Data-driven machine learning and optimization pipelines for real-world applications

Koch, M.

Samenvatting

Machine Learning wordt steeds meer een belangrijke technologie voor de industrie. Het biedt de mogelijkheid om te leren van voorgaande ervaringen op een geautomatiseerde wijze om zo beslissingen te maken die gebaseerd zijn op aangeleerde ervaring. Dit principe kan worden gebruikt op processen te optimaliseren en / of te automatiseren voor industriële doeleinden. Een dergelijk geschikt proces binnen de auto-industrie is bijvoorbeeld de kwaliteitsbepaling bij de lopende band. In plaats van deze kwaliteitsbepaling te verrichten in een tijdrovend, handmatig proces, kunnen beeldherkenningsmethoden gebaseerd op machine learning toegepast worden om storingen te detecteren.

Naast het verbeteren van de huidige situatie kan machine learning ook bijdragen aan de ontwikkeling van volledige nieuwe producten, zoals autonoom rijden of het aanbieden van diensten die volledig aangestuurd worden door data. De ontwikkeling van deze nieuwe data-gedreven producten is veelal een lange procedure en zelfs de toepassing van machine learning algoritmes voor specifieke problemen is vaak niet eenvoudig. Om dit te illustreren introduceren wij een data-gedreven dienst vanuit de auto-industrie genaamd Automated Damage Assessment. Gebaseerd op reeds vergaarde ervaringen van deze data-gedreven ontwikkeling van diensten, beschrijven wij hier een methodologie om data-gedreven diensten nauwkeurig en snel te ontwikkelen.

Met name data-gedreven diensten binnen het domein van de auto-industrie kunnen gebaseerd worden op sensor data, d.w.z. data welke wordt geregistreerd door sensoren binnenin de auto door de tijd heen, in zogenaamde tijdreeksen. In veel gevallen kunnen tijdreeksen van meer dan één sensor worden gebruikt, resulterende in zogenaamde multivariate tijdreeksen. De bestaande methoden om classificatie-problemen met multivariate tijdreeksen te verhelpen zijn vaak complex en ontwikkeld voor een specifiek probleem, zonder schaalbaar te zijn om

verschillende problemen op te lossen. Om dit te verhelpen demonstreren wij in dit proefschrift benaderingen van verschillende complexiteit, toepast op meerdere algemeen-verkrijgbare datasets, medische- en industriéle datasets. Als startpunt hebben wij een algemene pijplijn ontwikkeld, gebaseerd op tijdreeks-features en toepast en verbeterd binnen verschillende bestaande datasets uit de auto-industrie en medische domeinen.

Recent zijn een aantal AutoML methoden beschreven, welke er naar streven om geoptimaliseerde modellen op een automatische wijze te ontwikkelen. Wij hebben deze veelbelovende AutoML methoden versterkt met een andere methode om multivariate tijdreeks-problemen te verhelpen, met als resultaat dat enkele van deze technieken geschikt blijken te zijn voor dit doeleinde. We hebben al deze technieken toegepast op algemeen-verkrijgbare datasets, waaruit blijkt dat met name twee AutoML methodes, namelijk GAMA en ATM, evenals onze eigen PHCP methode het beste geschikt zijn om deze multivariate tijdreeks-problemen te verhelpen.

Om rekening te houden met de interactie van features uit verschillende tijdreeksen is een methode ontwikkeld, gebaseerd op de evolutionaire algoritmen techniek genaamd genetisch programmeren. Het gebruikt het principe van biologische evolutie om combinaties van features te berekenen. Gebaseerd op onze resultaten kunnen wij stellen dat de aanname dat de combinatie van verschillende features betere prestaties levert dan bij het gebruik van de pure features, waar is.