Tegenwoordig verzamelt ieder zichzelf respecterend bedrijf data van al zijn processen; we zijn in het tijdperk van de Big Data beland, en 'machine learning' is het toverwoord. Op het NGB-seminar van 18 januari 2018 was iedereen het erover eens dat big data grote kansen biedt voor de Operations Research, maar het is nog onduidelijk hoe we die kansen kunnen benutten. In dit artikel laten we zien hoe een basale vorm van datamining kan worden gecombineerd met lineaire programmering bij het oplossen van een belangrijk probleem bij de teelt van bloemen en planten.

STEKJES OOGSTEN

een combinatie van optimalisering en datamining


Calibrachoa Aloha Kona Dark Red van Dummen Orange

Han Hoogeveen

Maak kennis met persoon X: een enthousiaste amateur wielrenner. Hij wil geld ophalen voor een goed doel door de Alpe d'Huez op te fietsen. Omdat het voor hem geen uitdaging is om boven te komen (dat hij heeft hij al diverse malen gedaan) wil hij zich voor ieder van de 21 bochten laten sponsoren: voor iedere bocht is een deadline afgesproken, en hij verdient een behoorlijk bedrag door die deadline te halen en nog een extra bonus voor iedere minuut dat hij er is voordat de deadline is verstreken. De betreffende dag is aangebroken, en X voelt zich in topvorm. Soepel rondt hij de eerste paar bochten en voelt dat hij nog veel over heeft: tijd om te versnellen om de extra bonus voor de volgende bochten te incasseren. Enthousiast pakt hij een flinke bonus in de volgende bochten, maar vanaf bocht 10 gaat het mis en haalt hij de deadlines niet meer: hij heeft teveel gegeven in het begin.

Voldoende stekjes

Een soortgelijk probleem speelt ook bij het oogsten van stekjes bij Dümmen Orange, een internationaal bedrijf dat zich onder andere bezig houdt met het kweken en veredelen van bloemen en planten. De stekjes worden geproduceerd door moederplanten te laten groeien waar iedere week één of meer stekjes van worden afgesneden. Vooraf zijn contracten afgesloten die aangeven hoeveel stekjes per week geleverd moeten worden (vergelijkbaar met de deadlines per bocht) en een eventueel overschot wordt op de markt verkocht (vergelijkbaar met de extra bonus per bocht). Het aantal benodigde moederplanten wordt bepaald op basis van de bekende leververplichtingen met een buffer van 10%. Wanneer het bestaan van deze buffer wordt onthuld, gaat de afdeling Verkoop enthousiast aan de slag, maar wanneer er teveel extra worden verkocht, kan later niet aan de verplichtingen worden voldaan (de wielrenner mist de deadlines van de latere bochten). Dümmen Orange heeft dit probleem aangemeld bij Wiskunde voor de Industrie 2016[1]; dit artikel is gebaseerd op de toen voor dit probleem gevonden oplosmethode.

Lineaire programmering

Simpel gesteld komt het probleem van Dümmen Orange erop neer dat we moeten bepalen hoeveel moederplanten we nodig hebben en hoe vaak we welke snijpatronen moeten gebruiken, waarbij een snijpatroon aangeeft hoeveel stekjes we gemiddeld afsnijden per moederplant per

week; noem dit aantal a_{jt} voor snijpatroon A_j in week t. Deze a_{jt} waarden kunnen fractioneel zijn, omdat het gaat om een gemiddelde opbrengst over alle moederplanten. Uiteraard mogen we alleen snijpatronen gebruiken die niet te veel vragen van een moederplant; zulke snijpatronen noemen we *geldig*. In de context van de fietser komt een snijpatroon overeen met een schema voor een beklimming, waarbij alle doorkomsttijden zijn gespecificeerd; we bekijken alleen schema's die X vol kan houden. Als alle mogelijke geldige snijpatronen bekend zijn, dan kunnen we het minimale aantal benodigde moederplanten bepalen door het volgende lineair programmeringsprobleem (LP) op te lossen:

$$\min M = \sum_{j=1}^{n} x_j$$

onder de voorwaarden

$$\sum_{j=1}^{n} a_{jt} x_j \geq b_t \qquad \forall t = 1, \ldots, T$$

$$x_j \geq 0 \qquad \forall j = 1, \ldots, n$$

Hierin geeft x_j het aantal moederplanten aan waarop we snijpatroon A_j toepassen, en b_t is het aantal stekjes dat nodig is in week t. Aangezien het om grote aantallen gaat is het niet nodig om te eisen dat x_j geheeltallig is; een toegelaten oplossing kan worden gevonden door naar boven af te ronden. Deze oplossing blijkt ook optimaal te zijn, omdat er in een optimale oplossing precies één snijpatroon wordt gebruikt, zoals we later zullen zijn.

Datamining

Helaas is er geen lijst beschikbaar met snijpatronen die gebruikt kunnen worden. Geen nood, een impliciete lijst in de vorm van een verzameling beperkingen waar een geldig snijpatroon aan moet voldoen is ook goed: we kunnen dan met behulp van *kolomgeneratie* (Ford & D Fulkerson, 1958) het bovenstaande LP oplossen, net zoals bij het *cutting stock* probleem (Gilmore & Gomory, 1961; 1963). Echter, die aanpak gaat voor Dümmen Orange ook niet op: een belangrijke onderzoeksvraag is juist het vinden van zo'n lijst van eisen die noodzakelijk en voldoende zijn voor de geldigheid van een snijpatroon. Wat wel beschikbaar is zijn de oogstdata van de afgelopen tien jaar; in de context van de fietser zijn dit de doorkomsttijden van tien eerdere beklimmingen. Uit deze data zijn voor de hand liggende vuistregels af te leiden: aangezien er in geen enkele week gemiddeld meer dan 2,0 stekjes zijn geoogst, mogen we eisen dat $a_{jt} \leq 2{,}0$. Uiteraard kunnen er ook andere redenen aan ten grondslag liggen dat a_{jt} altijd ten hoogste 2,0 is, maar volgens de domeinexpert was $a_{jt} \leq 2.0$ een redelijke eis. Gesterkt door dit succes hebben we vervolgens door middel van een basale vorm van *data mining* een hele lijst met beperkingen afgeleid waar ieder snijpatroon $(a_{1,} a_2, \ldots, a_T)$ aan moet voldoen, waarbij a_t de gemiddelde oogst per moederplant in week t is. Hiertoe bepalen we voor ieder k-tupel $(t_1, \ldots, t_k)$ met $t_1 < t_2 < \ldots < t_k$, waarbij $k \leq 6$ en $t_k - t_1 \leq 10$ de minimale waarde Q zodanig dat voor ieder jaar van de oogstdata geldt

$$a_{t_1} + a_{t_2} + \ldots + a_{t_k} \leq Q.$$

Dit leidt tot een hele lijst met beperkingen die de geldigheid van een snijpatroon beschrijven; een kleine bloemlezing hieruit (deze moeten gelden voor alle t):

$$\begin{aligned}
& a_t && \leq 2{,}0 \\
& a_t + a_{t+1} && \leq 3{,}90 \\
& a_t + a_{t+2} && \leq 3{,}85 \\
& a_t + a_{t+1} + a_{t+2} && \leq 5{,}75 \\
& a_t + a_{t+2} + a_{t+4} && \leq 5{,}71 \\
& a_t + a_{t+1} + a_{t+2} + a_{t+3} && \leq 7{,}60 \\
& a_t + a_{t+1} + a_{t+3} + a_{t+4} && \leq 7{,}61 \\
& a_t + a_{t+1} + a_{t+2} + a_{t+3} + a_{t+4} && \leq 9{,}46 \\
& a_t + a_{t+1} + a_{t+3} + a_{t+4} + a_{t+5} && \leq 9{,}21 \\
& a_t + a_{t+1} + a_{t+2} + a_{t+3} + a_{t+4} + a_{t+5} && \leq 11{,}15 \\
& a_t + a_{t+1} + a_{t+3} + a_{t+4} + a_{t+5} + a_{t+6} && \leq 10{,}90
\end{aligned}$$

Hoeveel moederplanten zijn er minimaal nodig?

Met behulp van de volledige lijst met beperkingen is het mogelijk om met behulp van kolomgeneratie het gegeven LP op te lossen, maar het blijkt dat er een elegantere en veel simpelere oplossing beschikbaar is. De gevonden beperkingen zijn allemaal lineair (want dit zijn de enige beperkingen waar we naar hebben gezocht) en dit impliceert dat iedere *convexcombinatie* $\lambda_1 A_1 + \ldots + \lambda_k A_k$ van geldige snijpatronen $A_1, \ldots, A_k$ ook een geldig snijpatroon is. Als $(x_1^*, \ldots, x_n^*)$ nu een optimale oplossing van het LP is met uitkomstwaarde $M = x_1^* + x_2^* + \ldots + x_n^*$, dan kunnen we hetzelfde resultaat bereiken door M maal het snijpatroon te gebruiken dat gelijk is aan

$$\sum_{j=1}^{n} \frac{x_j^*}{M} A_j.$$

Hieruit volgt dat we kunnen volstaan met precies één snijpatroon, en aangezien we in periode t minstens b_t stekjes nodig hebben, is het optimaal om gemiddeld $a_t = b_t/M$ stekjes te oogsten in periode t. Nu moeten we nog de waarde van M bepalen zodanig dat het snijpatroon $(b_1/M, \ldots, b_T/M)$ aan de beperkingen voldoet.

Beschouw een willekeurige eis, bijvoorbeeld $a_t + a_{t+1} \leq 3{,}9$. Als we hierin $a_t = b_t/M$ invullen en uitwerken, dan vinden we dat $b_t + b_{t+1} \leq 3{,}9\ M$, waaruit volgt

$$M \geq \frac{b_t + b_{t+1}}{3{,}9} \qquad \text{voor alle } t = 1, \ldots, T{-}1.$$

Op deze manier kan iedere beperking worden vertaald naar een ondergrens op M, waarna we M gelijk kunnen kiezen aan de maximale ondergrens.

Maximaliseren extra opbrengst

Uiteraard kan er worden besloten om meer moederplanten te planten. Indien bekend is hoeveel stekjes extra kunnen worden verkocht in week t (noem dit aantal D_t) tegen prijs p_t, dan rijst natuurlijk de vraag: hoeveel moederplanten moeten we planten en hoeveel stekjes moeten we per week oogsten? Ook nu kunnen we bewijzen dat alle moederplanten volgens hetzelfde snijpatroon moeten worden gesneden. We gebruiken M als de variabele die aangeeft hoeveel moederplanten moeten worden geplant; dit brengt kosten $c(M)$ met zich mee, waarvan we eisen dat deze functie lineair is in M. Definieer z_t als het aantal stekjes dat wordt geoogst in week t. Aangezien we maar één snijpatroon gebruiken, worden er gemiddeld $a_t = z_t/M$ stekjes per moederplant geoogst in week t; deze waarden a_t moeten aan de geldigheidsbeperkingen voldoen. Dit is niet lineair, aangezien M een variabele is, maar als we iedere beperking met M vermenigvuldigen, dan houden we lineaire beperkingen over: de beperking $a_t + a_{t+1} \leq 3{,}9$ gaat dan over in $z_t + z_{t+1} \leq 3{,}9\ M$.

Van deze z_t stekjes hebben we er b_t nodig voor de vaste contracten, en de rest kan worden verkocht. De opbrengst moet uiteraard worden gemaximaliseerd, en dit leidt tot het volgende LP:

$$\begin{gathered}
\max \textstyle\sum_t p_t\,(z_t - b_t) - c(M) \\
\text{onder de voorwaarden} \\
b_t \leq z_t \leq b_t + D_t \quad \forall t \\
\text{‘}z_1, \ldots, z_T, M \text{ voldoen aan de geldigheidsbeperkingen’}
\end{gathered}$$

In het bovenstaande LP wordt er van uit gegaan dat alle extra stekjes voor dezelfde prijs kunnen worden verkocht.

Dit kan uiteraard worden veralgemeniseerd door een stuksgewijs-lineaire opbrengstfunctie te gebruiken. Evenzo kunnen we voor $c(M)$ een stuksgewijs-lineaire kostenfunctie gebruiken.

Conclusie

In het bovenstaande hebben we aangegeven hoe datamining kan worden gebruikt binnen de Operations Research; voor zover wij weten is dit de eerste toepassing van datamining in OR in plaats van omgekeerd. Onze basale vorm van datamining levert simpele, lineaire beperkingen op die eenvoudig te combineren zijn met LP, wat een groot voordeel is ten opzichte van het gebruik van geavanceerde technieken uit de machine learning.

Dit artikel is gebaseerd op het rapport van Han Hoogeveen met co-auteurs Jakub Tomczyk en Tom C. van der Zanden. Verder wil ik nogmaals de organisatoren van Wiskunde voor de Industrie 2016 en Dümmen Orange bedanken.

Noot

1. Studiegroep Wiskunde met de Industrie. http://www.ru.nl/math/research/vmconferences/swi-2016/

Literatuur

Ford, J.L.R., & Fulkerson, D.R. (1958). A suggested computation for maximal multicommodity network flows. *Management Science*, 5, 97–101. http://dx.doi.org/10.1287/mnsc.5.1.97.

Gilmore, P.C., & Gomory, R.E. (1961). A linear programming approach to the cutting-stock problem. *Operations Research*, 9, 849–859. http://dx.doi.org/10.1287/opre.9.6.849.

Gilmore, P.C., & Gomory, R.E. (1963). A linear programming approach to the cutting stock problem: Part ii. *Operations Research*, 11, 863–888. http://dx.doi.org/10.1287/opre.11.6.863.

Hoogeveen, J.A., Tomczyk, J., & Van der Zanden, T.C. (2018). *Flower power: Finding optimal plant cutting strategies through a combination of optimization and data mining*. Technical Report Utrecht University, UU-CS-2018-003. http://www.cs.uu.nl/research/techreps/UU-CS-2018-003.html

Han Hoogeveen studeerde econometrie in Groningen en deed daarna bij het CWI onderzoek op het gebied van multicriteria scheduling resulterend in een promotie aan de Universiteit Eindhoven bij Jan Karel Lenstra. Hij is als universitair docent verbonden aan de afdeling Information and Computing Sciences van de Universiteit van Utrecht. In zijn onderzoek richt hij zich vooral op het oplossen van op de praktijk geïnspireerde problemen uit de transport en logistiek met behulp van technieken uit de combinatorische optimalisering.
E-mail: j.a.Hoogeveen@uu.nl