Trustworthy anomaly detection for smart manufacturing

Li, Z.

Samenvatting

Deze dissertatie richt zich op de ontwikkeling van betrouwbare anomaliedetectiemethoden ter verbetering van productieprocessen, met name in high-tech systemen. Hoewel de primaire focus ligt op slimme productie, hebben veel van de technieken en bevindingen een brede toepasbaarheid buiten dit domein, waarbij verschillende methoden zijn ontworpen voor algemeen gebruik. Gezien de complexiteit van betrouwbare anomaliedetectie, benadert dit werk de uitdagingen vanuit twee complementaire perspectieven: data-centrische AI en model-centrische AI, elk leidend tot een onderzoeksvraag die uit meerdere subvragen bestaat.

Vanuit het *data-centrische AI perspectief* behandelt deze dissertatie de uitdagingen met betrekking tot complexe en hoog-dimensionale gegevens, die beide veel voorkomen in slimme productiesystemen. De eerste onderzoeksvraag (Q1.1) betreft het detecteren en verklaren van anomalieën in systeemlogboeken. Om dit op te lossen, hebben we *Logs2Graphs* ontwikkeld, een nieuwe methode die gebeurtenislogboeken eerst omzet in informatieve grafen. Deze grafen worden vervolgens geanalyseerd met een gespecialiseerd graaf neuraal netwerkmodel, *OCDiGCN*, om anomalieën te identificeren. *Logs2Graphs* verbetert niet alleen de detectie nauwkeurigheid, maar biedt ook duidelijke verklaringen door de meest relevante knooppunten in de anomalie-graaf te identificeren, wat de analyse van grondoorzaken mogelijk zou kunnen maken.

De tweede onderzoeksvraag (Q1.2) richt zich op het omgaan met hoog-dimensionale gegevens, die traditionele log-anomaliedetectiemethoden kunnen ondermijnen. We hebben een nieuwe aanpak ontwikkeld om de complexiteit van loggegevens te verminderen door relevante loggebeurtenissen die sterk samenhangen met systeemfouten te identificeren. Deze methode verbetert de foutdetectie en voorspellingsnauwkeurigheid, zoals aangetoond in experimenten met echte datasets. Door relevante systeemgebeurtenissen te selecteren, wordt het model effectiever in het identificeren van geleidelijke foutpatronen die anders verborgen zouden blijven.

Vanuit het *model-centrische AI perspectief* richt deze dissertatie zich op het verbeteren van uitlegbaarheid, robuustheid, generaliseerbaarheid en automatisering van anomaliedetectiemodellen. De eerste modelgerichte onderzoeksvraag (Q2.1) onderzoekt hoe intrinsieke uitlegbaarheid van anomaliedetectiesystemen kan worden bereikt. We hebben *QCAD* geïntroduceerd, een contextuele anomaliedetectiemethode gebaseerd op *Quantile Regression Forests*, die afhankelijkheden tussen kenmerken verkent om anomalieën te identificeren en te verklaren. Deze methode verhoogt zowel de detectie nauwkeurigheid als de interpretatie, waardoor domeinexperts beter kunnen begrijpen waarom bepaalde objecten als afwijkend worden beschouwd.

Een gerelateerde onderzoeksvraag (Q2.2) onderzoekt de robuustheid van post-hoc verklaringen voor graaf neurale netwerken (GNNs) onder adversariële aanvallen. We ontdekten dat huidige post-hoc GNN-verklaringstechnieken zeer kwetsbaar zijn voor kleine verstoringen in graafstructuren, die de verklaringen drastisch kunnen veranderen zonder de modelvoorspellingen aan te passen. Om dit aan te pakken, hebben we *GXAttack* ontwikkeld, een van de eerste optimalisatie-gebaseerde aanvallen die zich specifiek richt op GNN-verklaringen. Dit onderzoek onthult de kwetsbaarheid van veelgebruikte GNN-explainers en benadrukt de noodzaak voor robuustere interpretatiemethoden in kritieke toepassingen.

De derde modelgerichte onderzoeksvraag (Q2.3) betreft de generaliseerbaarheid van anomaliedetectiemodellen naar nieuwe, onbekende omgevingen. We hebben *ARMET* ontwikkeld, een ongesuperviseerde domeinaanpassingsmethode die gelabelde normale grafieken uit een brondomein gebruikt om anomalieën te detecteren in een niet-gelabeld doeldomein. Door middel van adversarial learning leert ARMET geschikte graafrepresentaties en behaalt het superieure prestaties in cross-domein anomaliedetectie, waardoor het bijzonder nuttig is voor evoluerende systemen zoals software-updates of hardwarewijzigingen.

De laatste onderzoeksvraag (Q2.4) richt zich op het automatiseren van de hyperparameterafstemming in ongesuperviseerde anomaliedetectiesystemen, waar gelabelde gegevens schaars zijn. Veel aanpakken gebaseerd op zelf-gesuperviseerd leren (SSL) voor anomaliedetectie zijn gevoelig voor hyperparameterinstellingen, wat leidt tot overschatte prestaties wanneer labels onjuist worden gebruikt voor afstemming. Om dit te verhelpen, hebben we *AutoGAD* geïntroduceerd, de eerste geautomatiseerde hyperparameterselectiemethode voor SSL-gebaseerde anomaliedetectie in grafen. AutoGAD maakt gebruik van een interne evaluatiemetriek om hyperparameters te selecteren zonder afhankelijk te zijn van labels, waardoor labellekkage wordt verminderd en de modelbetrouwbaarheid wordt verbeterd.

Samengevat presenteert deze dissertatie een uitgebreid raamwerk voor het verbeteren van anomaliedetectie in slimme productie door data-centrische en model-centrische AI-benaderingen te integreren. We laten zien dat anomaliedetectie in slimme productie kan worden verbeterd door graaf neurale netwerken te gebruiken om complexe loggegevens te verwerken en selectie van variabelen toe te passen om hoge dimensionaliteit aan te kunnen. Bovendien helpen uitlegbare modellen zoals *QCAD* om anomaliedetectie beter interpreteerbaar te maken, terwijl *GXAttack* de robuustheid van post-hoc GNN-verklaringen onderzoekt en *ARMET* de aanpasbaarheid over verschillende domeinen mogelijk maakt. Tot slot ondersteunt geautomatiseerde hyperparameterafstemming via *AutoGAD* de ontwikkeling van betrouwbare anomaliedetectiesystemen zonder afhankelijk te zijn van labels. Deze bijdragen verbeteren niet alleen de betrouwbaarheid en efficiëntie van productieprocessen, maar bieden ook inzichten die toepasbaar zijn in een breed scala aan domeinen die afhankelijk zijn van anomaliedetectie.