Feature network models for proximity data : statistical inference, model selection, network representations and links with related models

Frank, L.E.

Summary in Dutch (Samenvatting)

Feature Netwerk Modellen (FNM) zijn grafische modellen die nabijheidsdata met behulp van features weergeven in een discrete ruimte. *Nabijheidsdata* ontstaan wanneer respondenten gevraagd wordt de gelijkenis tussen paren objecten of stimuli te beoordelen op bijvoorbeeld een 5-punts schaal, waarbij een hoge score aangeeft dat een respondent de twee objecten erg op elkaar vindt lijken. Wanneer een groot aantal respondenten dezelfde objectparen hebben beoordeeld, kunnen de scores gebruikt worden om meer inzicht te verkrijgen in de cognitieve processen die een rol spelen bij het onderscheiden van verschillen en overeenkomsten tussen stimuli. In de psychometrie wordt dit type data vaak met meerdimensionale schaaltechnieken (MDS) geanalyseerd. Bij deze technieken worden de objecten afgebeeld als punten in een laag-dimensionale ruimte en is het doel de geobserveerde nabijheidsdata voor de verschillende object paren zo goed mogelijk te benaderen met afstanden tussen de object punten in die ruimte. Er wordt dan aangenomen dat de *psychologische afstand* tussen objecten, in de vorm van nabijheidsdata voortkomend uit de ervaringen van de respondenten, benaderd kan worden met een *metrische afstand* in een laag-dimensionale ruimte.

De assumptie dat een nabijheidsmaat zich als een metrische afstandsfunctie zou gedragen is al in 1977 door onder anderen Tversky in twijfel getrokken. (Zo weten we allemaal dat de beleving van de lengte van dezelfde treinreis met en zonder spannend boek, duidelijk anders is en dat de benadering in kilometers geen goede weergave is van deze twee verschillende belevingen.) Als alternatief voor de metrische representatie van een nabijheidsmaat, stelde hij het Contrast Model voor, waarbij de afstand tussen objecten wordt weergegeven in termen van een verzameling (set) kwalitatieve eigenschappen en introduceerde hij de *features* die als basis dienen voor het model. Een feature is een prominent kenmerk van een object. De representatie van objecten als een set features leidt volgens Tversky tot betekenisvollere psychologische modellen aangezien de features beschouwd kunnen worden als de elementen van de mentale processen die een rol spelen wanneer respondenten gevraagd wordt objecten te vergelijken, en als zodanig afzonderlijk getoetst kunnen worden.

In het Contrast Model wordt de nabijheidsmaat voor twee objecten weergegeven als de som van de features die beide objecten gemeenschappelijk hebben, de *common features*, en van de features die beide objecten onderscheiden, de *distinctive features*. De nabijheidsmaat wordt in het Contrast Model niet door een afstand benaderd maar door een lineaire combinatie van een set theoretische constructen: de common features worden gevormd door de intersectie te nemen van de feature sets die elk

object beschrijven en de distinctive features worden gevormd door het symmetrisch set verschil (de vereniging minus de intersectie) van de twee feature sets.

Sinds de introductie van het Contrast Model, zijn verscheidene modellen ontwikkeld die ofwel het gemeenschappelijk deel van het model modelleren (het common features model, CF), ofwel het distinctieve gedeelte (het distinctive features model, DF), of een combinatie van beide. De inleiding van deze dissertatie (Chapter 1) geeft een overzicht van al deze modellen. Feature Netwerk Modellen (FNM) concentreren zich op het distinctieve gedeelte en onderscheiden zich van de andere modellen doordat zij een grafische representatie van het DF model geven in de vorm van een netwerk. De objecten (stimuli) worden voorgesteld als punten in een netwerk. De afstand tussen de twee objecten is nu de afgelegde afstand langs de lijnstukken in het netwerk die de twee objecten met elkaar verbinden. De beste benadering van de nabijheidsmaat is dan het kortste pad tussen de twee objecten, die als punten zijn gerepresenteerd in het netwerk.

Het gegeven dat de nabijheidsmaat tussen twee objecten benaderd kan worden door de kortste pad afstand in het netwerk tussen beide objecten, komt voort uit het feit dat een simpele optelling van het aantal elementen van het symmetrisch set verschil een maat oplevert die voldoet aan de axioma's van een metriek, zoals aangetoond door Goodman (1951, 1977) en Restle (1959, 1961). Heiser (1998) heeft aangetoond dat deze afstand in termen van het symmetrisch set verschil ook in termen van coördinaten kan worden voorgesteld en noemde deze de *feature distance*. Deze afstand is gelijk aan de city-block afstand, ook wel de *Manhattan*-metriek genoemd, die zijn naam ontleent aan de manier waarop afstanden worden bepaald in een stad met uitsluitend rechthoekig op elkaar staande straten. De afstand van A naar B in een dergelijk stadsplan is de som van de lengte van de afzonderlijke *blocks* die gepasseerd worden. Dit, in tegenstelling tot de meer gangbare Euclidische metriek, die de afstand van A naar B niet in blocks meet, maar via een directe lijn van A naar B, zoals dit op landkaarten gebeurt. In Manhattan is de Euclidische afstand van A naar B alleen in "vogelvlucht" af te leggen.

In het kader van FNM houdt de city-block metriek in dat de waargenomen nabijheidsmaat tussen stimuli wordt bepaald door de som van het symmetrisch setverschil op elk afzonderlijk feature, dat gelijk staat aan een dimensie in de ruimte. De feature afstand is dan gelijk aan een city-block metriek in een ruimte met binaire coördinaten die gevormd wordt door de features (features zijn immers binaire variabelen die aangeven of een object een bepaalde eigenschap wel of niet heeft). Dit specifieke geval van de city-block metriek wordt ook wel de *Hamming* afstand genoemd. Dat de nabijheidsmaat benaderd wordt door de som van de ongelijkheden op iedere afzonderlijke dimensie, is een uniek kenmerk van de city-block afstand die daarom ook een additieve metriek wordt genoemd.

Als psychologisch model is de additieve metriek aannemelijk indien de stimuli verschillen op discrete, niet vermengbare dimensies (features), zoals in 1950 al aangetoond door Attneave. Een voorbeeld van dergelijke stimuli zijn de bloempot data van Tversky en Gati (1982), waarbij de twee niet vermengbare dimensies gevormd worden door type plant en type bloempot (zie hoofdstuk 1). Het waargenomen verschil tussen de stimuli (verschillende typen planten in verschillende typen bloempotten) is dan afhankelijk van het waargenomen verschil in type

bloempot *plus* het waargenomen verschil in type plant. Wanneer een gewogen optelling van de elementen van het symmetrisch setverschil wordt genomen, kan het relatieve belang van elk feature voor de oplossing worden bepaald aan de hand van het bijbehorende gewicht, de *feature discriminability parameter* (Heiser, 1998). Elk feature splitst de objecten in twee klassen en de feature discriminability parameter geeft aan hoe zeer deze twee klassen van elkaar verschillen. Deze parameters kunnen geschat worden met behulp van een verliesfunctie gebaseerd op het kleinste kwadraten criterium.

De bijdrage van dit proefschrift bestaat onder andere uit de introductie van statistische inferentie in FNM en in het algemeen voor modellen die gebaseerd zijn op features, aangezien er tot op heden in dergelijke modellen nauwelijks tot geen aandacht is besteed aan het beoordelen van de stabiliteit van de oplossingen met behulp van bijvoorbeeld standaardfouten en betrouwbaarheidsintervallen voor de parameters. Hoofdstuk 2 gaat in op de vraag hoe de bijdrage van elk feature beoordeeld kan worden op basis van de feature discriminability parameters voor het geval dat de features a priori bekend zijn op basis van theorie of eerder onderzoek. Het netwerk in FNM biedt een grafische representatie van de relaties tussen de objecten in termen van features en zou tegelijkertijd beschouwd kunnen worden als een psychologisch model voor de mentale representatie van de relaties tussen de objecten zoals deze naar voren komen in de geobserveerde nabijheidsmaten. De netwerk representatie zelf is echter niet toereikend om het psychologisch model te toetsen. Voor dit doel dienen de feature discriminability parameters, die aangeven welk feature het meest bijdraagt aan de beschrijving van de nabijheidsmaten.

Wanneer de features beschouwd worden als (binaire) predictoren kunnen FNM gezien worden als een univariaat multipele regressie model met als regressiegewichten de feature discriminability parameters. Het multipele regressie model biedt weliswaar een uitgangspunt voor statistische inferentie, maar de standaard procedures gaan niet op voor de FNM, aangezien er positiviteits restricties gelden voor de feature discriminability parameters: deze stellen namelijk lijnstukken in het netwerk voor en negatieve lijnstukken hebben immers geen betekenis en leveren dus ook geen adequate beschrijving van een psychologische theorie op. Daarom worden in FNM de parameterschattingen verkregen met behulp van het kleinste kwadraten criterium met positiviteitsrestricties, bekend als *nonnegative least squares*.

Data analyse met restricties op de waarden van de parameters is een veel voorkomend probleem in de statistische literatuur, echter, statistische inferentie voor dit soort problemen is niet eenvoudig omdat in veel gevallen geen statistische theorie beschikbaar is. De theorie over standaard fouten bij least squares met positiviteitsrestricties, beschreven in een bijna vergeten artikel door Liew (1976), blijkt goed toepasbaar te zijn in de context van FNM. In hoofdstuk 2 worden in een Monte Carlo studie deze theoretische standaard fouten, die nog niet eerder getoetst waren in de praktijk, vergeleken met empirische standaard fouten verkregen met de bootstrap methode. De resultaten zijn bemoedigend: de theoretische standaard fouten presteren over het algemeen even goed als de empirische standaard fouten, hetgeen betekent dat volstaan kan worden met het berekenen van een theoretische standaard fout in plaats van een meer tijdrovende bootstrap uit te voeren.

De resultaten van hoofdstuk 2 beperken zich tot het geval dat de features van

tevoren bekend zijn, op basis van theorie of eerder onderzoek. Hoofdstuk 3 biedt een uitbreiding van de statistische inferentie theorie op twee onderdelen. De eerste uitbreiding betreft het geval waarin de features niet van tevoren bekend zijn. Ten tweede blijken de behaalde resultaten voor de theoretische standaard fouten in FNM ook toepasbaar op aan FNM verwante modellen, namelijk de in de psychologie veel gebruikte additieve bomen (*additive trees*), die in de biologie en andere gerelateerde wetenschappen bekend staan als phylogenetische bomen (*phylogenetic trees* of *phylogenies*) en die veelvuldig gebruikt worden om genetische verwantschap aan te tonen tussen organismen. Deze boomstructuren zijn een speciaal geval van FNM wanneer de features een bepaalde structuur hebben, namelijk wanneer de set features bestaat uit uitsluitend geneste features en aangevuld met een set interne nodes, die aangeven waar clusters van objecten zich voordoen, of in het geval van de phylogenetische bomen, een afsplitsing in verschillende organismen plaatsvindt.

De resultaten van hoofdstuk 3 laten zien dat de methode om theoretische standaard fouten en bijbehorende 95% betrouwbaarheidsintervallen te berekenen voor de feature discriminability parameters in FNM ook toepasbaar is voor additieve bomen of phylogenies. Waarbij moet worden vermeld dat de positiviteitsrestricties voor de lijnstukken (geschat als de feature discriminability parameters) bij boomstructuren nog veel meer van belang zijn dan bij FNM aangezien elk lijnstuk exact 1 feature voorstelt. De resultaten berperken zich niet tot het geval waarbij de featurestructuur al bekend is, maar gelden ook voor nog niet bekende featurestructuren. In dit laatste geval is een extra stap nodig in het bepalen van de standaardfouten en bijbehorende 95% betrouwbaarheidsgebieden. In een cross-validatie opzet wordt de steekproef opgedeeld in twee sets, een training set om de boomstructuur (feature structuur) te vinden en een test set waarop de gevonden feature structuur gefit wordt om de standaardfouten en de betrouwbaarheidsintervallen te verkrijgen.

Deze resultaten zijn van belang omdat tot op heden in de psychologische literatuur nog geen statistische inferentie is toegepast op boomstructuren. Het phylogenitisch domein kent echter wel een traditie van statistische inferentie. Opvallend is dat in vrijwel geen enkele methode om phylogenies te fitten, gebruik gemaakt wordt van positiviteitsrestricties op de parameters die de lijnstukken voorstellen. Het gebruik in FNM van het multipele regressie raamwerk gecombineerd met features kan een waardevolle aanvulling voor phylogenetische bomen betekenen. Niet alleen kunnen afsplitsingen van verschillende organismen getoetst worden door een feature aan de set toe te voegen, maar de multipele regressie context biedt ook de mogelijkheid eenvoudig een algemene cross-validatie statistiek te bereken waarmee de fit van verschillende boomstructuren systematisch vergeleken kan worden, ook als deze niet genest zijn. Nadeel van de voorgestelde theoretische standaardfouten is dat zij berusten op de assumpties van normaal verdeelde error termen en homogeniteit van de varianties, beide niet altijd aannemelijk in de praktijk van de data analyse.

Hoofdstuk 4 van dit proefschrift bouwt voort op de situatie waarin de features niet a priori bekend zijn en introduceert een methode waarmee een adequate subset features gevonden kan worden op een manier die verwant is aan het predictor selectie probleem in de context van multipele regressiemodellen. De voorgestelde methode begint met het opstellen van de complete set distinctieve features voor

een gegeven aantal objecten. Aangezien features binaire variabelen zijn, kunnen zij eenvoudig gegenereerd worden met binaire codering. Om een aantal practische redenen, is gekozen voor de Gray code, een speciale vorm van de standaard binaire codering waarbij elk opeenvolgende binaire vector slechts op één bit verschilt van de voorafgaande vector. De tweede stap van de methode is de selectie van een subset features uit de totale set met behulp van de *Lasso*, uitgevoerd met het recentelijk ontwikkeld Least Angle Regression (LARS) algoritme (Efron et al., 2004). De Lasso is een predictor selectie methode voor multipele regressie modellen die een subset predictoren (in dit geval features) selecteert op basis van een compromis tussen model fit en model complexiteit. Het uitgangspunt is niet een optimale subset voor de gegeven data, maar een subset features die goede predictieve eigenschappen bezit, dat wil zeggen, een goede fit zal hebben op een nieuwe steekproef.

Voor gebruik met FNM moest de Lasso eerst aangepast worden om aan de positiviteitsrestricties te voeldoen en dit heeft geleid tot de ontwikkeling van de Positieve Lasso. De methode is ook toe te passen op a priori gegeven features en aangezien de beste subset features geselecteerd wordt, kan dit dienen als alternatief voor het kiezen van de features die het meest bijdragen aan de oplossing met behulp van betrouwbaarheidsintervallen. Een nadeel is echter dat de methode alleen goed werkt voor aantallen objecten niet groter dan 22 omdat de te genereren complete set distinctieve features dan 2 miljoen bedraagt en deze in de huidige applicatie niet door de computer bewerkt kan worden.

Naast de verschillende bijdragen op het gebied van statistische inferentie en modelselectie in de hoofdstukken 2 tot en met 4, richt het laatste hoofdstuk van dit proefschrift zich op de netwerk representatie van FNM. Hoofdstuk 5 laat zien dat de netwerk representatie universeel is voor alle city-block modellen. Dit resultaat komt voort uit het feit dat de city-block afstand een additieve metriek is en maakt gebruik van een aantal kernelementen van de netwerk representatie, zoals metrische segment additiviteit, betweenness en de interne nodes. Deze netwerk representatie is niet alleen universeel voor alle city-block modellen gebaseerd op distinctieve features, maar geldt ook voor modellen gebaseerd op common features, zoals additief clusteren, hierarchische boomstructuren en additieve boomstructuren. Een overzicht van de relaties tussen deze modellen is te zien in Figuur 5.10.

Eerder in deze samenvatting werd gemeld dat FNM een *distinctive features* model (DF) is en in zekere zin tegenovergesteld aan het *common features model* (CF). Er is een duidelijke relatie tussen de twee modellen: Sattath en Tversky (1987) en later Carroll en Corter (1995) hebben aangetoond dat het CF model en het DF model in elkaar vertaald kunnen worden. Echter dit theoretisch resultaat was nog niet in de praktijk van data analyse uitgeprobeerd. In hoofdstuk 5 wordt de translatie van CF naar DF toegepast op empirische data. Het blijkt mogelijk te zijn voor elk gefit CF model een even goed fittend DF model te vinden met gebruik making van dezelfde common features en feature gewichten en hetzelfde aantal onafhankelijke parameters. Een belangrijk resultaat dat hieruit volgt is dat een model dat het CF model met het DF model combineert, ook uitgedrukt kan worden als een combinatie van twee afzonderlijke DF modellen, waarmee het DF model een algemener model blijkt te zijn dan het CF model.

Hoofdstuk 6 sluit het proefschrift af met een algemene conclusie en een discussie.

De voor- en nadelen van het gebruik van de theoritsche standaard fouten in FNM worden uiteengezet en vergeleken met de bootstrap standaard fouten, waarbij het vooral gaat om de assumpties van normaliteit, homogeniteit van de varianties en het probleem van alpha inflatie bij het gebruik van betrouwbaarheidsintervallen voor meerdere features. Mogelijke oplossingen en ideeën voor vervolgonderzoek worden aangedragen. Naast de onderwerpen van statistische inferentie en model selectie, worden ook de netwerkrepresentatie, de embedding van het netwerk in een ruimte van lagere dimensionaliteit besproken.