden bewerkstelligd door ook op overeenkomst met de voorganger van de voorganger en diens voorganger te controleren. Dit is echter met het oog op snelheid en overzichtelijkheid nagelaten.

Hoewel een bewijs van convergentie niet gegeven wordt blijkt de PG-heuristiek voor alle geteste problemen te convergeren. Zoals verwacht blijkt ook uit de resultaten dat er voor de PN-heuristiek meer iteraties nodig zijn alvorens convergentie optreedt, voor zover deze überhaupt plaatsvindt, omdat het toewijzingscriterium verder van de doelstellingsfunctie (een kleinste gewicht) afstaat.

4 2−Optimaliteit

Wanneer een gegeven oplossing niet te verbeteren is door twee items uit verschillende groepen met elkaar te verwisselen is deze *2−optimaal*.

Een willekeurige oplossing is 2−optimaal te maken door in de oplossing naar zo'n items-paar op zoek te gaan en indien dat aanwezig is, zulke items te verwisselen; door dit te herhalen tot zo'n paar niet meer bestaat wordt een 2−optimale oplossing verkregen.

De rekentijd van deze verbeteringsmethode bedraagt minstens $O(n^2)$: er zijn $(\frac{1}{2}n)^2$ items-paren waarvoor per paar een constante rekentijd nodig is om te controleren of de oplossing te verbeteren valt. Per verbetering zijn O(n) rekenstappen nodig om verschillende variabelen aan te passen. De rekentijd van deze methode hangt dus van het aantal verbeteringen af.

In het algemeen neemt het aantal verbeteringen toe naarmate de onderzochte oplossing slechter is. Om die reden is de methode niet in de heuristieken verwerkt. Wel is de methode gebruikt om verbeterde oplossingen te berekenen waarmee de oplossingen van de heuristieken kunnen worden vergeleken.

In het geval langere rekentijden aanvaardbaar zijn, dan wel dat betere oplossingen hoge prioriteit hebben, valt te overwegen *k−optimale* (k even) oplossingen te bepalen, dat wil zeggen verwisselingen van k items tegelijk. Nadere rekenresultaten zijn nodig om na te gaan in hoeverre de kwaliteit van de oplossingen toeneemt ten koste van hoeveel extra rekentijd.

5 Rekenresultaten

Door de drie methodes uit sectie 3.1 om de seedpoints te berekenen te combineren met de drie heuristieken beschreven in de secties 3.2 en 3.3, ontstaan 9 verschillende heuristieken. Een instantie van het handelsreizigersprobleem heeft dezelfde vorm als het 2GCPP en test-problemen voor het handelsreizigersprobleem zijn daarom geschikt om de heuristieken op te toetsen. Deze zijn getest bij verschillende probleemgroottes (n = 50, 100 en 150) met respectievelijk 28, 27 en 20 willekeurige testproblemen, welke afkomstig zijn uit de problemen-bibliotheek TSPLIB, Reinelt (1991). De daaruit afkomstige testproblemen zijn aangepast door weglating van steden.

De gebruikte kwaliteitsnormen zijn de rekentijd, het aantal benodigde iteraties, de afwijking ten opzichte van de best bekende oplossing of een ondergrens en het aantal malen dat de heuristiek een beste oplossing geeft.

Bij het testen van de heuristieken is gebruik gemaakt van een personal computer met een 486 DX2 - 66 MHz processor; tijden zijn vermeld in tienden van seconden.

Het bepalen van een goede ondergrens voor het 2GCPP blijkt niet eenvoudig te zijn. Bij n = 10 en 20 kan binnen een aanvaardbare tijd (circa 1½ minuut bij n = 20) door alle mogelijke oplossingen met elkaar te vergelijken (*complete enumeration*) een optimale

oplossing bepaald worden, $O(n^{n-1})$. Voor n = 50, 100 en 150 is dit (praktisch) onmogelijk: voor een probleem met 50 steden zou met de gebruikte processor een tijd van ruim 3 millennia nodig zijn !

Zoals eerder genoemd levert een maximale snede een uitstekende ondergrens op. Het maximale-snedeprobleem is zelf echter NP-compleet en daarvoor is dus nog geen polynomiale oplossingsmethode gevonden.

Soms wordt de oplossing vergeleken met de laagst verkregen oplossing van de andere methodes. We hebben gekozen voor een betere vergelijking, namelijk met de beste 2−optimale oplossing (zie sectie 4). In tabel 1 zijn de resultaten vermeld. De gemiddelde afwijking van een oplossing ten opzichte van deze beste 2−optimale oplossing (de *gap*) wordt alleen berekend over het aantal oplossingen dat van deze getallen afwijkt. De rekentijd (in tienden van seconden) en het aantal iteraties zijn een gemiddelde over de testproblemen van een bepaalde omvang; bij het aantal iteraties is eventueel als superindex vermeld het aantal malen dat de methode niet binnen 250 iteraties convergeert.

Wanneer gekeken wordt naar de *kwaliteit van de oplossing* ten opzichte van de optimale of beste 2−optimale waarde blijkt de penaltyheuristiek op basis van groepsafstand de beste resultaten op te leveren. Deze heuristiek is daarbij in hoge mate onafhankelijk van de gebruikte seedpoints bepalingsmethode, wat eenvoudig is te verklaren. Waar de andere heuristieken afhankelijk zijn van de volgorde van de data-items, is deze dat niet omdat alle items afgelopen worden alvorens een item aan een cluster wordt toegevoegd.

Het lijkt erop dat de drie penaltyheuristieken op basis van groepsafstand minder goed presteren als veel data-items onderling ongeveer dezelfde afstand hebben, dat wil zeggen datasets die binnen het vlak waarin ze zich bevinden op een uniforme wijze zijn verdeeld. Het onderscheidende vermogen van de penaltywaarden neemt in die gevallen af, zodat minder goede beslissingen genomen worden. De kwaliteit van de verkregen oplossingen blijft evenwel acceptabel.

Een datastructuur die bestaat uit twee of meer niet even grote en ver van elkaar verwijderde puntenwolken geeft moeilijkheden voor de heuristieken die gebruik maken van de seedpointmethode gebaseerd op de dichtheden van de punten. Doordat *d*, de straal van de cirkel om een punt op basis waarvan de dichtheden worden bepaald, nadelig wordt beïnvloed door de afstanden tussen de puntenwolken en daarom relatief te hoog zal zijn, is het onderscheidende vermogen van de dichtheden klein. Bovendien zal seedpoint 2, nadat seedpoint 1 is gekozen, niet snel in de andere puntenwolk worden geplaatst omdat de afstand tussen seedpoint 1 en de puntenwolken tijdens de keuzeronde voor seedpoint 2 niet wordt overbrugd. Deze situaties kunnen verholpen worden door *d* aan te passen. Overigens komen deze onvolkomenheden bij de aangepaste dichtheidsmethode minder vaak voor.

seedpointbepaling	eerste 2 uit set			maximale afstand			maximale dichtheid		
aantal items 50	heuristiek			heuristiek			heuristiek		
28 testproblemen	NN	PG	PD	NN	PG	PD	NN	PG	PD
aantal beste	2	25	1	3	24	1	2	25	1
gap (in %)	10.2	1.2	12.6	7.4	1.5	11.2	7.9	1.2	12.6
tijd (in 0.1 sec)	0	1	2	1	1	1	1	1	2
aantal iteraties	4.8	2.7	23.2 [2]	12.6 [1]	2.1	5.5	20.4[2]	2.7	23.2[2]
aantal items 100 27 testproblemen									
aantal beste	2	23	2	2	23	2	2	24	2
gap (in %)	10.2	2.0	13.8	9.9	1.7	14.2	10.5	0.6	14.7
tijd (in 0.1 sec)	1	4	2	1	3	2	2	3	3
aantal iteraties	4.9	2.6	5.9	4.4	2.0	5.9	13.3[1]	2.2	5.5
aantal items 150 20 testproblemen									
aantal beste	2	18	0	1	19	1	2	18	1
gap (in %)	8.9	2.2	11.8	8.5	0.0	10.6	8.7	1.2	10.9
tijd (in 0.1 sec)	4	8	14	4	7	14	5	10	15
aantal iteraties	30.2[2]	2.7	18.4	28.8	2.0	17.9	28.8[2]	2.3	18.0[1]

bij aantal iteraties geeft [1] of [2] aan hoe vaak de methode niet binnen 250 iteraties convergeert

Tabel 1: Rekenresultaten voor de heuristieken

Op basis van *snelheid* verslaat de combinatie van de eerste twee items als seedpoints met de aangepaste nearest neighbour heuristiek alle andere heuristieken. Voor problemen met minder dan 50 items (waarvoor de resultaten wel zijn bepaald maar niet zijn vermeld) kan echter beter gebruik gemaakt worden van de penaltyheuristieken omdat de bijbehorende rekentijden niet veel hoger zijn (0.1 sec) en de afwijking bij de nearest neighbour heuristiek erg groot wordt.

Waar alle methodes bij kleine probleemomvang nog redelijk scoren is dat bij grotere omvang niet meer het geval. Dan blijkt de kracht van *betere toewijzingscriteria*, zoals bij de penaltyheuristieken. De rekentijden zijn overigens van alle methoden zo kort, dat overwogen kan worden verscheidene methoden te benutten en de best verkregen oplossing te gebruiken.

6 Conclusies

De **beste oplossingsmethode** voor het competitie partitioneringsprobleem van de onderzochte heuristieken is ongetwijfeld de penaltyheuristiek op basis van groepsafstand met initiële seedpoints bepaald met de aangepaste dichtheidsmethode (zie sectie 3.3). Daarbij geeft deze heuristiek vaak de laagst bekende 2−optimale oplossing (91% voor 50, 100 en 150 clubs), met slechts een kleine afwijking van gemiddeld 0.7% wanneer die oplossing niet gevonden wordt. De rekentijd (voor een computer met 486 DX2-66 MHz processor) ligt rond 0.1 sec voor problemen met 50 clubs tot 0.8 sec voor problemen met 150 clubs.

Voor **snelheideisende** competitie partitioneringsproblemen, waarbij de kwaliteit van de oplossingen op de tweede plaats komt, is de nearest neighbour heuristiek met als initiële seedpoints de eerste twee items uit de dataset het meest geschikt voor problemen met 50, 100 en 150 steden. Bij zulke problemen bedraagt de rekentijd gemiddeld 0.1 sec en de afwijking meer dan 10%.

Voor problemen met minder dan 50 steden kan beter gebruik gemaakt worden van de penaltyheuristieken, omdat ten eerste de rekentijd niet veel hoger is (0.1 sec) maar vooral omdat de afwijking van de nearest neighbour heuristiek erg groot wordt.

Het is een onjuiste gedachte dat het 2GCPP alleen te gebruiken zou zijn om een sport-competitie in tweeën te delen. Wanneer niet-euclidische afstanden genomen worden en er meer dan twee groepen gevormd worden, dan is het toepassingsgebied aanmerkelijk uitgebreid. Van de beschreven heuristieken verwachten we dat de penaltyheuristieken op basis van groepsafstand het geschiktste zijn om niet-euclidische problemen goed op te lossen. Nadere rekenresultaten kunnen deze verwachting onderbouwen.

Op verzoek is het computerprogramma beschikbaar.

7 Referenties

Anderberg, M.R., 'Cluster Analyses for Applications', *Academic Press*, 1973.

Balas, E., & C. Ho, 'Set Covering Algorithms Using Cutting Planes, Heuristic and Subgradient Optimization: A Computational Study', *Mathematical Programming* 12 (1980) 37-60.

Christofides, N. & P. Brooker, 'The optimal partitioning of graphs', *SIAM Journal of Applied Mathematics* 30 (1976) 55-69.

Fisher, M.L. & P. Kedia, 'Optimal Solution of Set Covering / Partioning Problems using Dual Heuristics', *Management Science* 36 (1990) 674-688.

Garey, M.R. & D.S. Johnson, 'Computers and Intractability: A Guide to the Theory of NP-Completeness', *Freeman, San Francisco*, 1979.

Holm, S. & M.M. Sørensen, 'The optimal graph partitioning problem, Solution method based on reducing symmetric nature and combinatorial cuts', *ORSpektrum* 15 (1993) 1-8.

Lavoie, S, M. Minoux & E. Odier, 'A New Approach for Crew Pairing Problems by Column Generation with an Application to Air Transportation', *European Journal of Operational Research* 35 (1988) 45-58.

Marsten, R.E., 'An Algorithm for Large Set Partitioning Problems', *Management Science* 20 (1974) 779-787.

Murphy, C.M. & J.P. Ignizio, 'A Methodology for Multicriteria Network Partitioning', *Computers & Operations Research* 11 (1984) 1-11.

Reinelt, G., 'TSPLIB - A Traveling Salesman Problem Library', *ORSA Journal on Computing* 3 (1991) 376-385.

Russell, R.A. & J.M.Y. Leung, 'Devising a cost effective schedule for a baseball league', *Operations Research* 42 (1994) 614-625.

Schreuder, J.A.M., 'Combinatorial aspects of construction of competition in dutch professional football leagues', *Discrete Applied Mathematics* 35 (1992) 301-312.

De Werra, D., 'On the multiplication of divisions: the use of graphs for sports scheduling', *Networks* 15 (1985) 125-136.

De Werra, D., 'Some models of graphs for scheduling sports competitions', *Discrete Applied Mathematics* 21 (1988) 47-65.

Willis, R.J. & B.J. Terrill, 'Scheduling the Australian state cricket season using simulated annealing', *Journal of the Operational Research Society* 45 (1993) 276-280.

Ontvangen: 15-6-1995
Geaccepteerd: 23-8-1996