Similarity coefficients for binary data : properties of coefficients, coefficient matrices, multi-way metrics and multivariate coefficients

Warrens, M.J.

Summary in Dutch (Samenvatting)

We spreken van binaire of dichotome data als er sprake is van een reeks getallen die slechts twee waardes aannemen. De twee waardes kunnen gezien worden als twee, elkaar uitsluitende, categorieën die voor het gemak als 1 en 0 kunnen worden gecodeerd. Een binaire reeks, bijvoorbeeld $\{0, 1, 1, 0, 0, 1, 0, 1, 1, 1, 0\}$, kan verkregen worden door van een aantal personen het geslacht vast te stellen, waarbij bijvoorbeeld, 1=vrouw en 0=man. Een binaire reeks kan ook verkregen worden door voor een persoon te coderen welke vragen hij of zij goed of fout had op een toets. Duidelijk is dat binaire reeksen simpel te verkrijgen zijn door allerlei tweeledigheden te verzamelen: goed/fout, voor/tegen, wel/niet, nieuw/oud, ja/nee, aanwezig/niet aanwezig, of PSV-fan/geen kampioen worden.

Als er twee of meer binaire reeksen beschikbaar zijn kan het interessant zijn om te weten in hoeverre de twee reeksen op elkaar lijken. Een bioloog die voor twee gebieden heeft gecodeerd welke diersoorten er wel of niet leven, bijvoorbeeld $\{0, 1, 1, 0\}$ en $\{1, 0, 1, 0\}$, kan zich afvragen in hoeverre de twee gebieden (reeksen) op elkaar lijken. Om twee reeksen te kunnen vergelijken moeten de posities van de reeksen wel dezelfde diersoorten weergeven. De eerste reeks geeft bijvoorbeeld aan dat in het eerste gebied geen vogels, wel paarden, wel muizen, maar geen schildpadden leven; de tweede reeks geeft aan dat in het tweede gebied wel vogels, geen paarden, wel muizen, en geen schildpadden leven. Twee reeksen kunnen nu vergeleken worden met elkaar door na te gaan hoeveel 1n of 0n ze gemeenschappelijk hebben in dezelfde posities. In de biologie zijn twee leefomgevingen meer in overeenstemming naarmate er meer diersoorten in beide gebieden aanwezig zijn (het is niet gebruikelijk om overeenstemming te definiëren in termen van afwezigheid).

Essentieel bij het bestuderen van binaire reeksen is het uitgangspunt dat alle informatie in twee reeksen van gelijke lengte uitputtend kan worden samengevat in vier getallen: $a = \#(1, 1)$ = het aantal posities dat beide reeksen een 1 hebben $(1, 1)$, $d = \#(0, 0)$ = het aantal posities dat beide reeksen een 0 hebben $(0, 0)$, en $b = \#(1, 0)$ en $c = \#(0, 1)$ = de aantallen posities dat er een 1 staat in de ene reeks en een nul in de andere reeks. Willen we de overeenstemming van twee binaire reeksen quantificeren dan kan dat met behulp van overeenstemmingsmaten

of gelijkheidscoëfficiënten. Een overeenstemmingsmaat drukt de gelijkheid van twee binaire reeksen uit in een getal. Voor de vergelijking van twee binaire reeksen is door de jaren heen echter een groot aantal maten voorgesteld. Ondanks het groot aantal verschillende overeenstemmingsmaten, zijn ze allemaal een of andere functie van de getallen of variabelen a, b, c, en d. Een voorbeeld is de Jaccard coefficient met de formule $a/(a+b+c)$.

Omdat het niet altijd duidelijk is wat nu in welke situatie de meeste geschikte coëfficiënt of associatiemaat is, is het nuttig om de coëfficiënten en hun eigenschappen te bestuderen. Dit kan op een veelvoud van manieren, maar in dit proefschrift is gekozen voor een mathematische bestudering van de coëfficiënten en hun eigenschappen. Kortweg wordt hiermee bedoeld dat het niet echt uitmaakt welke waardes de getallen a, b, c, en d aannemen, maar dat het alleen van uitmaakt hoe a, b, c, en d zich (in een formule) tot elkaar verhouden. Eigenlijk worden in het gehele proefschrift verschillende combinaties (formules), allemaal functies van a, b, c, en d, met elkaar vergeleken.

Dit proefschrift bestaat uit negentien hoofdstukken verdeeld in vier delen. In deel I worden steeds eigenschappen van coëfficiënten bestudeerd waar alleen de individuele formule voor nodig is. In dit deel worden er slechts twee binaire reeksen tegelijk met elkaar vergeleken. In deel II en IV worden twee benaderingen besproken voor het geval dat we meer dan twee reeksen tegelijk beschouwen. In deel II worden niet individuele coëfficiënten maar matrices van coëfficiënten bestudeerd. Een matrix wordt verkregen door, bij meer dan twee binaire reeksen, tussen alle paren van reeksen de associatiemaat te bepalen. Deze coëfficiënten kunnen dan worden weergegeven in een coëfficiëntmatrix. In deel IV van dit proefschrift worden coëfficiënten gedefiniëerd die de mate van associatie of overeenstemming reflecteren van twee of meerdere binaire reeksen tegelijk. Voordat de meerweg coëfficiënten in deel IV worden behandeld, wordt deel III gebruikt om een aantal meerweg concepten te definiëren en te bestuderen.

Deel I bestaat uit vijf hoofdstukken. Notatie en enkele basisconcepten van overeenstemmingsmaten worden geïntroduceerd in Hoofdstuk 1. Hoofdstuk 2 stelt de overeenstemmingsmaten voor binaire data in een breder perspectief. De formules die in dit proefschrift worden behandeld zijn in veel gevallen een speciaal geval van een formule die geschikt is voor algemenere data dan binaire gegevens. In dit hoofdstuk wordt aangetoond dat men de Hubert-Arabie adjusted Rand index kan uitrekenen door eerst de 2×2 tabel te formeren door het aantal objectparen te tellen dat in hetzelfde cluster is geplaatst door beide methodes, dat in een cluster is geplaatst door een methode maar in verschillende clusters door de andere methode, en het aantal objectparen te tellen dat in verschillende clusters door beide methodes is geplaatst, en vervolgens Cohen's kappa uit te rekenen voor deze 2×2 tabel. Hoofdstuk 3 laat zien dat een aantal coëfficiënten behoren tot families van coëfficiënten. Het bestuderen van families in plaats van individuele coëfficiënten geeft ons vaak algemenere inzichten en resultaten. Een hoge waarde van een coëfficiënt kan ook komen door toeval. In hoofdstuk 4 worden coëfficiënten en correctie voor toeval bestudeerd en wordt, bijvoorbeeld, aangetoond dat de simple matching coëfficiënt, Cohen's kappa,

Goodman en Kruskal's lambda, Scott's pi, Hamann's eta, en overeenstemmingsmaten geintroduceerd door Gleason/Dice/Sørenson en Rogot en Goldberg, equivalent worden na correctie voor toeval, ongeacht de verwachte waarde die gebruikt wordt.

De maximale waarde van een coëfficiënt gegeven de marginale distributies (totaal aantal 1n van de ene en andere binaire reeks) wordt bestudeerd in hoofstuk 5. Voor sommige coëfficiënten is de maximale waarde niet onder alle omstandigheden gelijk aan 1. De formule van deze coëfficiënten wordt een andere formule na correctie voor de maximale waarde. Iedere overeenstemmingsmaat voor binaire data, waarvan de teller gelijk is aan de covariantie en de noemer een functie is van de marginale distributies, wordt gelijk aan de Loevinger coëfficiënt na correctie voor maximale waarde gegeven de marginale dsitributies.

Deel II bestaat uit vijf hoofdstukken. Hoofdstuk 6 beschrijft een aantal manieren waarop de 1n en 0n van twee of meer binaire reeksen aan elkaar gerelateerd kunnen zijn. De modellen en data structuren die hier beschreven worden, dienen in hoofdstukken 7 en 8 als voldoende voorwaarden voor bepaalde matrices van coëfficiënten om zekere eigenschappen te bezitten. Hoofdstuk 7 betreft Robinson matrices. Een vierkante coëfficiëntenmatrix wordt een Robinson matrix genoemd als de hoogste waardes in iedere rij en kolom op de hoofddiagonaal liggen, en wegbewegend van de hoofddiagonaal zijn de waardes nooit oplopend. In hoofdstuk 8 worden eigenwaardes en eigenvectoren van coëfficiëntenmatrices bestudeerd. Als het double monotonicity model voor binaire items opgaat, dan wordt de correcte ordering van de items weerspiegeld in de elementen van de eigenvector behorende bij de grootste eigenwaarde van de matrix met elementen $a(i,j)/p(j)$, waar $a(i,j)$ de proportie 1n is dat items i en j in dezelfde posities hebben, en $p(j)$ is de proportie item correct van item j. In hoofdstuk 9 wordt een systematische vergelijking gemaakt tussen een eigenwaarde techniek, homogeniteitsanalyse, en het logistische item response theory model met twee parameters. Hoofdstuk 10 is het eerste hoofdstuk waar metrische eigenschappen van coëfficiënten worden bestudeerd. Een functie wordt metrisch genoemd als deze voldoet aan de driehoeksongelijkheid. Dit hoofdstuk dient als opstap naar deel III, waar allerlei generalisaties van driehoeksongelijkheid worden gedefiniëerd en besproken.

Deel III bestaat uit vijf hoofdstukken. Voordat meerweg coëfficiënten bestudeerd kunnen worden in deel IV, wordt eerst een aantal meerweg concepten gedefiniëerd en bestudeerd in deel III. Ideeën voor de meerweg concepten zijn vooral verkregen door te kijken naar literatuur over drieweg data-analyse. Hoofdstuk 11 behandelt axioma's en basiseigenschappen die kunnen opgaan voor meerweg coëfficiënten en hun complementen, afstandsmaten. In dit hoofdstuk wordt onder andere bestudeerd wat mogelijk de kleinste sets van axioma's zijn. In hoofdstuk 12 wordt geëxploreerd op welke manieren de driehoeksongelijkheid kan worden gegeneraliseerd naar ongelijkheden voor vier of meer objecten. Een voorbeeld is hier een ongelijkheid gebaseerd op de tetraëder, waarbij het oppervlakte van een van de zijdes van de tetraëder altijd kleiner of gelijk is aan de som van de oppervlaktes van de drie overige zijdes. Deze ongelijkheden definiëren verschillende meerweg metrieken. In hoofdstuk 13

worden meerweg ultrametrieken bestudeerd en hoofdstuk 14 gaat over hoe twee specifieke drieweg functies gegeneraliseerd kunnen worden. Hoofdstuk 15 beschrijft twee manieren om een resultaat uit hoofdstuk 10 te generaliseren. Dit resultaat vertelt ons dat als een afstandsmaat k aan de driehoeksongelijkheid voldoet, dan voldoet de functie $k/(e+k)$ daar ook aan, waarbij e een positief getal is.

Als laatste bestaat deel IV uit vier hoofdstukken. In dit laatste deel worden meerweg formuleringen van coëfficiënten behandeld. In hoofdstuk 17 zijn de formuleringen functies van de tweeweg informatie, ofwel de coëfficiënten uit deel I. De meerweg coëfficiënten in hoofdstuk 16 zijn geen functies van de tweeweg informatie, maar in dit hoofdstuk wordt een poging om coëfficiënten te formuleren die een kern of basiseigenschap van de tweeweg coëfficiënten generaliseren. Metrische eigenschappen van de meerweg coëfficiënten worden onderzocht in hoofdstuk 18. Hoofdstuk 19 behandelt de drieweg uitbreiding van de Robinson matrices uit hoofdstuk 7, Robinson kubussen genoemd.