DATA ANALYTICS VOOR PROCES VERBETERING

FRANK VAN DER MEULEN
PROJECTSONE & TU DELFT

Abstract. Big data, data-analytics, big data analytics, machine learning... het zijn allemaal termen die we geregeld in de media horen. En zo nu en dan lijkt er bijna een soort mythe te ontstaan dat deze technieken voorgoed alle problemen in de wereld oplossen. Hoe belangrijk is het om kennis van deze ontwikkelingen te hebben (en houden)? Statistische methoden die standaard binnen Six Sigma aan de orde komen doen nogal wat aannamen op de data; aannamen die bij grote, ongestructureerde data zelden realistisch zijn. Machine-learning methoden, veelal ontwikkeld over de afgelopen 20 jaar door informatici en statistici, gebruiken een andere invalshoek en doen veel minder sterke aannamen op de data. Zeker voor het genereren van hypothesen over je data kunnen deze methoden heel erg nuttig zijn en tot nieuwe inzichten leiden.

Waarom geen gebruik maken van deze "nieuwe" technieken terwijl de software daarvoor beschikbaar is? Uiteraard moet je kennis en vaardigheden ontwikkelen voor het correct toepassen van de technieken. Echter dit lijkt me onontbeerlijk voor elke Black Belt die het maximale uit zijn data-analyse wil halen binnen een Six Sigma project.

Big data, data-analytics, big data analytics, machine learning... het zijn allemaal termen die we geregeld in de media horen. En zo nu en dan lijkt er bijna een soort mythe te ontstaan dat deze technieken voorgoed alle problemen in de wereld oplossen. Veel informatici claimen dat de ontwikkelde technieken voortkomen uit de "artificial intelligence", wiskundigen claimen dat dit allemaal nooit mogelijk was geweest zonder fundamentele wiskunde, numeriek wiskundigen claimen dat numerieke lineaire algebra essentieel is, statistici claimen dat veel fundamenten al lang geleden ontdekt zijn en natuurkundigen claimen waarschijnlijk dat varianten van deze methoden al in de jaren 60 ontwikkeld zijn binnen de natuurkunde. Uiteraard is dit wat gechargeerd en zullen vele onderzoekers uit de diverse velden een minder ruwe inschatting maken. Interessant is dat alle claims op zijn minst een kern van waarheid bevatten en dat ik daarbij ongetwijfeld nog voorbijga aan belangrijke bijdragen uit andere vakgebieden. Naast het algoritmische deel, is veel vooruitgang uiteraard ook te danken aan hedendaags veel betere hardware dan enkele decennia geleden. In dit artikel ga ik in op de rol van statistiek en machine learning (om maar eens één van de vele termen erbij te pakken) binnen Lean Six Sigma. Hoe belangrijk is het om kennis van deze ontwikkelingen te hebben (en houden)? Om een antwoord te geven op deze vraag zal ik eerst het gebruik van een aantal statistische methoden binnen Six Sigma evalueren. Vervolgens zal ik aangeven dat deze methoden nogal wat aannamen doen op de data; aannamen die bij grote, ongestructureerde data zelden realistisch zijn. Machine-learning methoden, veelal ontwikkeld over de afgelopen 20 jaar door informatici en statistici, gebruiken een andere invalshoek en doen veel minder sterke aannamen op de data. Zeker voor het genereren van hypothesen over je data kunnen deze methoden heel erg nuttig zijn en tot nieuwe inzichten leiden. Kennis

Date: October 12, 2017.

hierover lijkt me onontbeerlijk voor een ieder die data-analyse binnen een Six-Sigma project uitvoert.

Laat ik beginnen met een beeld te geven van de stand van zaken zonder het onderwerp big-data. Veel statistiekvakken op hogescholen en universiteiten bestaan voor een groot deel uit regressie-technieken en hypothese toetsen. Binnen Lean Sig-Sigma komen deze technieken terug, maar daarbij bijvoorbeeld ook nog onderwerpen als proces-prestatie analyse. Bij laatstgenoemd onderwerp heb ik veel gebruikers horen spreken over "het fitten van distributies". Wat is er van deze methoden nou daadwerkelijk van belang om tot een succesvol six-sigma project te komen? Als voorbeeld neem ik proces-prestatie analyse. Een visuele dan wel numerieke samenvatting van de status-quo lijkt mij zeer zinvol. We hebben het over beschrijvende statistiek wanneer we een dataset samenvatten met bijvoorbeeld het gemiddelde, de mediaan en de standaard-afwijking, of grafisch via een boxplot of histogram. Een uitvalspercentage valt ook in deze klasse. Inferentiële statistiek bestaat uit een collectie methoden om tot uitspraken te komen die zich niet beperken tot de verkregen dataset, met andere woorden om uitspraken te kunnen doen over data die we mogelijk ook hadden kunnen verkrijgen. Voor inferentiële statistiek is een statistisch model noodzakelijk. Dit beschrijft de klasse van kansverdelingen waaruit we veronderstellen dat de data zijn gegenereerd. Het vinden van een passend statistisch model is doorgaans heel lastig, zeker wanneer een dataset uit vele variabelen bestaat. Mogelijk kan voor de respons (vaak CTQ, Critical To Quality, genoemd), nog wel een verdeling gevonden worden, maar dit voegt aan de proces-prestatie analyse nagenoeg niets toe. Ja we kunnen een curve door het histogram trekken, maar vervolgens doen we er feitelijk niets mee. Voor regressie problemen gaat dit wel meespelen. In het regressieprobleem proberen we de respons voor de i-de experimentele eenheid, zeg y_i, te relateren aan voorspellende variabelen $(x_{i1}, x_{i2}, \ldots, x_{ip})$. Als de respons y_i numeriek is, en als we veronderstellen dat y_i een realisatie is van een stochastiche variabele Y_i, dan bestaat het lineaire regressie-model uit de aanname dat

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + Z_i, \qquad 1 \leq i \leq n, \tag{1}$$

waarbij n dus het aantal experimentele eenheden is en p het aantal invloedsfactoren ("n is het aantal rijen in je spreadsheet; $p+1$ het aantal kolommen"). In de gegeven vergelijking zijn $Z_1, \ldots, Z_n$ de "ruistermen". Doorgaans wordt hiermee bedoeld dat $Z_1, \ldots, Z_n$ onafhankelijk en normaal verdeeld zijn met verwachting gelijk aan nul. Het aantonen dat dit model daadwerkelijk geschikt is, volgt via residuen-analyse. Indien dit zo is, dan kan bijvoorbeeld aan de hand van de p-waarden voor de gefitte coefficiënten $\beta_1, \ldots, \beta_p$ bepaald worden welke invloedsvariabelen significant zijn. Indien de respons dichotoom is, bijvoorbeeld "op tijd"/"te laat", dan kan een vergelijkbare analyse uitgevoerd worden en dit wordt logistische regressie genoemd.

Toen ik buiten de universiteit met "echte" data in aanmerking kwam, merkte ik al snel dat de tekstboekvoorbeelden voor regressie toch vaak geïllustreerd worden aan de hand van net wat "schonere data", en dat het echte werk net wat "vuiler" is. Stel dat je $p = 100$ invloedsvariabelen hebt, sommige categorisch en sommige numeriek, is het dan reëel om te denken dat model (1) valide is? Is het uberhaupt mogelijk om een eenvoudige vergelijking te geven die ons vertelt hoe we de respons op grond van de invloedsvariabelen kunnen voorspellen? En verder, met zo veel variabelen zouden

interacties tussen invloedsvariabelen wel eens van belang kunnen zijn, maar het aantal van deze interacties is van de orde $p^2/2 = 5000$. Moeten we deze nu echt allemaal mee nemen in het model? En iets meer voor de specialisten: problemen rond collineariteit worden steeds waarschijnlijker naarmate p groter wordt. In het bijzonder, als er vele variabelen zijn die in feite uit ruis bestaan (en dus geen relatie met de respons hebben), dan ontstaan specifieke problemen bij het schatten van parameters in het veronderstelde model.

In het veel geciteerde artikel [1] uit 1995 betoogt de inmiddels overleden onderzoeker Leo Breiman dat we het idee van een passend statistisch model moeten laten varen. In plaats daarvan zouden we moeten proberen algoritmen te ontwikkelen waarmee we zo goed mogelijk de respons kunnen voorspellen op grond van de invloedsfactoren. Centraal staan hierbij de termen "predictive modeling" en "predictive performance". Zelf was Breiman vrij uitgesproken: de samenvatting van zijn artikel is als volgt:

> There are two cultures in the use of statistical modeling to reach conclusions from data. One assumes that the data are generated by a given stochastic data model. The other uses algorithmic models and treats the data mechanism as unknown. The statistical community has been committed to the almost exclusive use of data models. This commitment has led to irrelevant theory, questionable conclusions, and has kept statisticians from working on a large range of interesting current problems. Algorithmic modeling, both in theory and practice, has developed rapidly in fields outside statistics. It can be used both on large complex data sets and as a more accurate and informative alternative to data modeling on smaller data sets. If our goal as a field is to use data to solve problems, then we need to move away from exclusive dependence on data models and adopt a more diverse set of tools.

Reeds in de jaren 80 had hij met co-auteurs zogenaamde "Classification And Regression Trees" (CART) ontwikkeld met dit in gedachten. In de jaren 90 ontwikkelde hij "Random Forests": een algoritme gebaseerd op CART om tot een goed voorspellend algoritme te komen. Dit algoritme is computationeel veel zwaarder dan dat van regressie-modellen, maar heeft als groot voordeel dat er weinig aannamen ten aanzien van de data gedaan worden. Verder is de output van het algoritme ongevoelig voor het toevoegen van niet informatieve (ruis) variabelen. "Maar hoe zit dat dan met significantie?" zou je je af kunnen vragen. Dit doet geheel niet ter zake, aangezien hier niet het statistisch model, maar het voorspellende algoritme centraal staat.

In Lean Six-Sigma projecten is vaak het genereren van hypothesen op grond van de data van groter belang dan het feitelijk toetsen van hypothesen. Juist om een idee te krijgen welke invloedsvariabelen ertoe doen, zijn random forests zeer nuttig. Daarna kan weer verdere analyse plaatsvinden via beschrijvende statistiek, of zelfs uiteindelijk via regressie-analyse op een kleine deelverzameling van de invloedsfactoren.

Random forests is zeer zeker niet het enige algoritme dat voorgesteld is in de "machine learning community". Er zijn honderden algoritmen bedacht, waarbij vele varianten van elkaar zijn. Deze hebben vaak prachtige namen als "support vector machines", "boosted trees", "multivariate adaptive regression splines" en "kernel machines". Ik vermoed dat

daar maandelijks weer een aantal algoritmen bijkomt. Sommige algoritmen zijn varianten op regressie-analyse; andere zijn meer van het type random forests. Als toepasser is het lastig om kennis te hebben van al deze verschillende methoden en de vraag rijst of dit nou echt nodig is. Mijn mening is dat kennis van enkele van zulke algoritmen buitengewoon nuttig is en je kan helpen om snel informatie uit je data te halen waar dat met traditionele methoden zeer lastig of zelfs onmogelijk is. Temeer omdat het ook niet meer mogelijk is om op eenvoudige wijze je volledige dataset te visualiseren. Algoritmen uit de machine-learning kunnen helpen bij het genereren van hypothesen, wat in veel projecten waarschijnlijk belangrijker is dan het vaststellen of een geschatte coëfficient significant afwijkt van nul. Weet hebben van meer algoritmen kan natuurlijk zeker geen kwaad, maar bij toepassers van statistische methoden moet natuurlijk altijd de afweging gemaakt worden of de extra inspanning om je deze algoritmen eigen te maken daadwerkelijk loont. Als je hier toch meer over wil weten, dan zou je bijvoorbeeld kunnen beginnen met het lezen van het artikel [2]. Hier wordt reeds betoogd dat we geen honderden algoritmen nodig hebben in de praktijk. Maar kennis hebben van een strict positief aantal algoritmen lijkt me zeer wenselijk.

References

[1] Leo Breiman (2001) *Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author)* Statist. Sci. 16(3), 199-231.

[2] Manuel Fernández-Delgado, Eva Cernadas and Senén Barro (2014) *Do we Need Hundreds of Classifiers to Solve Real World Classification Problems?* Journal of Machine Learning Research 15 (2014) 3133-3181.