Intrinsically Interpretable Machine Learning In Computer Aided Diagnosis

S. Ghosh, Sreejita

DOI:
10.33612/diss.175627883

Samenvatting

Door de relatief gemakkelijke beschikbaarheid van allerlei soorten data en de technologische vooruitgang de afgelopen tijd vertrouwen we misschien op *AI-algoritmen*, net zoals onze voorouders of sommige van onze familie en/of vrienden die we als blinde 'gelovigen' beschouwen, vertrouwen op de zogenaamde goddelijke machinaties en de hogepriesters. Waarom zouden we onszelf dan minder blind en goedgelovig vinden als ook wij vertrouwen op dat wat ondoorgrondelijk is? Er is zeker complexe wiskunde bij gekomen. die is gevalideerd in de algemeen aanvaarde topbladen en conferenties. Maar wat is de zekerheid dat het black-box model geleerd heeft wat het werkelijk had moeten leren? Een alarmerend schadelijk effect van het vertrouwen op dergelijke black-box modellen is wanneer dergelijke modellen worden toegepast in de gezondheidszorg, zoals aan de orde komt in (Obermeyer et al. 2019). In deze studie werd onderzocht hoe een black-box model, dat werd vertrouwd en dus gemplementeerd in het gezondheidszorgsysteem in een bepaald land, feitelijk medisch irrelevante kenmerken had geleerd om kritieke medische beslissingen te nemen. De studie beschrijft hoe het model in feite bevooroordeeld was en had geleid tot het verlies van medische zorg voor mensen die tot een bepaalde etniciteit behoren. Om dergelijke vooroordelen te voorkomen hebben we AI modellen of algoritmen nodig waarvan de beslissingen zijn te verklaren.

Deze dissertatie presenteert een set intrinsiek interpretabele machine learning modellen die werden toegepast op medische datasets uit de echte wereld, een synthetische dataset, en een publiek beschikbare dataset uit de UCI repository, die de uitdagingen van heterogene metingen, onevenwichtige klassen, en systematische missingness. De interpreteerbaarheid van de gepresenteerde set van classifiers zijn in termen van (1) het vertrouwen van de classificator in het toekennen van een klasse-etiket aan een voorgesteld steekproef (in plaats van alleen maar crisp labels), (2) duidelijke visualisatie van de beslissingsgrenzen van een gepresenteerd probleem zoals aangeleerd door de classificator, (3) impliciete berekening van kenmerkrelevantie, en (4) extractie van typische profiel(en) van elk van de aangeleerde klassen (prototypes) door de classifier. Deze nieuwe gentroduceerde set classifiers zijn nearest prototype based classifiers (NPCs) die behoren tot de familie van de Learning Vector Quantization (LVQ). Tijdens de training leert een LVQ-model een of meer (door de gebruiker bepaald) representatief steekproef van elke klasse in de trainingsset; deze worden prototypes genoemd. Wanneer een nieuw steekproef aan het model wordt gepresenteerd, wordt de afstand of dissimilarteit

van dat steekproef ten opzichte van elk van de prototypes wordt berekend en krijgt het steekproef het klasse-etiket toegewezen van het prototype waarop het het meest lijkt, volgens de winner takes all (WTA)-strategie. Latere versies van LVQ hebben een adaptieve dissimilariteit gebruikt in plaats van absolute dissimilariteit, aangezien niet alle kenmerken van een trainingsset even relevant zijn voor een classificatieprobleem. Het leren van de feature-relevantie tijdens de training leidt tot een embedded feature selectie.

De meest populaire keuze van metriek voor de berekening van adaptieve dissimilariteit is de Euclidische afstand. Wij hebben echter onderzocht en geconstateerd dat in aanwezigheid van missingness en een laag aantal training samples angle basd (cosinus) dissimilariteit robuuster is dan Euclidische afstand. Dit proefschrift presenteert eerst de op angle-dissimilarity gebaseerde varianten van Generalized Relevance LVQ (GRLVQ), Generalized Matrix Relevance LVQ (GMLVQ), Local metric tensor LVQ (LGMLVQ) en Localized Limited Rank Metric (LLiRAM) LVQ. Vervolgens worden probabilistische varianten van de GMLVQ en hoek GMLVQ gepresenteerd. Deze modellen geven het vertrouwen van de classificator in het toekennen van verschillende klassen aan een steekproef. Om onevenwichtigheid van klassen te behandelen worden in dit proefschrift ook twee strategien gentroduceerd: (1) een geodetische variant van Chawla et al's techniek van synthetische oversampling van de minderheidsklasse steekproeven, en (2) door de gebruiker gedefinieerde bestraffing van misclassificatie van een steekproef uit een minderheidsklasse naar een onjuiste klasse. Deze nieuw ontwikkelde modellen leveren niet alleen prestaties die vergelijkbaar zijn met die van Random Forests, maar helpen ook bij het extraheren van medische kennis uit de dataset waarop ze getraind zijn. Random Forest is een ensemble van Decision Trees (DT's), wat het krachtig maakt, terwijl het compromitteert met de interpreteerbaarheid van DTs. In deze thesis introduceerden we een geodetische middelingstechniek die de kracht van ensembling combineert met behoud van de interpreteerbaarheid van de LVQ modellen.

De prestaties van deze nieuw ontwikkelde en gepresenteerde intrinsiek interpreteerbare modellen bleken veelbelovend voor bredere toepassingen in industrien van verschillende sectoren. Gehoopt wordt dat vooral industrien die werken aan antropocentrische toepassingen die gebruik maken van AI, baat zullen hebben bij deze modellen, vooral na de recente regelgeving van de Europese Commissie die stelt dat AI-systemen met een hoog risico, zoals systemen die een directe impact hebben op het leven van mensen, zo transparant mogelijk moeten zijn, waarbij de output van deze systemen gemakkelijk te interpreteren is. De dissertatie behandelt ook de verschillende niveaus van interpreteerbaarheid en tracht de lezers te motiveren om na te gaan wat een model heeft geleerd in plaats van zich uitsluitend tevreden te stellen met de resultaten van het model.