Reliable and fair machine learning for risk assessment

Pereira Barata, A.P.

Samenvatting

Het inschatten van risico's is het hoofdonderwerp van dit proefschrift. In Hoofdstuk 1 introduceren we in algemene zin de recente ontwikkelingen op het gebied van kunstmatige intelligentie (Artificial Intelligence, AI). We kijken daarbij vooral naar het gebruik van AI-technieken bij het inschatten van risico's. In het bijzonder besteden we aandacht aan de praktische inzetbaarheid van AI-programma's bij de Inspectie Leefomgeving en Transport (ILT). Binnen dit orgaan van de Nederlandse overheid is grote behoefte aan een *eerlijke* en *betrouwbare* inschatting van risico's. Het onderzoekswerk is een bijdrage aan de paradigma-verschuiving die in 2016 in gang gezet is bij de start van het programma Anders Omgaan Met Data (AOMD). Ons werk draagt dan ook bij aan de voortdurende verbetering van het data-gestuurd werken door de ILT-inspecteurs, met name door middel van de inzet van *machine learning*.

De focus van dit proefschrift ligt expliciet op classificatiemodellen. Data die door de inspectie wordt gegenereerd, verzameld en geclassificeerd staan vaak (a) in tabelvorm en (b) bevatten diverse datatypen. In het onderzoekswerk richten we ons met name op data in beslisbomen. Daarbij onderzoeken we twee problemen die regelmatig voorkomen bij classificatie: (1) *bias in de data*; en (2) *datakwaliteit*. Bij datakwaliteit kijken we in het bijzonder naar welke data er missen (*missing data*) en ruis (*noise*). Het doel is om adequaat presterende classificatiemodellen te ontwerpen die deze problemen oplossen, en dus zowel *eerlijk* als *betrouwbaar* zijn. De probleemstelling van dit proefschrift luidt dan ook als volgt.

> **Probleemstelling:** *Hoe kunnen we bijdragen aan eerlijke en betrouwbare machine learning methoden, die inzetbaar zijn voor de inschatting van risico's door de Inspectie Leefomgeving en Transport (ILT)?*

De probleemstelling wordt onderverdeeld in drie onderzoeksvragen. Hieronder bespreken we deze vragen, geven dan de precieze formulering en vatten tenslotte de belangrijkste bevindingen samen.

Bij *missing data* speelt een precieze karakterisering van het mechanisme (waardoor ontbreken de data eigenlijk?) een belangrijke rol. We onderscheiden drie mechanismen: (1) MCAR (Missing Completely At Random); (2) MAR (Missing At Random); en (3) MNAR (Missing Not At Random). Afhankelijk van het mechanisme waar we mee te maken hebben, zien we dat de efficiency van de *missing-data* technieken nogal varieert. In ons onderzoek meten we de performance van de verschillende technieken en kijken we naar de relatie tussen de performance en het gekozen *machine learning* model.

In een *niet-MCAR*-scenario (dus MAR en MNAR) lijkt een adequate benadering te zijn: het coderen van ontbrekende gegevens door een *extra variabele* (of attribuut). Dit is de zogeheten *missing-indicator* methode. Bij het *MCAR*-scenario heeft daarentegen —zo lijkt het— *imputatie* van de *missing data* de voorkeur. In de werkelijke wereld ontbreekt data echter zelden volgens het MCAR mechanisme; bovendien is het testen of data ontbreekt volgens het MCAR mechanisme moeilijk en soms ook nog onbetrouwbaar. Daarom luidt de eerste onderzoeksvraag als volgt.

> **Onderzoeksvraag 1:** *Welke combinatie van methoden voor het behandelen van missing data en machine learning algoritmen moet worden gebruikt om, los van het precieze missing data mechanisme, een adequaat presterend model te verkrijgen?*

In Hoofdstuk 2 wordt het MCAR mechanisme bestudeerd. Door op een gecontroleerde manier verschillende gradaties van *missing data* volgens het MCAR mechanisme te genereren, kan empirisch worden vergeleken hoe combinaties van bepaalde methoden voor het behandelen van *missing data* en *machine learning* algoritmen presteren. De resultaten laten zien dat onder het MCAR mechanisme *imputatiemethoden doorgaans beter presteren dan de missing-indicator methode*.

Een belangrijke bevinding is dat voor *gradient boosting* classificatie-algoritmen op basis van beslisbomen, de verschillen in prestaties verwaarloosbaar lijken. Het antwoord op de eerste onderzoeksvraag is dan ook dat de missing-indicator methode, in combinatie met een beslisboomalgoritme — met name op basis van *gradient boosting* — moet worden gebruikt, ongeacht het precieze mechanisme dat aangeeft welke data er ontbreekt.

Door ruis in de data zullen in het algemeen de classificatie-prestaties verslechteren. Ruis kan voorkomen als *featureruis* of als (klasse) *labelruis*. Vooral labelruis heeft een sterke invloed. Bij labelruis bestaan drie mechanismen: (1) NCAR (Noise Completely At Random); (2) NAR (Noise At Random); en (3) NNAR (Noise Not At Random). De aanpak bij ruis in de data richt zich in het algemeen op het bepalen van detectiescores voor de ruisdata.

De detectiescores worden traditioneel bepaald door gebruik te maken van classificatiescores die zijn geleerd op de ruis bevattende dataset. Binnen het scenario van de Inspectie kunnen data met labelruis juist een aanwijzing zijn voor afwijkingen, vooral wanneer er ook featureruis is (d.w.z., volgens NNAR mechanisme); bijv., bedrijven die afvaltransport-rapporten manipuleren om de kosten ervan te verlagen. Het is van belang om niet alleen deze afwijkingen op te sporen, maar ook om zo goed mogelijk classificatiemodellen te leren uit de beschikbare gegevens.

Binnen het scenario van de Inspectie kunnen data met labelruis juist een aanwijzing zijn voor afwijkingen, vooral wanneer er ook featureruis is (d.w.z., volgens NNAR mechanisme); b.v. bedrijven die afvaltransportrapporten manipuleren om de kosten ervan te verlagen. Het is van belang om niet alleen deze afwijkingen op te sporen, maar ook om zo goed mogelijk classificatiemodellen te leren uit de beschikbare gegevens.

De tweede onderzoeksvraag is samengesteld uit onderzoeksvraag 2(a) en onderzoeksvraag 2(b) en luidt in zijn geheel als volgt.

> **Onderzoeksvraag 2:** *Gegeven data met labelruis, hoe kan die data met ruis (a) adequaat worden opgespoord, en (b) worden gebruikt om een model met adequate performance te leren?*

In Hoofdstuk 3 introduceren we de term *crosslier*, die een sample aanduidt met afwijkende features. Preciezer gezegd crossliers zijn een speciaal type outlier die afwijkend zijn ten opzichte van een bepaalde categorie. Het zijn *samples* die labelruis vertonen met betrekking tot een bepaalde categorische feature, waarbij de ruis mogelijk het NNAR-mechanisme volgt. Om crossliers te detecteren, stellen we de onderzoeksmethode EXPOSE voor. De EXPOSE-methode evalueert *samples* op een cross-validatie (CV) manier, zodat uiteindelijk alle *samples* worden geëvalueerd. In de *CV loops* wordt elke training set gebruikt om een goed gecalibreerde *classifier* te produceren daarbij gebruik makend van Platt-schaling. De *classifier* wordt vervolgens ingezet op de bijbehorende *test samples*. Van de *classificatie output* $f(x)$, is de corresponderende crosslierscore $-\log_2[f(x)]$. Om de prestaties van onze aanpak te evalueren, valideren we de EXPOSE-methode in een gecontroleerde experimentele omgeving. Daarmee beantwoorden we onderzoeksvraag 2(a).

Hoofdstuk 4 bouwt voort op de uitgangspunten van EXPOSE en levert een aanpak op voor het leren van een classificatiemodel met gegevens die ruis bevatten. We noemen deze samengestelde methode DENOISE. De DENOISE methode bestaat uit twee stappen. Eerst worden goed gecalibreerde kansen berekend voor elk *sample* volgens de EXPOSE-methode. Ten tweede worden de kansen gebruikt om individuele *sample* gewichten te genereren als de *log-odds* van de gecalibreerde *sample* kans.

Tezamen met een logistische verliesfunctie die wordt toegepast op de leeralgoritme van de *gradient boosting* beslissingsboom ontstaat een adequaat presterende ruisbestendige *classifier*. In een gecontroleerde experimentele omgeving valideren we vervolgens onze methode. Daarmee beantwoorden we onderzoeksvraag 2(b).

Leren van gegevens met een bias leidt tot modellen met een bias. Zelfs wanneer de omstandigheden van het verzamelen van gegevens ideaal zijn, kan er nog steeds vertekening in de gegevens optreden als gevolg van historische factoren (bijv., de loonkloof tussen mannen en vrouwen). Om dit probleem aan te pakken, kunnen verschillende eerlijke (*fair*) *machine learning*-technieken worden gebruikt. Het doel van deze technieken is om modellen te genereren waarvan de output onafhankelijk is van een gevoelig attribuut, zoals geslacht; dat wil zeggen, eerlijke (*fair*) modellen. Let wel, er bestaan verschillende maatstaven voor eerlijkheid, waarvan *strong demographic parity* analoog is aan de AUC prestatiemaatstaf voor classificatie *performance*. Over het algemeen geldt dat *fairness* op gespannen voet staat met de voorspellende kracht van *performance*. Het betreft hier de *trade-off* tussen *performance* en *fairness* gegeven een bepaald criterium.

Een voorbeeld is als schepen onder de vlag van een bepaald land varen, dan kunnen ze worden beschouwd als schepen met een hoger risico. Dit kan betekenen dat ze daarom meer aandacht krijgen dan andere schepen, wat kan leiden tot *confirmation bias*. Aangezien de landsvlag gemakkelijk kan worden gewijzigd, kunnen bedrijven het inspectie-selectieprotocol betrekkelijk gemakkelijk omzeilen. Op grond van deze observaties (en gevolgtrekkingen) formuleren we de onderzoeksvraag 3 als volgt.

> **Onderzoeksvraag 3:** *Hoe kunnen we een model bouwen van vooringenomen gegevens, zodat het door de domeinexpert kan worden aangepast met betrekking tot de trade-off tussen performance en fairness?*

In Hoofdstuk 5, we stellen een eerlijk algoritme voor om een beslisboom te leren voor *strong demographic parity*. We doen dit door een samengesteld splitsingscriterium te definiëren, genaamd SCAFF (Splitsings Criterium AUC For Fairness). SCAFF kan worden *getuned* met betrekking tot de *trade-off* tussen *performance* en *fairness*. Tegelijkertijd kan het mechanisme gebruik maken van verschillende gevoelige attributen, waarvan de waarden multi-categorisch kunnen zijn.

SCAFF wordt derhalve gedefinieerd als een gewogen lineaire combinatie van (a) de traditionele AUC classificatieperformance , en (b) de *strong demographic parity* die geschaald is in overeenstemming met het bereik van de AUC. We noemen deze AUC een *sensitive AUC*. Hoe dichter de sensitive AUC bij 0.5 ligt, des te eerlijker is het model.

Door een orthogonaliteitsparameter $\Theta \in [0, 1]$ op te nemen, die geïmplementeerd wordt als een *elastisch net*-achtig gewicht op de *trade-off* tussen *performance* en *fairness*, kan de waarde worden aangepast; $\Theta = 0$ genereert een traditioneel (potentieel) niet-eerlijke classificatie, en het verhogen van Θ vergroot de *fairness* van het uiteindelijke model. Door SCAFF te vergelijken met andere eerlijke splitsingscriteria in een gecontroleerd experiment, valideren we onze aanpak en beantwoorden we de derde onderzoeksvraag.

De *conclusie* van het proefschrift wordt gegeven in Hoofdstuk 6. Door de nadruk op betrouwbare AI in Europa zien we een verschuiving van het huidige paradigma van risicobeoordeling naar een meer data-gestuurde beoordeling. Het resultaat is een delicate maar haalbare onderneming via eerlijke en betrouwbare *machine learning*. Met inachtneming van het risicovolle karakter van de beoordelingsactiviteiten en de kenmerken van de data die door de beoordelingen worden gegenereerd, kunnen technische methoden de geschiktheid van de uiteindelijk geleerde modellen garanderen. Met name de problemen die samenhangen met het leren van een classificatiemodel uit data met een bias (en met lage kwaliteit) kunnen hiermee worden aangepakt.