Achtergrond 0 Reacties donderdag 20 augustus 2015 Kennislink

Verantwoord data recyclen

Paradox: als je gegevens niet inkijkt, kan je er meer relevante informatie uit halen

Het verband tussen genen en kanker, buitenaardse signalen opsporen in ruis – het komt allemaal neer op het speuren naar regelmaat in chaotische data. Maar als je dat niet goed doet, hou je jezelf voor de gek. Het over-interpreteren van data is schering en inslag in de wetenschap. Amerikaanse wiskundigen bedachten, dat je daarom de data beter voor jezelf geheim kan houden. “Op deze manier gebruik je de data als een soort orakel”, zegt hoogleraar kansrekening Peter Grünwald.

door Arnout Jaspers

Stel, je krijgt op een dag ongevraagd e-mail van een beursgoeroe met een beleggingsadvies: ‘koop nu aandelen KLM, want die gaan volgende week stijgen’. Je negeert dit advies natuurlijk, maar de volgende week stijgen de aandelen KLM wel degelijk. Bovendien krijg je weer een e-mail: ‘Koop nu goud, want de goudprijs gaat volgende week sterk omhoog’.

De week daarop stijgt de goudprijs inderdaad, en de week daarop net zo: iedere keer krijg je een beleggingsadvies waarmee je een mooie winst had kunnen boeken. Na drie weken begin je zijn adviezen op te volgen, en sindsdien heb je je spaargeld verdubbeld. Na een maand krijg je een mail waarin de beursgoeroe je een jaarabonnement op zijn beleggingsadviezen aanbiedt dat vijftig euro per week kost. Wat zou je doen?

In deze 3D-kaart van tienduizenden sterrenstelsels (de aarde zit in de punt van de

Deze website maakt gebruik van cookies.

verberg deze melding

of iets wat wij er graag in willen zien? De overige afbeeldingen bij dit artikel zijn representaties van abstracte, min of meer chaotische datasets.

Geluksvogels

Delen Printen

Vakgebieden

Geneeskunde, Sociale Wetenschappen, Wiskunde

Onderwerpen

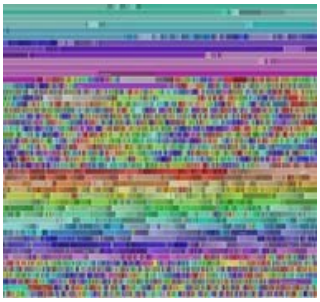
Mens & Maatschappij, Techniek & Natuurwetenschappen

Kernwoorden

significant, p-waarde, data, statistiek

Je zou er verstandig aan doen om toch geen abonnement te nemen, want waarschijnlijk ben je één van duizenden personen die de beursgoeroe bestookte met willekeurige voorspellingen. Sommige voorspellingen komen toevallig uit, en alleen die personen krijgen een tweede voorspelling. In de tweede ronde komen ook sommige daarvan uit, en alleen die gelukkigen krijgen een derde voorspelling gratis. Blijkbaar behoor je tot de selecte groep van geluksvogels die telkens de juiste voorspelling kregen.

Maar er is geen enkele garantie dat, als je een jaarabonnement aanschaft, de adviezen winstgevend blijven. De beursgoeroe heeft namelijk geen idee hoe de koersen zich in de toekomst ontwikkelen. Wiskundig bekeken is dit een voorbeeld van *oversampling,* overinterpretatie. Net zo lang blijven vissen in een bak met data, en telkens je veronderstellingen aanpassen, totdat je daar een schijnbaar relevante wetmatigheid uit haalt. Veel onderzoek in de sociale en medische wetenschappen resulteert in een dergelijke bak met data, waar de onderzoekers graag een of andere wetmatigheid in ontdekken, want dan kunnen ze die publiceren in een vakblad.



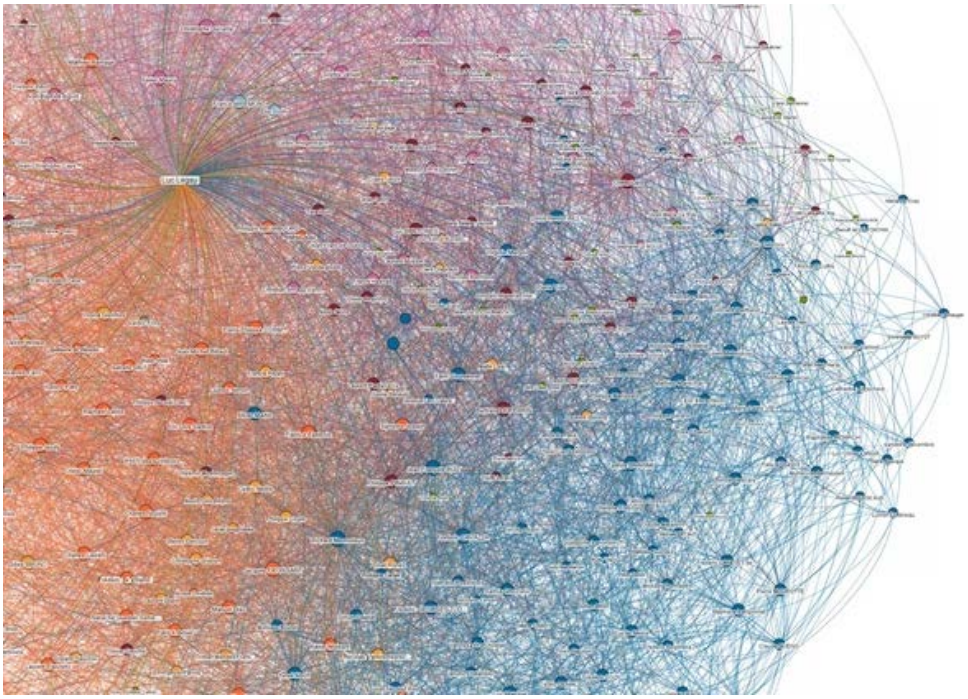
Psychologen testen honderden studenten op allerlei eigenschappen, en vermoeden dan bijvoorbeeld dat linkshandigheid samenhangt met migraine. Of is er misschien een mooier verband tussen acné en tentamenangst bij studenten uit een-oudergezinnen? Het komt echter ook in de ‘harde’ bèta-wetenschappen voor. Zo zoekt men al jaren met grote detectoren naar zwaartekrachtsgolven uit het heelal. Die zouden ontstaan als, bijvoorbeeld, twee neutronensterren heel dicht om elkaar heen draaien.

Als zulke signalen bestaan, registreren deze detectoren ze als mechanische trillingen, die maar moeilijk te onderscheiden zijn van allerlei trillingen uit de aardse omgeving van de detector. Bovendien weet je niet van te voren wat de frequentie en sterkte van het buitenaardse signaal zal zijn. Als je in data waarin ongetwijfeld veel alledaagse rommel zit op zoek gaat naar ‘een of andere trilling’, hoe weet je dan wanneer je echt iets gevonden hebt? Het gevaar is levensgroot, dat je aannames over hoe zo’n signaal eruitziet telkens aanpast, totdat je iets vindt waarvan je jezelf wijsmaakt dat het significant is.

Toevallige patronen

“Overfitting is de gesel van data-analysten, zelfs als er meer dan genoeg data zijn”, schrijven Cynthia Dwork en haar collega’s in een artikel dat ze in juni 2015 publiceerden op arXiv, een website voor natuurwetenschappelijke artikelen. Dwork is een gerenommeerd computerwetenschapper en cryptograaf die nu bij Microsoft werkt. Als je voorafgaand aan het verzamelen van de data één hypothese formuleert, en vervolgens kijkt hoe goed je data daaraan voldoen, dan bestaan er statistische tests om te bepalen of je iets hebt gevonden wat significant is. Maar als je keer op keer je hypothese aanpast en test op dezelfde data, dan weet je niet meer of je een echte wetmatigheid op het spoor bent, of slechts een toevallig patroon in data met veel ruis.

Maar dat is juist de trend in Big Data: eerst geautomatiseerd enorme hoeveelheden data verzamelen, en dan allerlei hypotheses verzinnen over welke patronen er in die data zouden kunnen zitten. Levert hypothese één niets op, dan probeer je hypothese twee, en drie, en vier, net zo lang tot je iets vindt wat er interessant uitziet.

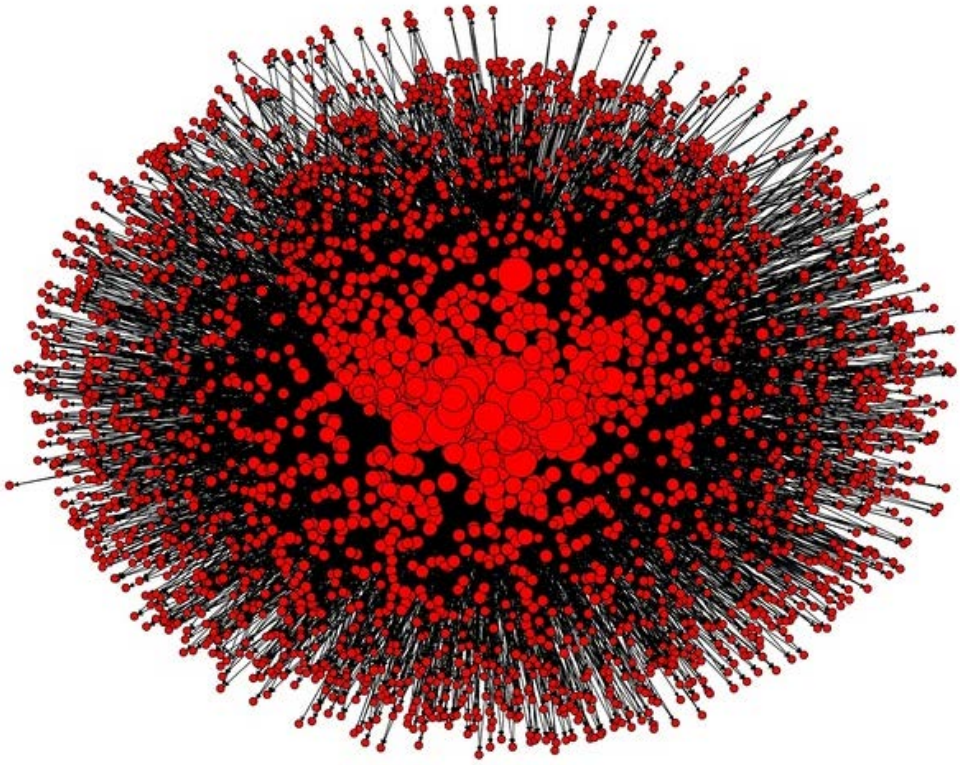


Het gevaar van overinterpretatie is zeker ook aanwezig bij het zoeken naar de genetische basis van kwalen als kanker, autisme, dementie of schizofrenie. Als je het DNA van een groot aantal patiënten afleest en in duizenden genen speurt naar combinaties van gen-varianten die iets met de ziekte te maken kunnen hebben, test je eigenlijk een groot aantal hypotheses tegelijk op één dataset. Eigenlijk is men pas vrij recent gaan nadenken over de statistische haken en ogen van dit soort onderzoek. In Nederland kreeg de Leidse wiskundige Aad van der Vaart daar dit jaar nog een Spinoza-premie voor.

De kist openmaken

Het gevaar van overinterpretatie wordt al langer erkend, en in principe is er ook een remedie tegen. Je moet vooraf je dataset splitsen in een trainings-set en een *holdout*-set, een set met data die als het ware in een afgesloten kist wordt bewaard. Dat moet je *random* doen, dat wil zeggen zodanig dat er geen enkel relevant verschil tussen de twee datasets ontstaat.

In het voorbeeld van de zwaartekrachtsdetector zou je elke maand een trainings-set kunnen maken met waarnemingen op de oneven uren van de dag, op vijf dagen die je met de dobbelsteen uitloot, terwijl de rest van de waarnemingen de kist in gaat. Iedereen mag in de trainings-set speuren naar signalen die mogelijk afkomstig zijn van zwaartekrachtsgolven uit het heelal.



Omdat de data voor minstens 99 procent uit ruis bestaan, is dat ook een kwestie van interpretatie en gokken. Het lijkt op het onderscheiden van één fluisterstem in een luidruchtige menigte. Pas als iedereen zijn favoriete kandidaat-signaal heeft bepaald, gaat de kist open. Dat is het moment van de waarheid: als je afstemt op een fluisterstem met precies die toonhoogte en timbre, hoor je die dan ook in de holdout-set? Pas dan weet je vrij zeker dat je een echt signaal hebt gevonden, en niet een hersenschim.

Dit principe kun je ook toepassen bij onderzoek naar het verband tussen genen en ziektes. In dat geval verdeel je de patiënten met de dobbelsteen in een trainingsgroep en een holdout-groep. Pas als je meent in de trainingsgroep genen te hebben geïdentificeerd die sterk samenhangen met een bepaald type kanker, check je of dat statistische verband in de holdout-groep overeind blijft.

Voor eenmalig gebruik

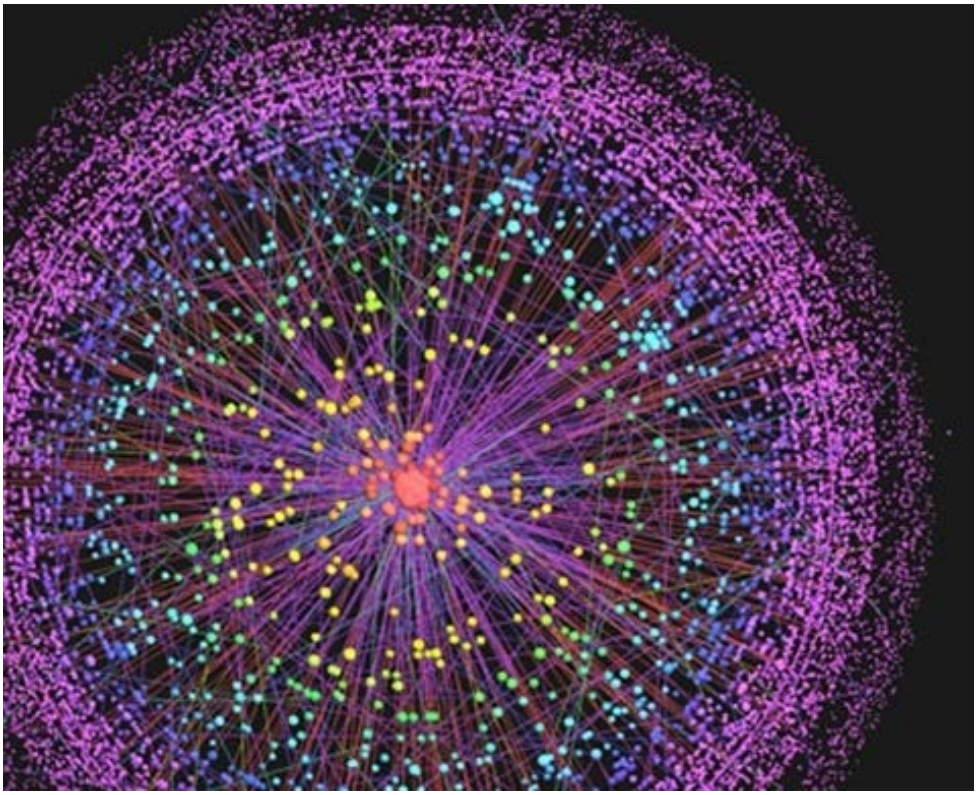
Het principe van een gescheiden trainings-set en holdout-set is mooi, maar onpraktisch. Want als de kist eenmaal open is gegaan, kun je strikt genomen de holdout-set maar één keer gebruiken. Daarna heeft iedereen die data gezien, en kunnen onderzoekers – bewust of onbewust – hun aannames over hoe het signaal eruitziet daarop aanpassen. Zo strikt neemt men het niet in de rommelige praktijk; het verzamelen van zulke data is duur en tijdrovend, dus worden ze hergebruikt, soms vele malen.

Computerwetenschapper en cryptograaf Cynthia Dwork publiceerde met haar collega's echter deze maand in vakblad *Science* een methode die het mogelijk maakt om holdout-data een groot aantal malen te hergebruiken. Het basisidee is, dat je nooit rechtstreeks in de kist kijkt, maar via een algoritme (een app, zou je kunnen zeggen) met de woordspelige naam *Thresholdout* (*threshold* = drempel).

De procedure is: eerst formuleer je op grond van de trainings-set een hypothese ('genen x, y en z zijn betrokken bij het ontstaan van leukemie') en giet die via *Thresholdout* in een wiskundige formule die één getal oplevert. Dat is een rapportcijfer voor hoe goed je

hypothese op de trainingsdata past. Maar dat getal krijg je nog niet te zien, want eerst toetst *Thresholdout* dezelfde wiskundige formule aan de holdout-data, wat ook een rapportcijfer oplevert. Vervolgens doet het algoritme iets merkwaardigs. Als het verschil tussen deze twee rapportcijfers klein genoeg is, geeft het als antwoord het trainings-rapportcijfer. Is het verschil groter, dan is het antwoord het rapportcijfer van de holdout-set binnen een onzekerheidsmarge (daar zit ook nog een toevalsfactor in, zodat je niet kunt concluderen dat het echte rapportcijfer precies in het midden van de marge zit).

Peter Grünwald, hoogleraar kansrekening aan de Universiteit Leiden en onderzoeker aan het Centrum Wiskunde & Informatica, vindt het een uiterst origineel idee. “Op deze manier gebruik je de data als een soort orakel, dat een ja/nee antwoord geeft op de vraag: past deze hypothese op deze data?”



Orakel uitgeput

Een onderzoeker komt dus nooit precies te weten hoe goed zijn of haar hypothese op de holdout-data past, maar slechts ongeveer. Dit nadeel wordt echter ruimschoots gecompenseerd doordat hij of zij herhaaldelijk de hypothese mag aanpassen, die dan weer via *Thresholdout* aan dezelfde holdout-data mag worden getoetst. “Ieder keer dat je je hypothese opnieuw toetst aan de holdout-data, krijg je een antwoord met een grotere onzekerheidsmarge. Uiteindelijk wordt die marge zo groot, dat het antwoord niks meer zegt. Dan is het orakel uitgeput”, zegt Grünwald.

Je kunt wiskundig bewijzen, dat je op deze manier vrijwel nooit over-interpreteert. Als je een patroon in de data vindt, weet je vrijwel zeker dat je dit ook aantreft in een volledig nieuwe dataset. Zo spoor je dus met grote zekerheid alleen genen op die echt te maken hebben met het ontstaan van leukemie, en geen toevallige, zelf bedachte combinatie van genen die alleen bij deze ene groep patiënten afwijkt.

Tot zover de theorie, want de methode is alleen nog getest in computersimulaties, niet op echte patiëntengegevens of detectordata. “Daarom is het nog te vroeg om te zeggen

of deze techniek de belofte ook zal waarmaken," aldus Grünwald.

Privacy beschermen

Volgens Richard Gill, hoogleraar statistiek in Leiden, is het heel verrassend dat inzichten uit de cryptografie gebruikt kunnen worden bij dit soort problemen. De kunstgreep om alleen inzicht in de data te geven via een algoritme dat een toevalsgetal aan het resultaat toevoegt, wordt *differential privacy* genoemd. Deze techniek is oorspronkelijk mede door Dwork ontwikkeld om de privacy te beschermen van patiënten in een medische database, terwijl je er toch allerlei nuttige informatie aan kunt onttrekken.

Wel constateert Gill, dat *Thresholdout* in deze vorm alleen werkt met de simpelste vorm van statistiek, waarbij de hypothese niet ingewikkelder is dan het nemen van een gemiddelde van een stel data. Dwork en haar groep stellen echter dat de methode ook werkt voor ingewikkelder statistische operaties, al is dat in het artikel in *Science* nog niet verder uitgewerkt.

"Andere onderzoekers zullen dit zeker gaan uitproberen met echte data", benadrukt Grünwald. "Over een jaar of twee weten we echt of dit een bruikbare methode is. Wat dat betreft komt deze publicatie in *Science* eigenlijk een beetje vroeg."

Bron

- Dwork, C. *The reusable holdout: Preserving validity in adaptive data-analysis,* Science (7 augustus 2015). DOI: 10.1126/science.aaa9375

Deel deze publicatie

Dit is een publicatie van **Kennislink**

meer informatie | website

© Kennislink, sommige rechten voorbehouden

Stuur ons een reactie, vraag, suggestie