Exceptional Model Mining

Duivesteijn, W.

Nederlandse Samenvatting

Wanneer we een grote verzameling ruwe data hebben, poogt het deelgebied van Informatica genaamd *Datamining* er informatie uit te destilleren die kan worden geïnterpreteerd door de eindgebruiker, en bij voorkeur kan worden benut in het domein waar de gebruiker in geïnteresseerd is. Met de toegenomen rol van het internet in het dagelijks leven, en in het bijzonder de opkomst van smartphones, genereert iedere persoon en ieder bedrijf enorme hoeveelheden gegevens. Er ligt een taak voor de wetenschappelijke gemeenschap om methoden te ontwikkelen die een onoverzichtelijke hoeveelheid data nemen en er klompjes zinnige informatie uitpersen.

Elementen aanwijzen die afwijken van de norm is een belangrijke taak. Het meeste dataminingonderzoek in deze richting concentreert zich op afwijkingen *detecteren*. In Lokale Patroonmining zijn we echter niet tevreden met het aanwijzen van afwijkende elementen in de data. In plaats daarvan zoeken we naar *subgroepen*: coherente deelverzamelingen die kunnen worden *beschreven* door een klein aantal voorwaarden op attributen van de data. Het bestaan van zulke beschrijvingen maakt de resulterende afwijkende subgroepen meer actiegericht: als we bijvoorbeeld een farmaceutisch bedrijf vertellen dat vijf bepaalde personen slecht reageren op een medicijn, dan kan het bedrijf daar minder mee dan ze zouden kunnen als we ze kunnen vertellen dat de groep rokers slecht reageert op het medicijn.

"Afwijken van de norm" is multi-interpretabel. Traditioneel is zo'n uitzonderlijkheid gedefinieerd op basis van veelvoorkomendheid, of op basis van een afwijkende verdeling van één doelattribuut. Deze concepten omvatten niet alle potentieel interessante afwijkingen. Om deze algemene interesse vorm te geven, hebben we Exceptional Model Mining (het graven naar uitzonderlijke modellen) ontwikkeld.

De eerste stap van het EMM-raamwerk is het verdelen van de attributen in twee delen: een deel om de subgroepen op te *definiëren* (de *beschrijvers*), en een deel om de subgroepen op te *evalueren* (de *doelwitten*). Dan selecteren we een *modelklasse* over de doelwitten, en ontwerpen we een *kwaliteitsmaat* over de modelklasse. Ten slotte gebruiken we de al bestaande Subgroup Discovery methode om de beschrijver-ruimte te doorzoeken naar subgroepen die goed presteren volgens de kwaliteitsmaat. De modelklasse vertegenwoordigt het samenspel tussen de doelwitten, en de kwaliteitsmaat meet de uitzonderlijkheid van modelparameters. Bijvoorbeeld kunnen we zo subgroepen vinden waarvoor twee doelwitten ongebruikelijk correleren, waarvoor een Bayesiaans netwerk een afwijkende structuur heeft, of waarvoor een regressiemodel een uitzonderlijke parametervector heeft.

Het doorzoeken van de beschrijver-ruimte kost veel rekenkracht: we zoeken interessante deelverzamelingen van de data, en je kunt je voorstellen dat er van een grote dataverzameling veel kandidaat-deelverzamelingen zijn. Voor de EMM-instantie met regressie als modelklasse hebben we bovengrenzen afgeleid op de kwaliteit van een kandidaat, die we kunnen uitrekenen zonder de parametervector zelf uit te rekenen. Met behulp van deze bovengrenzen kunnen we deze dure laatste rekenstap vermijden voor maximaal 40% van de kandidaten, waardoor het hele proces sneller verloopt.

Door EMM-instanties hebben we subgroepen gevonden die weersomstandigheden betreffen waaronder voedselketens ontsporen, subgroepen die de economische wet van vraag en aanbod tarten, subgroepen die het dempende effect van vakbonden op de salarisverdeling illustreren, et cetera. Daarnaast hebben we een test ontwikkeld of zulke subgroepen valse ontdekkingen zijn: wanneer we een toets vaak uitvoeren zullen we uiteindelijk iets schijnbaar significants vinden, puur door toevalseffecten. Door een model te bouwen van kunstmatig gegenereerde valse ontdekkingen, en gevonden subgroepen hiermee te vergelijken, kunnen we inschatten of het waarschijnlijk is dat onze subgroepen ook valse ontdekkingen zijn.

We hebben bepaald of de resultaten van één EMM-instantie (met een Bayesiaans netwerk als modelklasse) ook kunnen worden gebruikt om de doelwitten beter te *voorspellen* wanneer we nieuwe data binnenkrijgen. We laten zien dat we door uitzonderlijk samenspel tussen de doelwitten met EMM te vatten, soms in staat zijn deze voorspellingen te verbeteren.