Exploring graph-based clustering and outlier detection algorithms

Li, J.

Samenvatting

In deze tijd van big-data, blijft het verkrijgen van betekenisvolle inzichten uit complexe data en datasets die onderlinge verbanden hebben, een van de belangrijke uitdagingen in verschillende toepassingsgebieden. In dit proefschrift hebben we methoden bestudeerd voor clustering en detectie van uitschieters (outliers) ter ondersteuning van data-analyse. Traditionele methoden voor clustering en outlier-detectie, zoals k-means, DBSCAN en statistische anomalie scoring, zijn sterk afhankelijk van Euclidische afstand, aannames of dichtheidsverdeling of globale grenswaardes. Daarbij hebben deze methoden moeite met het vastleggen van ingewikkelde relaties en contextuele anomalieën zoals die voorkomen in hoog-dimensionale, ruis-rijke, of dynamisch-voortschrijdende data. Graaf-gebaseerde benaderingen zoals methoden gebaseerd op de Minimum Spanning Tree (MST) en spectrale modellen, bieden een krachtig alternatief door het benutten van structurele eigenschappen en het blootleggen van verborgen patronen en afwijkingen. Ons onderzoek laat de superioriteit van de graaf-gebaseerde methoden zien in vergelijking met andere methoden. Daarbij wordt de aanpasbaarheid, de efficiëntie en de toepasbaarheid in de dagelijkse praktijk benadrukt, zonder dat deze methoden berusten op Deep Learning of Neurale netwerken. MST-gebaseerde clustering excelleert in het detecteren van clusters die onregelmatig van vorm zijn en ook voor het detecteren van hiërarchische structuren; beiden worden niet gezien door traditionele methoden. Door het construeren van grafen waar de knopen de data punten representeren en de overeenkomsten worden gecodeerd door de segmenten, i.e. de lijnen die de knopen verbinden zoals om kNN of Gaussian Kernels, kunnen algoritmen als spectraal clustering de knopen partitioneren door de Laplacian graaf te analyseren. De Laplacian graaf is een matrix waarin de patronen van globale connectiviteit worden vastgelegd. Anders dan

k-means clustering, waarbij sferische clusters worden verondersteld, is er bij spectraal clustering sprake van aanpassing aan de complexe geometrieën, hierdoor is het zeer geschikt voor toepassingen zoals segmentatie in beelden waar we veelal niet-lineaire grenzen zien. Deze methoden overtreffen dichtheid-gebaseerde methoden zoals DBSCAN waar het schaarse of onevenwichtig gedistribueerde data betreft; in dergelijke verdelingen van de data worden clusters gefragmenteerd of onjuist samengevoegd. Outlier-detectie heeft op een soortgelijke manier baat bij graaf structuren. In traditionele statistiek, i.e. Z-score, IRQ, krijgen anomalieën een label dat is gebaseerd op de globale afwijking van het gemiddelde waarbij geen rekening wordt gehouden met contextuele outliers die acceptabel zijn in een lokale omgeving maar afwijkend in een globale context. De Spectral-Analysis-Outlier-Detection (SAOD) methode introduceert een nieuwe, graaf-gebaseerde aanpak voor het vinden van outliers in hoog-dimensionale ruimtes. Er worden k Nearest Neighbor grafen geconstrueerd waarvan de eigenwaarde distributies van de Laplacian matrices worden geanalyseerd, hieruit kan SAOD op een effectieve manier de inliers van de outliers onderscheiden op basis van hun onderscheidende spectrale patronen. Deze methode maakt gebruik van een dichtheid schatting multivariate uit een Gaussische kernel om de mate van outlierness te kwantifieren. Experimenten met bestaande datasets laten zien dat SAOD is zee waardevol voor toepassingen in de medische diagnostiek en cybersecurity waar hoog-dimensionale data overheersend zijn. Veel outlier-detectie methoden zijn niet in staat om datasets met een variabele dichtheid of onregelmatige vorm te verwerken. Er is dan voorkennis vereist over de verhoudingen met betrekking tot de outliers waardoor de praktische toepasbaarheid in medische toepassingen beperkt is. Het adaptieve Mini-MST Outlier-detection (MMOD) is een algoritme dat rekening houdt met de dichtheid-afhankelijkheid de beperkend vormt bij andere algoritmen. Daarbij kunnen outliers effectief worden gevonden zonder voorkennis over de verdeling van de outliers. De nieuwe afstandsmetriek van het algoritme schaalt dynamisch in overeenkomst met lokale dichtheid waardoor juist voor complexe datasets met een variabele clustervorm en -dichtheid de outliers met nauwkeurigheid kunnen worden gevonden. Een validatie op elektronische-patiënt-data laat zien dat het MMOD algoritme een superieure performance heeft. Methoden voor outlier-detectie gebaseerd op clustering zijn gevoelig voor ruis en willekeurige selectie van het medoide punt in het cluster hetgeen resulteert in fout-

positieven in complexe datasets. In een algoritme met een geschaalde MST met Medoide selectie (MS2OD) wordt een combinatie gemaakt van de robuustheid van de MST en een innovatief ruis-bewust medoide selectie proces. In deze methode worden de data op iteratieve wijze gepartitioneerd naar aanleiding van langste-segment snede in een geschaalde MST, terwijl het geavanceerde medoide selectie algoritme rekening houdt met vervuiling door ruis in de cluster centers. Naar aanleiding van het onderzoek dat in dit proefschrift wordt gepresenteerd kunnen we vaststellen dat graaf-gebaseerde clustering en outlier-detectie onontbeerlijke gereedschappen zijn in moderne data analyse, die een ongeëvenaarde flexibiliteit bieden in het verwerken van complexe en onderling verbonden dataset. Voortgang in theoretische onderbouwing, het optimaliseren van computationele efficiëntie en de impact op "echte" data, maken dat dit werk een brug bouwt tussen academische innovatie en praktische inzetbaarheid. In toekomstige verder ontwikkelingen zal worden gekeken naar uitbreiden van deze methoden voor streaming-data waardoor real-time analyse kan worden uitgevoerd, en daarnaast het ontwikkelen van interpreteerbare raamwerken om het vertrouwen in graaf-gebaseerde beslissingen te vergroten waarmee hun belang wordt gezekerd in deze data-gedreven wereld.

Samenvatting