Stochastisch koelen en de voetbalpool

Voordracht gehouden in de gewone vergadering van de Afdeling Natuurkunde der Koninklijke Nederlandse Akademie van Wetenschappen op 27 april 1992

STOCHASTISCH KOELEN EN DE VOETBALPOOL

J.H. van Lint

Faculteit Wiskunde en Informatica, TU Eindhoven

Inleiding

De rol van de wiskundige discipline die bekend staat als *optimalisering* is in de laatste 40 jaar voortdurend toegenomen. Er zijn industriële toepassingen op allerlei gebieden. Vooral wat betreft de lineaire programmering is er in de laatste tijd veel gebeurd, met o.a. de opvolging van het simplex algoritme door de ellipsoïde methode en het algoritme van Karmarkar.

Vooral van *discrete* optimaliseringsproblemen is bekend dat ze zeer lastig zijn. Hier is het meest bekende voorbeeld het zgn. *handelsreizigerprobleem*. Bij dit probleem zijn een collectie steden gegeven, met diverse verbindingswegen en hun lengte (de "kostenfunctie"). Het doel is om een route te vinden waarlangs de reiziger alle steden één maal bezoekt en er voor te zorgen dat de afgelegde weg minimale lengte heeft. Om rekentijd te sparen neemt men vaak genoegen met een redelijke benadering van het optimum bij dit soort problemen. Ook met snelle computers is het vinden van een absoluut minimum of maximum voor combinatorische problemen veel te tijdrovend.

Het type optimaliseringsproblemen waar ik het in deze voordracht over wil hebben wordt beschreven door het volgende model. Gegeven is een verzameling *toestanden* (of configuraties, zoals een mogelijke route voor de handelsreiziger) en een *kostenfunctie* die aan iedere toestand een waarde toekent. Er is een afstandsbegrip gedefinieerd waardoor men kan spreken over een "omgeving" van een toestand bestaande uit alle toestanden die voldoende dicht bij de gegeven toestand liggen. Men wil de minimale (of maximale) waarde van de kostenfunctie vinden. De meest simpele algoritmen beginnen op een willekeurig gekozen punt, zoeken in een omgeving naar een betere waarde, gaan daar naartoe en herhalen het proces. Het is duidelijk dat dit lokale zoeken eindigt als men in een lokaal optimum is aangeland. Betere resultaten kunnen worden verkregen door het algoritme vele malen uit te voeren of door in het begin met grote omgevingen te werken en die tijdens de uitvoering langzaam kleiner te laten worden, enz.

In het laatste decennium zijn er diverse zoekalgoritmen ontwikkeld die een stochastisch element bevatten. Ze zijn vaak zeer succesvol. Het idee is ruwweg het volgende. In het vorige algoritme werden alleen stappen gedaan van een toestand naar een andere als daarbij de waarde van de kostenfunctie "beter" werd. Bij de stochastische algoritmen is er een probabilistisch mechanisme dat met een in de tijd afnemende waarschijnlijkheid ook stappen in de verkeerde richting accepteert. Zo is er in het begin de mogelijkheid om uit een lokaal optimum te ontsnappen. De variant die we hier beschouwen is ontleend aan een fysisch model.

Simulated annealing

Annealing (langzaam afkoelen) is het fysische proces waarbij een vaste stof tot smelten wordt gebracht en daarna zeer langzaam afgekoeld. In de vloeibare toestand vormen de deeltjes een random configuratie. Bij het koelen wordt bij iedere temperatuur T op thermisch evenwicht gewacht voor verder wordt gekoeld. Het evenwicht wordt beschreven door de *Boltzmann distributie*: de waarschijnlijkheid dat de stof zich in een toestand met energie E bevindt is

$$\mathrm{Prob}(\mathbf{E} = E) = \frac{1}{Z(T)} \cdot \exp\left(-\frac{E}{k_B T}\right),$$

waarin k_B de constante van Boltzmann is en $Z(T)$ een normaliserende functie (om de totale waarschijnlijkheid 1 te laten zijn). Als T afneemt worden toestanden met lage E steeds waarschijnlijker en tenslotte hebben alleen toestanden met minimale energie nog een positieve waarschijnlijkheid. Bij zeer snel afkoelen kan van alles en nog wat gebeuren. Voor dit verhaal is dat onbelangrijk.

Door Metropolis et al. (1953) is een Monte Carlo methode ontwikkeld om het fysisch proces bij vaste T te simuleren. De toestand van de stof wordt gekarakteriseerd door de positie van alle deeltjes. Een random deeltje wordt gekozen en dit wordt over kleine afstand verplaatst. Laat ΔE de verandering in energie zijn bij deze perturbatie. Als deze negatief is (de nieuwe toestand heeft lagere energie) dan gaan we gewoon verder; zo niet, dan wordt de nieuwe toestand geaccepteerd met een waarschijnlijkheid

$$\exp(-\underline{\Delta E}).$$

Men kan aantonen dat zo de toestandsverdeling inderdaad tot de $k_B T$ Boltzmann distributie nadert.

Kirkpatrick et al. (1982) hebben een combinatorisch optimaliseringsalgoritme bedacht dat het idee van het Metropolis algoritme vertaalt naar het wiskundig probleem. De toestanden zijn de boven genoemde collectie R, de energie wordt vervangen door de kostenfunctie en er wordt een parameter geïntroduceerd die *temperatuur* heet. Men moet ook nog regelen hoe snel het koelproces zal gaan. Daartoe moet worden vastgelegd hoeveel stappen worden gedaan bij elke (vaste) temperatuur en volgens welke regel een temperatuur wordt vervangen door een lagere. We geven hieronder een voorbeeld van deze zgn. *simulated annealing* (stochastisch koelen).

Algoritme

Gegeven zijn een ruimte R van toestanden, een definitie van *"buren"* van een gegeven toestand, en een kostenfunctie c voor de toestanden.

1. Parameter β is de *temperatuur*; kies beginwaarde β_0.
2. Kies parameter L. Dit is het aantal *pogingen* om de toestand te veranderen per *serie* (uitgevoerd bij vaste β).
3. Kies afbreekparameter h: als in h opeenvolgende series de toestand steeds niet verandert, dan eindigt het proces.
4. Kies de "koelfactor" α. Neem als afkoelregel (na een serie van L pogingen) : $\beta' = \alpha\beta$.
5. Definitie van een poging:
   (a) kies een random punt in R voor de eerste poging; anders het laatst geaccepteerde punt;
   (b) kies een random buur van dit punt;
   (c) bereken de verandering Δc van de kostenfunctie;
   (d) laat een random generator een getal y kiezen in [0,1);
   (e) accepteer het punt uit (b) als $\exp(-\frac{\Delta c}{\beta}) > y$ en anders blijft het punt uit (a) het laatst geaccepteerde punt.

Voor alle details: zie Van Laarhoven en Aarts (1987).

Het "football-pool" probleem

In een echte voetbal pool voorspelt men de uitslag van een aantal, zeg n, voetbalwedstrijden. We gebruiken 0 voor een gelijk spel, 1 als de thuisploeg wint en 2 als de bezoekers winnen. Eén zo'n voorspelling van een deelnemer is dus een *rijtje* $x = (x_1, x_2, \ldots, x_n)$, met $x_i \in \{0,1,2\}$, $(i = 1,2,\ldots,n)$. Om de eerste prijs te winnen moet men alle voorspellingen goed hebben. Wil een deelnemer absolute zekerheid hebben dat hij de eerste prijs zal winnen, dan moet hij 3^n rijtjes insturen. Voor de tweede prijs mag men één foute voorspelling hebben. Een domme manier om te garanderen dat men een rijtje heeft met ten hoogste één fout is om alle rijtjes die met 0 beginnen in te sturen. Dat zijn er dus 3^{n-1}. Het feit dat het veel beter kan, maakt dit tot een interessant wiskundig probleem. Gevraagd is de waarde van $\delta(n,3)$, het minimale aantal rijtjes dat men moet insturen bij een voetbal pool met n wedstrijden om zeker te zijn dat er tenminste één bij is met niet meer dan één foute voorspelling. Het algemene wiskundige probleem wordt als volgt geformuleerd. Laat Q een alfabet zijn met k symbolen en beschouw de collectie $R := Q^n$, d.w.z. de verzameling van alle rijtjes van n symbolen uit Q. In R voeren we een afstandsfunctie d in door

$$d(\mathbf{x},\mathbf{y}) := |\{i : x_i = y_i,\ 1 \leq i \leq n\}| ,$$

dat is het aantal posities waarin de rijtjes verschillen. We zeggen dat een deelverzameling C van R de verzameling R overdekt als er voor iedere $\mathbf{x} \in R$ tenminste één $\mathbf{c} \in C$ is met $d(\mathbf{x},\mathbf{c}) \leq 1$. Dan is

$$q(n,k) := \min \ \{ \ | C | \ : C \text{ overdekt } R\}.$$

We geven eerst een voorbeeld dat aantoont dat er subtiele overdekkingen bestaan. Voor 4 wedstrijden weten we dat met $27=3^3$ rijtjes een triviale overdekking mogelijk is. Wat is minimaal nodig? Elk rijtje is goed voor een prijs als alles klopt of één van de wedstrijden één van de twee andere mogelijke uitslagen heeft. Samen zijn dit 1 + 4 . 2 = 9 mogelijke series van vier uitslagen.
Daar in totaal 3^4 = 81 series a priori mogelijk zijn ziet men direct dat $\delta(4,3) \geq 81/9 = 9$. Dat hier gelijkheid geldt berust op de beroemde perfecte fouten-verbeterende ternaire [4,2,3] *Hamming code*. We geven hieronder het schema:

0	0	0	0
0	1	1	1
0	2	2	2
1	0	1	2
1	1	2	0
1	2	0	1
2	0	2	1
2	2	1	0
2	1	0	2

Men kan eenvoudig nagaan dat in dit schema ieder tweetal rijtjes op drie plaatsen verschilt. Als men in elk van deze twee rijtjes één getal verandert dan kunnen zo niet dezelfde rijtjes ontstaan. Dit impliceert dat de negen rijtjes samen 81, *dus alle*, mogelijkheden overdekken. Zo'n fraaie zgn. *perfecte* oplossing bestaat pas weer bij 13 wedstrijden. Voor $5 \leq n \leq 12$ is het voetbal pool probleem een uitzonderlijk lastig combinatorisch probleem. Het heeft bijna 20 jaar geduurd voor eindelijk de waarde $\delta(5,3)$ = 27 werd vastgesteld (Kamps en Van Lint (1967)). Nog steeds is voor $6 \leq n \leq 12$ het probleem open.

We geven hieronder de resultaten van het onderzoek naar $\delta(n,3)$ voor $6 \leq n \leq 9$ in de laatste 25 jaar. voor deze voordracht is het belangrijk om op te merken dat met S.A. wordt aangegeven dat de grens werd bereikt door van een simulated annealing algoritme gebruik te maken.

Football pool world records

$\delta(6,3) \leq 79$ (E.W. Weber, 1983)
$\delta(6,3) \leq 74$ (L.T. Wille, 1987)...(S.A.)
$\delta(6,3) \leq 73$ (E.H.L. Aarts et al., 1989)...(S.A.)
$\delta(7,3) \leq 225$ (H. Fernandes en E. Rechtschaffen, 1985)
$\delta(7,3) \leq 216$ (A. Blokhuis en C.W.H. Lam, 1984)
$\delta(7,3) \leq 186$ (E.H.L. Aarts et al., 1989)...(S.A.)
$\delta(8,3) \leq 567$ (H. Fernandes, zie boven)

$\delta(8,3) \leq 486$ (E.H.L. Aarts et al., 1989)...(S.A.)
$\delta(9,3) \leq 2.3^6$ (H.J.L. Kamps en J.H. van Lint, 1970)

Een manier om hier simulated annealing te gebruiken gaat als volgt. Beschouw weer de ruimte R van alle rijtjes van n symbolen 0,1, of 2. Kies een getal δ en begin met een random gekozen deelverzameling C van R met grootte δ. De kostenfunctie c is het aantal punten van R dat afstand groter dan 1 heeft tot alle elementen van C (de niet overdekte rijtjes). Bij een perturbatie (een poging in het algoritme) wordt een element van C vervangen door een punt op afstand 1 (mits dit niet reeds in C ligt). Als tijdens het algoritme de functie c de waarde 0 aanneemt, dan is de laatst geaccepteerde C een overdekking en dan is bewezen dat $\delta(n,3) \leq \delta$. Vervang δ door δ-1 en begin opnieuw. We houden op als het algoritme er niet meer in slaagt om een overdekking te vinden. Zo zien we hierboven dat voor $n=6$ nog wel een overdekking werd gevonden bij $\delta=73$ maar helaas niet bij $\delta=72$.
Het zou ons te ver voeren om hier details te geven maar we vermelden wel een interessant detail van het onderzoek naar $\delta(8,3)$. Zoals verwacht mag worden van het simulated annealing algoritme is het gevonden optimum niet een configuratie die duidelijk structuur vertoont. Desalniettemin was de output van het programma niet totaal chaotisch. Door langdurig staren naar deze output is een idee ontstaan over een mogelijke systematische oplossing (niet geheel van optimisme ontbloot). Het idee werkte inderdaad en de hierboven genoemde publikatie van Aarts et al. geeft dan ook een volledig van het computer algoritme onafhankelijke afleiding van $\delta(8,3) \leq 486$.

Referenties

E.H.L. Aarts, P.J.M. van Laarhoven, J.H. van Lint en L.T. Wille (1989), New Upper bounds for the Football Pool Problem for 6,7, and 8 Matches,
J. Combinatorial Theory (A) **52**,pp. 304-312.

A. Blokhuis en C.W.H. Lam (1984), Coverings by Rook Domains,
J. Combinatorial Theory (A) **35**,pp. 240-244.

H. Fernandes en E. Rechtschaffen (1985), The Football pool Problem for 7 and 8 Matches, J. Combinatorial Theory (A) **35**,pp. 109-114.

H.J.L. Kamps en J.H. van Lint (1967), The Football pool Problem for 5 Matches, J. Combinatorial Theory (A), **3**,pp. 315-325.

H.J.L. Kamps en J.H. van Lint (1970), A covering problem, Colloquia Mathematica Societas János Bolyai, **4**,pp. 679-685.

S. Kirkpatrick, C.D. Gelatt,Jr. en M.P. Vechi (1983), Optimization by Simulated Annealing, Science **220**,pp 671-680.

P.J.M. van Laarhoven en E.H.L. Aarts, Simulated Annealing: Theory and Applications, Reidel, Dordrecht.

N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller en E. Teller (1983), Equation of State Calculations by Fast Computer Machines, J. of Chem. Physics **21**,pp 1087-1092.

E.W. Weber (1983), On the Football Pool Problem for 6 Matches, J. Combinatorial Theory (A) **35**,pp 109-114.

L.T. Wille (1987), The Football Pool Problem for 6 Matches: A New Upper Bound Obtained by Simulated Annealing. J. Combinatorial Theory (A) **45**,pp. 171-177.