ONDER PROFESSOREN

AIDA: Analytic Isolation and Distance-based Anomaly Detection

Kees Oosterlee

Prof. dr. ir. C.W. Oosterlee is hoogleraar aan de Universiteit Utrecht.

Machine learning en kunstmatige intelligentie worden meer en meer gebruikt voor verschillende toepassingen binnen de financiële en ook de actuariële wereld. Denk aan het construeren van optimale beleggings-strategieën, econometrische voorspellingen vanuit historische tijdreeksen, maar ook als computationeel gereedschap om uitdagende, dure berekeningen te versnellen. In deze context werkt mijn onderzoeksgroep Computational Finance in het Mathematische Instituut van de Universiteit Utrecht aan snelle, robuuste en betrouwbare rekenmethoden voor verschillende financiële toepassingen. Naast optimale portefeuilles en financieel risicomanagement (zoals counterparty kredietrisico, en total valuation adjustment, xVA), werken wij ook aan machine learning algoritmen voor fraudedetectie en anti-witwaspraktijken. Anomaliedetectie (AD) is een fundamenteel onderzoeksgebied in machine learning, vanwege de relevantie ervan voor veel real-life toepassingen, van netwerkinbraakdetectie tot fraudeanalyse. Dit onderwerp bespreken we hier in meer detail.

INTRODUCTIE

In dit stuk beschrijven we onderzoek naar data-gedreven anti-witwas-onderzoek (anti-money laundering, AML) en fraudedetectie, om criminele activiteiten te bestrijden. Anomaly-detectie is het vakgebied waar afwijkende datapunten (zogenaamde outliers) in hoog-dimensionale sets opgespoord dienen te worden. De dataset representeert bijvoorbeeld het aantal cliënten in een klantenbestand en relevante features zijn de aspecten van klanten en transacties die belangrijk kunnen zijn voor de detectie. Deze features bepalen de dimensionaliteit van de dataset, en met 25 tot 50 features spreken we van een hoogdimensionale dataset.

Iedereen heeft wel een (visuele) notie van outliers in datasets. Hawkins [5] geeft een intuïtieve definitie van een outlier als ‘een waarneming die zo sterk afwijkt van andere waarnemingen dat het vermoeden bestaat dat ze door een ander mechanisme is gegenereerd’. Outliers zijn al eeuwenlang een statistisch relevant onderwerp, maar sinds de datawetenschap ongeveer tien jaar geleden begon te bloeien, zijn outliers ook een hot topic binnen datamining. De belangrijkste focus is het vinden van schaalbare methoden voor het detecteren van outliers, omdat men met enorm grote datasets te maken heeft met vele features. Hierdoor worden niet-schaalbare methoden te duur in rekentijd en dataopslag. De standaardpraktijk om outliers op te sporen is een op regels van de regelgever gebaseerde detectie, met andere woorden, als voorgeschreven regels overtreden worden moet er gecheckt worden. Echter, regels kunnen omzeild worden en we focusseren hier daarom op anomaliedetectie zonder voorgeschreven regels (patroonherkenning), ook wel unsupervised machine learning genaamd. In het bijzonder presenteren we het parametervrije Analytic Isolation and Distance-based Anomaly (AIDA)-detectiealgoritme [3] door Luis Souto Arias, Pasquale Cirillo en Kees Oosterlee. Op basis van AIDA is ook het Tempered Isolation-based eXplanation (TIX)-algoritme gedefinieerd, dat de meest relevante kenmerken voor de anomalie identificeert, waardoor de verklaarbaarheid van het detectiemechanisme wordt verbeterd.

INTRODUCTIE WITWASSEN

De afgelopen jaren zijn er meerdere incidenten gemeld waarbij het toezicht op witwassen tekortschoot, en die tot hoge boetes voor banken leidden. De Financial Action Task Force (FATF) is de wereldwijde waakhond tegen witwassen en terrorismefinanciering. Zij controleert landen op het naleven van afspraken en doet aanbevelingen. Het United Nations Office on Drugs and Crime (UNODC) schatte de omvang van de witwasactiviteiten in 2009 op 2,7% van het mondiale BNP, ofwel ongeveer 1,6 biljoen dollar, die via het financiële systeem werd witgewassen. Dat bedrag is steeds verder gegroeid. De meeste landen hebben AML-instanties met als doel het bewustzijn binnen de publieke en private sector te vergroten en regelgevende instrumenten te bieden om de problemen te bestrijden. De financiële sector speelt een cruciale rol in de strijd tegen het witwassen en moet systemen opzetten voor het melden van financiële transacties en het screenen en monitoren van klanten. Een onderdeel van het AML-proces is de voortdurende due diligence en het monitoren van zakelijke relaties, waarbij ongebruike-lijke en mogelijk verdachte transacties gemonitord moeten worden, en elk daadwerkelijk vermoeden zal moeten leiden tot een meldingsplicht.

Het monitoren van klanten gebeurt vaak met behulp van een op regels gebaseerd systeem waarbij klantactiviteit wordt vergeleken met vaste grenswaarden die bepalen of de activiteit ongebruikelijk is of niet. Deze grenswaarden worden bepaald op basis van historische activiteit en zijn vaak beperkt tot bepaalde groepen klanten. De op regels gebaseerde systemen zijn goed interpreteerbaar en werken goed bij het onderscheppen van de meest voorkomende witwasscenario’s. Er zijn echter nadelen aan een dergelijk systeem, zoals de mogelijkheid dat criminelen de regels en grenswaarden testen en leren hoe ze deze kunnen omzeilen en dat de regels op slechts een kleine subset van alle beschikbare gegevens gebaseerd zijn.

Als machine learning algoritmen worden gebruikt om de regels te checken, wordt gewoonlijk voor supervised learning gekozen. In het geval van supervised learning bevatten de gegevens waarmee het algoritme getraind wordt, de gewenste output (met andere woorden: de regels en grenswaarden zijn de outputlabels). Een doelfunctie moet geminimaliseerd worden, waarbij het algoritme de voorspeller is, die is getraind om de juiste output te voorspellen op basis van de invoergegevens. Supervised learning kan worden onderverdeeld in classificatie en regressie. Classificatiemodellen worden gebruikt als de gewenste output een klasse is, zoals het voorspellen of een transactie frauduleus is (1) of niet (0), terwijl regressiemodellen worden gebruikt om numerieke waarden te voorspellen, zoals de prijs van huizen (gegeven de locatie, marktomstandigheden, enz.).

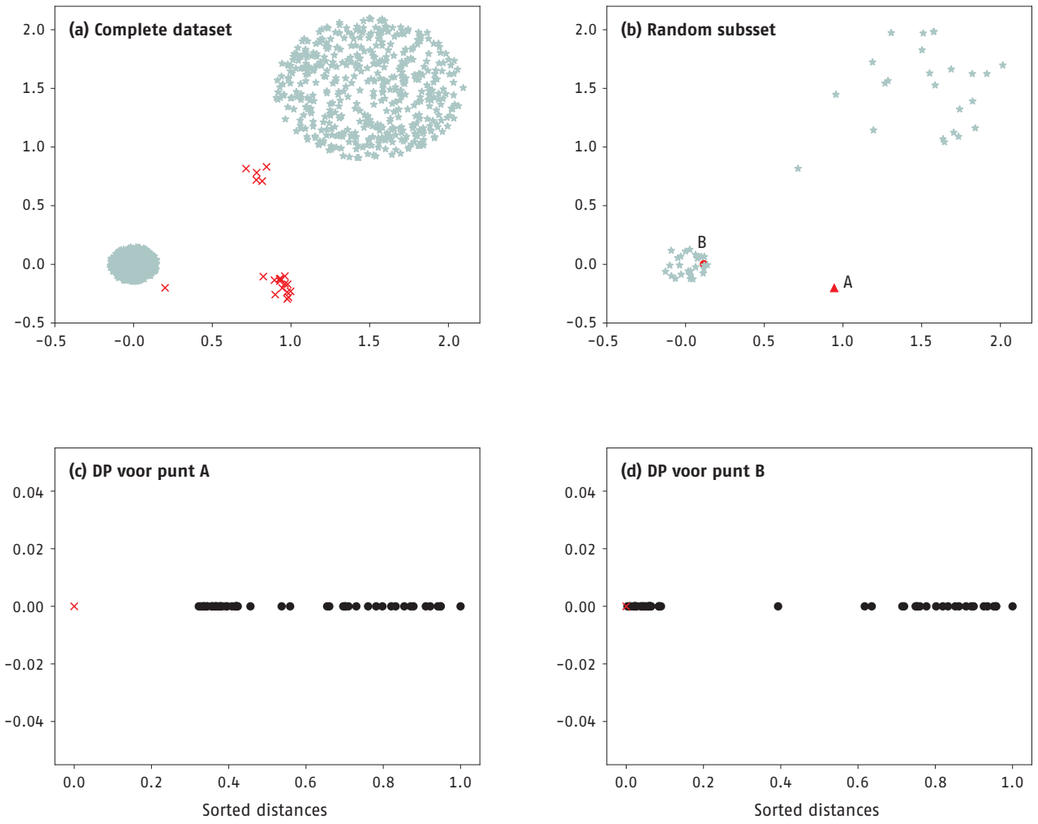
UNSUPERVISED LEARNING

In het geval van unsupervised learning wordt de training van het algoritme uitgevoerd zonder specifieke informatie over de gewenste resultaten. Unsupervised AD-technieken hebben de voorkeur in scenario’s waarin gelabelde outliers niet aanwezig zijn. Van de unsupervised learning methoden zijn methoden die gebaseerd zijn op afstand tussen datapunten of dichtheid van dataclusters populair vanwege hun interpreteerbaarheid en prestaties. Afwijkingen zijn punten die ver verwijderd zijn van ‘normale’ punten [7], of die zich bevinden in gebieden met een geringe lokale dichtheid [4]. Populaire methoden in deze klasse zijn Local Outlier Factor (LOF) [4] en k-Nearest Neighbours (kNN) [7], omdat er geen aannames zijn over de data. Het voorgeschreven aantal (k) buurpunten heeft echter grote invloed op de prestaties van dit algoritme, maar kan niet a priori worden bepaald omdat iedere dataset weer anders is. In hoge dimensies is verder het concept afstand niet meer eenduidig gedefinieerd (zie bijvoorbeeld [2])! Een ander nadeel is dat de afstanden tussen alle paren data-punten moeten worden berekend, wat resulteert in een kwadratische rekentijd en dus in een duur algoritme. Dit kan gedeeltelijk worden opgelost door subsets en steekproeven te gebruiken, zodat de rekentijd wordt teruggebracht.

Een tweede klasse unsupervised learning methoden zijn de op isolatie gebaseerde algoritmen, waarvan Isolation Forest (IF) [6] de meest bekende variant is. IF bouwt vele binaire bomen, waarbij iedere boom geconstrueerd wordt door de dataset herhaaldelijk in tweeën (langs een dimensie/feature) te splitsen, waarbij de gemiddelde lengtes van de paden in de bomen van het bos indicatoren zijn voor ‘outlierness’ van datapunten. Het inzicht is dat outliers sneller te isoleren zijn in zo’n boom dan normale datapunten en zich dus gemiddeld dichtbij de wortel van een boom bevinden, ofwel, korte padlengtes betekenen een grote kans op isolatie en vice versa. Omdat berekeningen van afstanden en dichtheden niet nodig zijn, is de rekencomplexiteit laag. Bovendien is IF eenvoudig schaalbaar naar grote datasets. Echter, voor hoog-dimensionale datasets werkt IF minder goed omdat het splitsen van de datatset op basis van een random keuze van de te splitsen dimensie verloopt en het kan gebeuren dat in hoge dimensies de outlier niet wordt gedetecteerd.

AIDA ALGORITME

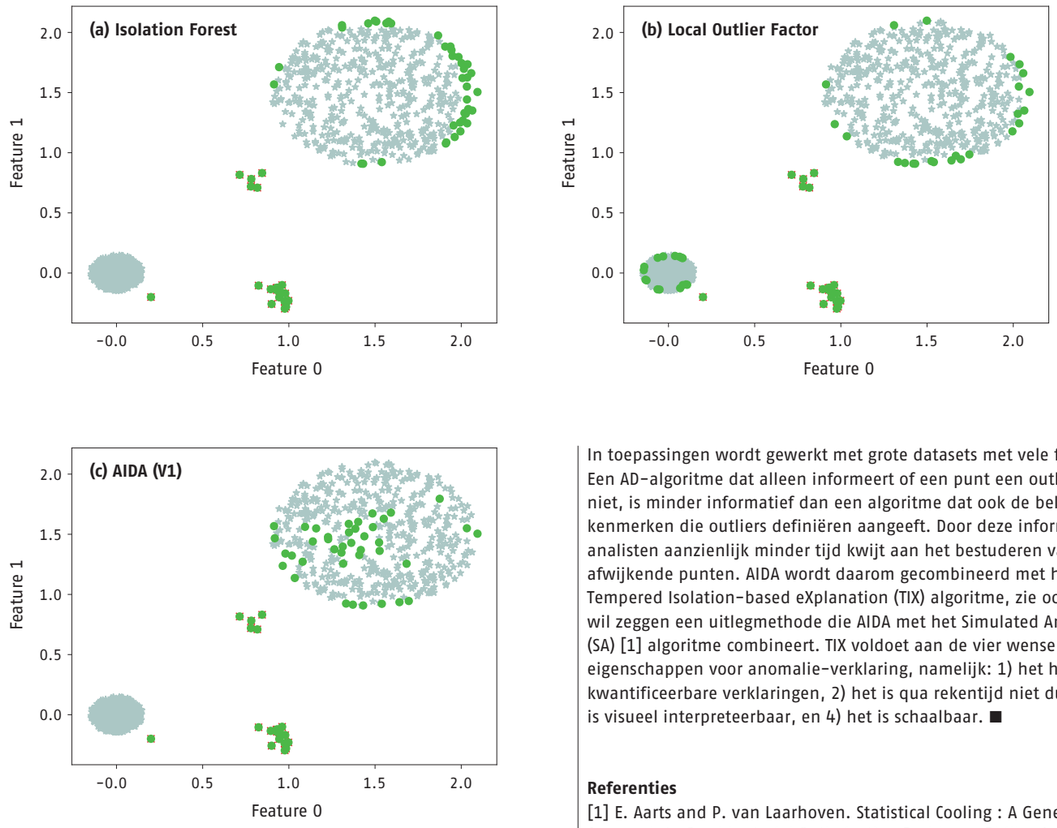
Als alternatief introduceren we een nieuw, op afstand gebaseerd, AD-algoritme: Analytic Isolation and Distance-based Anomaly (AIDA) [3]. AIDA baseert zich niet slechts op het concept van buurpunten om anomalieën/uitschieters te detecteren, maar combineert dit met isolatie. AIDA maakt gebruik van N random subsets van de complete dataset om de rekencomplexiteit te reduceren. Voor elk punt in de subset wordt de afstand tot ieder ander datapunt bepaald (de afstand is een één-dimensionale metriek). Het nulpunt komt overeenkomt met de afstand van het punt tot zichzelf. Nadat de afstanden zijn gesorteerd in een afstandsprofiel (DP, distance profile), passen we hierop het IF-algoritme toe totdat het linkerrandpunt geïsoleerd is. Deze stappen geven ons een outlierscore per subset, en de uiteindelijke score wordt verkregen door aggregatie van deze resultaten. Het idee van het AIDA-algoritme wordt geïllustreerd in figuur 1.



Figuur 1: Korte weergave AIDA algoritme

De plot linksboven toont de volledige dataset, die bestaat uit 1000 waarnemingen met twee features, terwijl de plot rechtsboven een willekeurige subset van 50 samples toont met twee referentiepunten gemarkeerd met een rode driehoek (A) en een rode cirkel (B). De onderste plots presenteren de DP's van punt A (links) en punt B (rechts), waar de linkerrandpunt is gemarkeerd met een rood kruis om te benadrukken dat dit het punt is dat we willen isoleren. Het is duidelijk dat het linkerrandpunt gemakkelijker te isoleren is in de DP van punt A dan in de DP van punt B, vandaar dat punt A een hogere outlierscore krijgt.

AIDA is in staat om andere soorten outliers te detecteren dan bijvoorbeeld IF [6] of LOF [4]. Dit is met name relevant voor ensemblemethoden, waarbij de scores van verschillende detectiemodellen worden gecombineerd om de robuustheid van de uiteindelijke schattingen te vergroten, zie figuur 2. Terwijl in de IF literatuur de isolatiescore wordt verkregen met simulatie, is voor de scorefunctie van AIDA een analytische uitdrukking bepaald, waardoor berekeningen worden vereenvoudigd.



Figuur 2: Vergelijking van de 60 meest afwijkende punten gedetecteerd door AIDA, IF en LOF. Inliers zijn gemarkeerd met grijze sterren, gedetecteerde outliers met groene cirkels en daadwerkelijke outliers met rode kruisjes.

In toepassingen wordt gewerkt met grote datasets met vele features. Een AD-algoritme dat alleen informeert of een punt een outlier is of niet, is minder informatief dan een algoritme dat ook de belangrijkste kenmerken die outliers definiëren aangeeft. Door deze informatie zijn analisten aanzienlijk minder tijd kwijt aan het bestuderen van mogelijk afwijkende punten. AIDA wordt daarom gecombineerd met het Tempered Isolation-based eXplanation (TIX) algoritme, zie ook [3], dat wil zeggen een uitlegmethode die AIDA met het Simulated Annealing (SA) [1] algoritme combineert. TIX voldoet aan de vier wenselijke eigenschappen voor anomalie-verklaring, namelijk: 1) het heeft kwantificeerbare verklaringen, 2) het is qua rekentijd niet duur, 3) het is visueel interpreteerbaar, en 4) het is schaalbaar. ■

Referenties

[1] E. Aarts and P. van Laarhoven. Statistical Cooling : A General Approach to Combinatorial Optimization Problems. Philips Journal of Research, 40(4): 193–226, 1985.

[2] C. C. Aggarwal, A. Hinneburg, and D. A. Keim. On the Surprising Behavior of Distance Metrics in High Dimensional Space. In J. Van de Bussche and V. Vianu, editors, Lecture Note in Computer Science, volume 1973. Springer, Berlin, Heidelberg, 2001. doi:10.1007/3-540-44503-X_27.

[3] L. A. S. Arias, C. W. Oosterlee, and P. Cirillo. Aida: Analytic isolation and distance-based anomaly detection algorithm. Pattern Recognition, page 109607, 2023.

[4] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander. LOF: Identifying density-based local outliers. SIGMOD Rec., 29(2):93–104, May 2000. doi:10.1145/335191.335388.

[5] D. M. Hawkins. Identification of outliers, volume 11. Springer, 1980.

[6] F. T. Liu, K. M. Ting, and Z.-H. Zhou. Isolation-Based Anomaly Detection. ACM Transactions on Knowledge Discovery from Data, 6(1):1–39, 2012. doi:10.1145/2133360.2133363.

[7] S. Ramaswamy, R. Rastogi, and K. Shim. Efficient Algorithms for Mining Outliers from Large Data Sets. SIGMOD Rec., 29(2):427–438, 2000. doi: 10.1145/335191.335437.