Universiteit Leiden
The Netherlands

Massively collaborative machine learning

Rijn, J.N. van

Dutch Summary

Veel wetenschap is gericht op het bouwen van modellen. Modellen zijn schematische (vaak versimpelde) weergaven van de werkelijkheid. Deze modellen zijn gebaseerd op waargenomen data uit het verleden, en maken vaak voorspellingen voor de toekomst (nog niet waargenomen data). De mens werkt onbewust met modellen; vrijwel alle informatie die wij tot ons nemen verwerken wij tot een model. Een klassiek Nederlands voorbeeld hiervan is het weer. Wanneer we 's ochtends uit het raam kijken en er hangen donkere wolken, zouden we geneigd zijn te denken dat het een regenachtige dag wordt. Andersom, wanneer de lucht helder blauw is, zouden we wellicht denken dat het een mooie dag gaat worden. Het model dat wij gebruiken stelt dus: "Als het bewolkt is, is er een grote kans op regen; en als de lucht helder is, is er een grote kans op een mooie dag." Op een of andere manier hebben mensen de capaciteit om ervaringen uit het verleden om te zetten in zulke modellen, en daarmee voorspellingen te maken voor de toekomst. Dit noemen we ook wel 'leren'.

Tegenwoordig zijn er veel technieken waardoor computers ook kunnen leren. Dit vakgebied heet Machine Learning, en wordt in het dagelijks leven op veel plekken toegepast. Voorbeelden hiervan zijn:

- sociale media: de computer bepaalt welke advertenties er aan een specifieke gebruiker getoond moeten worden
- aandelenmarkt: de computer bepaalt welke aandelen gekocht of verkocht moeten worden
- spamfilters: de computer bepaalt voor inkomende berichten of het spam of geen spam is

Aangezien veel van deze onderwerpen 'hot topics' zijn, is er veel onderzoek gestoken in het ontwikkelen van computer programma's die zulke modellen kunnen bouwen

(algoritmes). Er bestaan bijzonder veel van zulke algoritmes, en vaak hebben deze algoritmes ook nog verschillende opties die de werking subtiel beïnvloeden (parameters). Daarnaast is er sluitend wiskundig bewijs dat er niet één algoritme bestaat dat goed werkt op alle datasets. Dat maakt het voor eindgebruikers die deze algoritmes op hun data willen toepassen niet makkelijk om het juiste algoritme te kiezen.

Het vakgebied van 'meta-learning' heeft als doel orde in deze chaos te scheppen. Het houdt zich bezig met het in kaart brengen van welke algoritmes het goed doen op wat voor soort data. Dit wordt veelal op experimentele basis gedaan. Een veel gebruikte methodologie is om zo veel mogelijk algoritmes op zo veel mogelijk datasets uit te proberen, en op basis van de resultaten een model te leren welke op welke datasets bepaalde algoritmes goed werken (vandaar de naam: meta-learning). Hoewel dit in de praktijk goed werkt is de tijd een limiterende factor. Vaak nemen deze experimenten veel tijd in beslag, waardoor het onderzoek noodgedwongen kleinschaliger is dan wenselijk.

Om dit probleem op te lossen hebben we OpenML ontwikkeld. Dit is een online database waarop onderzoekers hun experimentele data met elkaar kunnen delen, om op die manier grootschaliger meta-learning onderzoek te kunnen doen. Wanneer eerdere experimentele resultaten opgeslagen en vrij toegankelijk beschikbaar zijn, is het niet langer nodig om deze tijdrovende experimenten op te zetten, maar kunnen vragen over de werking van algoritmes direct grondig worden beantwoord.

In dit proefschrift wordt beschreven hoe fundamentele meta-learning problemen als 'welk algoritme heeft de voorkeur op een bepaald soort data?', 'wat is het effect van een bepaalde parameter op de prestaties van een algoritme?' en 'hoe werkt een zojuist nieuw ontwikkeld algoritme ten opzichte van alle bestaande algoritmes?' met behulp van OpenML met relatief weinig moeite kunnen worden beantwoord.