Beter prevalenties meten door het combineren van schattingsmethoden

Machine learning algoritmes worden veel ingezet voor classificatiedoeleinden. Met een hoge nauwkeurigheid kunnen algoritmes veel objecten juist classificeren. Maar wat als we niet geïnteresseerd zijn in de classificatie van ieder object, maar juist in de groepsgrootte van iedere klasse? Classificeren en optellen lijkt een logische aanpak, zeker als het algoritme een hoge nauwkeurigheid heeft. In dit artikel vinden we uit waarom dit geen goede manier is om te kwantificeren en, nog veel belangrijker, hoe we dat op kunnen lossen.

Kevin Kloos

In de officiële statistiek zijn veel statistieken gewijd aan het schatten van prevalenties. Deze lopen uiteen van het percentage 65-plussers met een internetverbinding tot het percentage werknemers dat lid is van een vakbond. De laatste jaren zijn *machine learning* algoritmes enorm in opkomst, ook binnen de officiële statistiek. Nieuwe toepassingen liggen bijvoorbeeld in het bepalen of een woning zonnepanelen heeft aan de hand van satellietbeelden, of bepalen of een retailer een webwinkel heeft aan de hand van diens webpagina. Deze algoritmes zijn echter imperfect en kunnen dus classificatiefouten maken. Als er in de ene groep meer fouten worden gemaakt dan in de andere groep ontstaat er *misclassification bias,* oftewel vertekening door misclassificatie. Deze vertekening ontstaat niet alleen bij slimme algoritmes. In de *STAtOR*-editie van juni 2021 schreef Anton Meijburg (Meijburg, 2021) welke invloed classificatiefouten van PCR-testen hebben op het schatten van de prevalentie positieve uitslagen. Om de prevalentie juist te schatten is het vanzelfsprekend om de vertekende schatting te corrigeren.

Paradox

De vertekening van de geschatte prevalentie hangt af van drie parameters: 1. de daadwerkelijke prevalentie, 2. het aantal juist geclassificeerde positieven (sensitiviteit) en 3. het aantal juist geclassificeerde negatieven (specificiteit). Het lijkt logisch om te denken dat een verbetering in de sensitiviteit en/of specificiteit leidt tot een betere schatting van de prevalentie. Aan de hand van een simpel voorbeeld is het gemakkelijk om te zien dat dat niet altijd hoeft te zijn. Algoritme A schat 20 personen fout in: 10 fout positieven en 10 fout negatieven. Algoritme B schat 15 personen fout in: 9 fout positieven en 6 fout negatieven. Algoritme B kan beter classificeren, omdat zowel de sensitiviteit als de specificiteit van algoritme B hoger ligt dan die van algoritme A. Algoritme A kan echter beter de prevalentie schatten, omdat de fout positieven en de fout negatieven tegen elkaar weggestreept worden, terwijl dat in algoritme B niet het geval is. In algoritme B zijn, in tegenstelling tot algoritme A, de fouten niet gelijk over de groepen verdeeld en ontstaat er dus een vertekening van de geschatte prevalentie.

Steekproef

Om de prevalentie beter te kunnen schatten, is er meer informatie nodig over de specificiteit en sensitiviteit. Deze twee waarden kunnen worden bepaald met een aselecte steekproef. We nemen een willekeurige groep bestaande uit *n* observaties waar naast de geschatte classificatie ook de daadwerkelijke classificatie bekend is. In het voorbeeld van de webpagina's en de zonnepanelen kan er bijvoorbeeld handmatig gecontroleerd worden wat de daadwerkelijke classificatie is. Deze informatie kan worden samengevat in een kruistabel (tabel 1). De sensitiviteit en de specificiteit kunnen beiden uit deze kruistabel worden geschat, alsmede een schatting van de daadwerkelijke prevalentie.

Corrigeren

Aan de hand van deze steekproef (tabel 2) kan de vertekende schatting van de prevalentie ((TP + FP) / n = (900 + 300) / 2000 = 0,6) gecorrigeerd worden. Er zijn drie bekende en makkelijk toepasbare manieren om de prevalentie te corrigeren. De makkelijkste manier om de prevalentie te corrigeren is simpelweg het aantal daadwerkelijke positieven tellen en te delen door de omvang van de steekproef ((TP + FN) / n = (900 + 100) / 2000 = 0,5). Een andere bekende manier om de prevalentie te corrigeren is door de vertekende schatting te corrigeren met een kansenmatrix (tabel 3). Een kansenmatrix kan makkelijk geconstrueerd worden vanuit de steekproef door simpelweg de rijen te laten optellen tot 1, ook wel het rij-normaliseren van een matrix genoemd. De gecorrigeerde schatting kan gemaakt worden door de inverse

	Geschat positief	Geschat negatief
Daadwerkelijk positief	TP (terecht positief)	FN (fout negatief)
Daadwerkelijk negatief	FP (fout positief)	TN (terecht negatief)

Tabel 1. Voorbeeld van een kruistabel

	Geschat positief	Geschat negatief
Daadwerkelijk positief	900	100
Daadwerkelijk negatief	300	700

Tabel 2. Fictief voorbeeld van een steekproef

	Geschat positief	Geschat negatief
Daadwerkelijk positief	TP / (TP + FN) = 0,9	FN / (TP + FN) = 0,1
Daadwerkelijk negatief	FP / (TN + FP) = 0,3	TN / (TN + FP) = 0,7

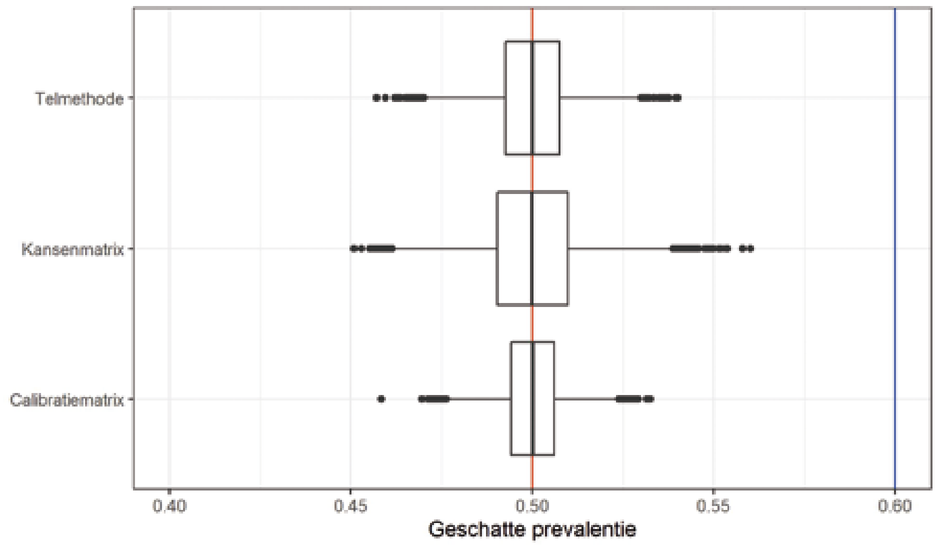
Tabel 3. Kansenmatrix van fictieve steekproef

	Geschat positief	Geschat negatief
Daadwerkelijk positief	TP / (TP + FP) = 0,75	FN / (TN + FN) = 0,125
Daadwerkelijk negatief	FP / (TP + FP) = 0,25	TN / (TN + FN) = 0,875

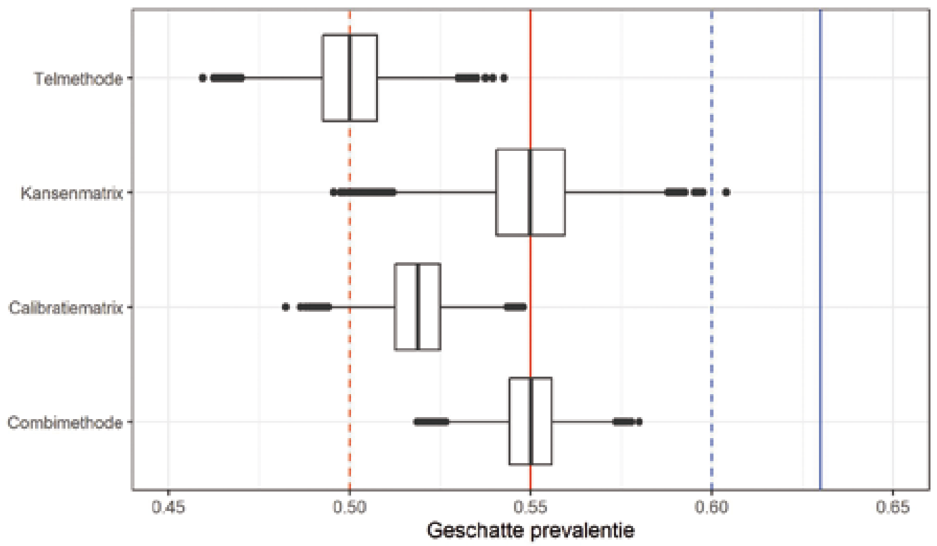
Tabel 4. Kalibratiematrix van fictieve steekproef

van de getransponeerde kansenmatrix te berekenen en te vermenigvuldigen met de vertekende schatting van de prevalentie. Een minder bekende manier om de prevalentie te corrigeren is door de vertekende schatting te corrigeren met een kalibratiematrix (tabel 4). Waar de kansenmatrix geconstrueerd wordt door het rij-normaliseren van de kruistabel van de steekproef, kan de kalibratiematrix geconstrueerd worden door het kolom-normaliseren van diezelfde kruistabel. De gecorrigeerde schatting kan gemaakt worden door de kalibratiematrix te vermenigvuldigen met de vertekende schatting van de prevalentie.

De drie correctiemethoden geven een onvertekende schatting van de prevalentie, maar de variantie is verschillend bij ieder van deze methoden. Een simulatie van 10.000 steekproeven laat zien dat alle schattingen van iedere methode dichter bij de daadwerkelijke prevalentie zitten dan de vertekende prevalentie (figuur 1). De methode met de kalibratiematrix varieert het minst en de methode met de kansenmatrix varieert het meest onder deze set van parameters. Het is bewezen dat de methode met de kalibratiematrix altijd minder varieert dan de methode met de kansenmatrix. Dit komt omdat er bij de



Figuur 1. Boxplot van 10.000 simulaties om de prevalentie te schatten. De daadwerkelijke prevalentie is 0.50 (rood), terwijl de vertekende prevalentie 0.6 (blauw) is. Voor de simulatie zijn de dezelfde gegevens als in de tabellen zijn gebruikt: sensitiviteit = 90%, specificiteit = 70%, daadwerkelijke prevalentie = 50%, steekproefomvang van 2.000 op een zeer grote populatie



Figuur 2. Boxplot van 10.000 simulaties om de prevalentie te schatten. De nieuwe daadwerkelijke prevalentie is 0,55 (rood, solide), terwijl de vertekende prevalentie 0,63 (blauw, solide) is. De oude daadwerkelijke prevalentie is 0,50 (rood, stippellijn) en de oude vertekende prevalentie is 0,60 (blauw, stippellijn). Voor de simulatie zijn de dezelfde gegevens als in de tabellen zijn gebruikt: sensitiviteit = 90%, specificiteit = 70%, daadwerkelijke prevalentie = 50%, steekproefomvang van 2.000 op een zeer grote populatie

methode met de kalibratiematrix niet geïnverteerd hoeft te worden, in tegenstelling tot de methode met de kansenmatrix. Een hogere sensitiviteit of specificiteit leidt tot minder variantie bij de `matrixmethoden`, maar heeft geen invloed op de simpele telmethode. Verder is het vanzelfsprekend dat een grotere steekproef leidt tot minder variantie in alle correctiemethoden.

Verandering over tijd

De prevalentie blijft vanzelfsprekend niet constant over tijd. Om een prevalentie te schatten over tijd, is in een ideale situatie een nieuwe aselecte steekproef nodig. Hier is vaak geen tijd en/of geld voor. Een mogelijke oplossing is om de oude steekproef ook te gebruiken op het nieuwe tijdspunt. Aan de andere kant is het wél relatief makkelijk om een nieuwe vertekende schatting te maken met het slimme algoritme. De informatie van de oude steekproef en de nieuwe vertekende schatting worden gecombineerd tot een nieuwe gecorrigeerde schatting. Het is de vraag of de drie correctiemethoden ook werken op het nieuwe tijdspunt.

De situatie in figuur 1 wordt als uitgangspositie genomen. Stel dat de prevalentie verandert van 50% in de oude situatie naar 55% in de nieuwe situatie. De vertekende prevalentie stijgt van 60% naar 63%, wat betekent dat het verschil niet gecorrigeerd kan worden met alleen de stijging van de vertekende prevalentie. Het valt op dat twee van de drie correctiemethoden een vertekende schatting maken (figuur 2). Ten eerste is het logisch dat de telmethode een vertekende schatting maakt. De steekproef is notabene getrokken uit een populatie met een prevalentie van 50%. De schatting met de kansenmatrix blijft onvertekend, omdat de waarden van de specificiteit en sensitiviteit onveranderd blijven in de nieuwe situatie. De schatting met de kalibratiematrix is echter vertekend. De verandering van de prevalentie over tijd kan niet helemaal gecorrigeerd worden; de gemiddelde schatting ligt tussen de oude en de nieuwe prevalentie in. Dit komt omdat de kansen in de kalibratiematrix afhankelijk zijn van de daadwerkelijke prevalentie, terwijl de kansen in de kansenmatrix onafhankelijk zijn van de daadwerkelijke prevalentie. De methode met de kansenmatrix kan dus goed de verandering over tijd schatten, maar heeft veel variantie voor de initiële schatting. Daarentegen kan de methode met de kalibratiematrix een goede initiële schatting kan maken, maar kan het niet goed de verandering over tijd verwerken.

De verandering over tijd zorgt er dus voor dat er nog geen goede methode is om de prevalentie ook over tijd goed in te kunnen schatten. Een mogelijke oplossing is om de methodes met de kansenmatrix en de kalibratiematrix te combineren tot de combimethode. Als startpunt wordt de methode met de kalibratiematrix genomen, maar de verandering over tijd wordt beschreven door het verschil in de methode met de kansenmatrix. In figuur 2 valt het op dat de waarden van de combimethode ongeveer dezelfde variantie heeft als de methode met de kalibratiematrix uit figuur 1. De combimethode zorgt ervoor dat de verandering over tijd beter gemodelleerd kan worden, zonder dat er verdere ingewikkelde algoritmes gebruikt hoeven te worden. Het is bewezen dat de combimethode beter presteert dan de individuele correctiemethoden, op een paar extreme situaties na, en dus een goede toevoeging is op het bestaande palet van correctiemethoden.

Conclusie

De voornaamste conclusie uit dit artikel is dat prevalenties schatten veel meer is dan alleen goed classificeren en dan optellen. Zelfs goedwerkende classificatie-algoritmes kunnen vertekende uitkomsten genereren die gecorrigeerd moeten worden. Een aantal correctiemethoden is gepresenteerd in dit artikel. Klassieke correctiemethoden werken goed als de prevalentie constant blijft, terwijl de nieuwe combimethode goed werkt als de prevalentie verandert over tijd. Mocht u geïnteresseerd zijn in de wiskundige uitwerkingen van de methodes, wijs ik u graag door naar mijn master thesis (Kloos, 2021) of neem contact met me op.

Literatuur

Meijburg, A. (2021). Sensitiviteit, specificiteit en COVID-19. *STAtOR, 22*(2), 40–44. (https://www.vvsor.nl/wp-content/uploads/2021/07/STAtOR-2021-2-40-44-Meijburg.pdf)

Kloos, K. (2021). *Comparing correction methods to reduce misclassication bias*. Leiden University. Master Thesis. (https://www.github.com/kevinkloos/MasterThesis)

Kevin Kloos doet promotieonderzoek in Quantification Learning aan de Universiteit van Leiden. Gedurende zijn masteropleiding Statistical Science volgde Kevin het deeltraject European Master of Official Statistics (EMOS), waardoor hij zijn scriptie kon schrijven bij het Centraal Bureau voor de Statistiek (CBS). Zijn wetenschappelijke artikel over de combimethode leverde de tweede prijs op in de internationale IAOS Young Statisticians Prize. Dit artikel is een samenvatting van zijn masterscriptie.
E-mail: k.kloos@fsw.leidenuniv.nl