Minimum Description Length Model Selection

de Rooij, S.

Publication date
2008

Samenvatting

Modelselectie is een ongrijpbaar onderwerp in de leertheorie en statistiek. Op het eerste gezicht lijkt het probleem duidelijk: hoe moeten we beslissen welke verzameling van kansverdelingen het best overeen komt met de beschikbare observaties. Deze vraag duikt telkens weer op in allerlei verschillende contexten, waaronder het toetsen van hypothesen in het algemeen (welke van deze psychologische modellen beschrijft het best hoe mensen zich gedragen?) tot meer concrete toepassingen (een polynoom van welke graad moeten we kiezen om de trend in deze gegevens te beschrijven? Welke "lossy" representatie van dit plaatje beschrijft de structurele eigenschappen van het origineel het best?). Kortom, modelselectie is een sleutelprobleem in vele verschillende toepassingen. De one-size-fits-all-oplossingen die gebaseerd zijn op het Minimum Description Length (MDL) principe en de nauw verwante Bayesiaanse statistiek worden daarom veel gebruikt.

Bij nadere beschouwing blijkt dat de vele toepassingen van modelselectie op essentiële punten verschillen. Ze kunnen worden onderscheiden op basis van technische eigenschappen (zijn de modellen in elkaar bevat? Parametrisch? Telbaar?), maar ook op basis van a priori aannames (nemen we aan dat het proces dat de gegevens genereert een element is van een van onze modellen of niet?), alsmede de oorspronkelijke motivatie voor het doen van modelselectie (willen we het model identificeren dat het proces bevat dat de gegevens genereert, of willen we een model selecteren waarvan we mogen hopen dat het toekomstige uitkomsten goed zal voorspellen?). De meest wenselijke methodologie hangt in de praktijk vaak af van dergelijke kwesties, nog los van praktische afwegingen zoals of de relevante grootheden al dan niet efficiënt kunnen worden uitgerekend.

Op allerlei manieren is aangetoond dat het gebruik van MDL/Bayesiaanse modelselectie leidt tot goede prestaties in vele contexten; in dit proefschrift wordt een aantal van de raadselachtige problemen en beperkingen van de methodologie onderzocht, die niettemin in de loop van de tijd aan het licht zijn gekomen. Ook wordt het toepassingsdomein van MDL/Bayesiaanse modelselectie uitgebreid door het te koppelen aan andere onderwerpen in de machine learning en statistiek.

Om MDL toe te kunnen passen moeten zogenaamde universele codes of uni-

versele kansverdelingen worden toegewezen aan alle modellen die worden overwogen. De code die daarbij volgens sommige literatuur de voorkeur heeft is de Normalised Maximum Likelihood (NML) of Shtarkov code. Het blijkt echter dat deze code voor vele modellen leidt tot oneindige codelengtes, waardoor de prestaties van de verschillende modellen niet meer met elkaar vergeleken kunnen worden. Dit eerste probleem met MDL modelselectie wordt onderzocht in hoofdstuk 2, waarin we computerexperimenten uitvoeren om de prestaties te meten van enkele van de beschikbare alternatieven voor de NML code. Een van de meest interessante resultaten is dat de zogenaamde prequentiële plug-in code leidt tot inferieure modelselectieprestaties. De prequentiële plug-in code wordt vaak gezien als een handig alternatief voor andere codes zoals de NML code, omdat het vaak veel makkelijker is uit te rekenen. Het werd vrij algemeen aangenomen dat de resulterende codelengtes vergelijkbaar waren met die van andere universele codes zoals NML of 2-part codes, maar uit onze experimenten blijkt dus dat dit niet onder alle omstandigheden het geval is. In hoofdstuk 3 wordt aangetoond dat de redundantie van de prequentiële plug-in code fundamenteel verschilt van die van andere universele codes in het geval dat het proces dat de gegevens produceert geen element is van het model. Dit betekent dat prequentiële plug-in codes met beleid moeten worden toegepast in modelselectie.

Het derde probleem dat wordt behandeld in dit proefschrift is dat MDL en Bayesiaanse modelselectie normaal gesproken geen rekening houden met het volgende: zelfs in het ideale geval waarin een van de beschikbare modellen "waar" is (het proces dat de gegevens produceert bevat), en zelfs als het gegevens producerende proces stationair en ergodisch is, dan nog *hangt het af van de hoeveelheid beschikbare gegevens* welk van de beschikbare modellen de beste voorspellingen van toekomstige uitkomsten levert. Ruwweg kan worden gesteld dat, als de hoeveelheid beschikbare gegevens klein is, dat dan eenvoudige modellen vaak betere voorspellingen leveren dan het complexere "ware" model (i.e., het kleinste model dat het gegevens producerende proces bevat). Als vervolgens op een gegeven moment de hoeveelheid beschikbare gegevens zo groot is geworden dat het ware model het beste begint te voorspellen, dan heeft het eenvoudigere model al zo lang zoveel beter gepresteerd dat het soms zeer lang kan duren voordat het door de MDL/Bayesiaanse modelselectieprocedure wordt verworpen. Hoofdstuk 5 beschrijft dit verschijnsel, alsmede een nieuwe modelselectieprocedure die het ware model prefereert vrijwel zodra er voldoende gegevens beschikbaar zijn dat dat model de beste voorspellingen begint te leveren. Deze nieuwe procedure is *consistent*, wat betekent dat (onder milde condities) het ware model wordt geselecteerd met kans 1 mits er voldoende gegevens beschikbaar zijn. We tonen ook aan dat voorspellen op basis van deze modelselectieprocedure leidt tot optimaal snelle convergentie: de cumulatieve KL-risk is bewijsbaar zo laag als die van willekeurig welke andere modelselectieprocedure. De methode is gebaseerd op de *switch-verdeling*, die kan worden uitgerekend met behulp van een efficiënt algoritme voor expert tracking. Meer eigenschappen van deze switch-verdeling

worden behandeld in hoofdstuk 4, waarin we een overzicht geven van deze en andere algoritmes voor expert tracking, en waarin we laten zien hoe zulke algoritmes handig kunnen worden geformuleerd in termen van Hidden Markov Models.

In hoofdstuk 6 tenslotte evalueren we de nieuwe theorie van algoritmische rate-distortion experimenteel. Deze theorie werd recentelijk voorgesteld door Vitányi en Vereshchagin als een alternatief voor klassieke rate-distortiontheorie. Ze maakt analyse mogelijk van de structurele eigenschappen van individuele objecten, en vereist niet dat er een objectbron wordt gespecificeerd in de vorm van een kansverdeling. In plaats daarvan wordt algoritmische rate-distortion gedefinieerd in termen van Kolmogorovcomplexiteit, die niet berekenbaar is. Om deze theorie toch in de praktijk te kunnen toetsen benaderen we de Kolmogorovcomplexiteit met de gecomprimeerde grootte van een algemeen toepasbaar datacompressiealgoritme. De zo verkregen praktische aanpak is in feite een generalisatie van MDL modelselectie.

De perspectieven die in dit proefschrift worden geboden op vele aspecten van MDL/Bayesiaanse modelselectie dragen bij tot een dieper begrip van de verbanden tussen modelselectie en diverse onderwerpen als universal learning, voorspellen met advies van experts, rate-distortiontheorie en Kolmogorovcomplexiteit.