



**FRAUDEDTECTIE
IN DE FINANCIËLE WERELD:
ONGEKENDE
NIEUWE MOGELIJKHEDEN**

Data en dataverkeer zijn niet meer weg te denken uit de financiële wereld. De snelle ontwikkeling van informatietechnologie levert, behalve een enorme toename van dataverkeer en -opslag, grote mogelijkheden voor methoden die gebaseerd zijn op wiskunde en machine learning. Onder dit laatste verstaan we de wetenschap (en de toepassing) van adaptieve systemen waarbij patroonherkenning en reacties op inputs kernthema's zijn. We hebben technieken die in een split second online een indicatie kunnen geven van afwijkingen, high risk, onregelmatigheden, marktinefficiëntie, enzovoorts. Nog krachtiger technieken liggen binnen handbereik.

Zowel in de banksector als de verzekeringssector kunnen deze methoden bijdragen aan succes. Het opsporen van fraude bij verzekeringsdeclaraties, geldtransacties, creditcardtransacties, onroerendgoedtransacties, enzovoorts, wordt veel succesvoller dan nu. Hoewel de instellingen zelf daar zeer weinig over kwijt willen, is het succes van nieuwe technieken voor de oplettende lezer waarneembaar. Zo is het systeem dat creditcardtransacties monitort bij Interpay uitzonderlijk goed en draagt het aanwijsbaar bij aan de veiligheid, betrouwbaarheid en het imago van Interpay. En soms melden verzekeraars het succes van deze methoden zelf: in augustus 2003 meldde Allianz-dochter FFIC dat één onderdeel van hun systemen op dit terrein alleen al \$700.000 per jaar oplevert door fraudegericht, alert en met moderne technieken op te treden¹. Interessante recente rapporten over fraudedetectie in Nederland zijn 'Kwetsbaarheid van de zorgsector voor georganiseerde fraude' (in opdracht van het Ministerie van Justitie, november 2003) en 'Risicosturing bijstandsfraude: een inventarisatie van methodieken' (in opdracht van het Ministerie van Sociale Zaken, juli 2003). Beide rapporten geven een goede doorkijk naar omvang, aard en methoden om mogelijke fraude en onregelmatigheden te detecteren.

Dezelfde technieken komen we tegen bij het opsporen van witwaspraktijken, bij het bijtijds detecteren van retention, bij eventdriven marketing en bij marketing in het algemeen. Dit artikel poogt een tipje van de sluier op te lichten van de huidige state-of-the-art en een beeld te schetsen van de enorme mogelijkheden voor bedrijfsleven en overheid.

BERT KERSTEN

De rekenkracht van computers ontwikkelt zich exponentieel. Voor geheugenopslag geldt eenzelfde ontwikkeling waarbij bovendien de prijs keldert. In Figuur 1 (pag. 14) wordt de ontwikkeling voor rekenkracht weergegeven, met zowel een logaritmische schaal als een gewone schaal.

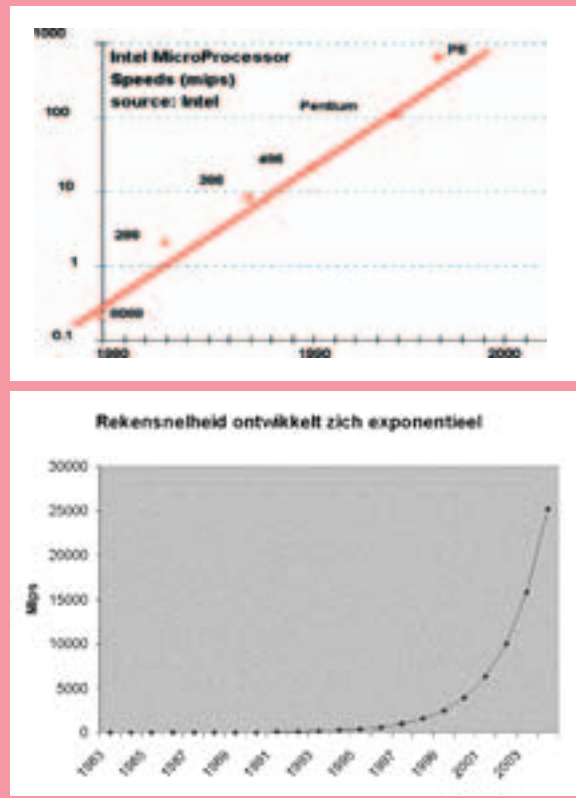
Deze grafieken laten zich eenvoudig illustreren met enkele voorbeelden. Indien het salaris van een medewerker zich op eenzelfde wijze zou ontwikkelen en hij zou in 1983 f 2500,- per maand verdienen, dan zou hij nu circa €28,5 miljoen per maand verdienen. Als dezelfde groei voor de snelheid van een auto zou gelden (stel voor het gemak in 1983: 100 km/h), dan zou de auto door de geluidsbarrière zijn gegaan in 1989 en wordt de snelheid van het licht bereikt in 2023. En tot slot, als het een opsporingsmedewerker in 1983 één dag zou kosten om onregelmatigheden te vinden in claimedrag bij verzekeringen of witwassen, dan zou het hem nu slechts één seconde kosten. Met andere woorden: hij zou nu 28.800 mogelijke cases per dag te verwerken krijgen.

De ontwikkeling van geheugencapaciteit gaat op vergelijkbare wijze: elke tien maanden verdubbelt de geheugencapaciteit. Maar ook de prijzen dalen gestaag. Op dit moment kan men over één Terabyte beschikken voor €1000,- en het ziet er niet naar uit dat de bodem in de prijzen is bereikt.

Beide verschijnselen openen de deur voor het gebruik van wiskundige technieken en methoden uit machine

learning voor de monitoring van grote datastreams en het opsporen van afwijkingen of vreemde patronen. Voor de fans van Star Trek en The Matrix komt dit niet als een verrassing, maar de eerste toepassingen laten verbluffende resultaten zien.

Figuur 1. De exponentiële ontwikkelsnelheid van rekenkracht van computers.



Veel toepassingen richten zich op profiling: het opstellen en gebruiken van profielen waaraan gebeurtenissen in het dataverkeer worden gematcht. Dit kunnen bijvoorbeeld klantprofielen zijn die worden gebruikt voor het signaleren van afwijkende aankooppatronen, voor het aanbieden van nieuwe producten en diensten, voor het focussen van direct mailing acties, enzovoorts. In het verlengde hiervan liggen profielen voor gedrag zoals claimgedrag bij verzekeringen, gebruik van creditcards, gebruik van mobiele telefoons, maar ook profielen over de wijze waarop gebruikers instrumenten gebruiken. Reeds begin jaren negentig werd op de Vrije Universiteit Amsterdam onderzocht in hoeverre men aan de wijze waarop toetsen op een toetsenbord werden aangeslagen, kon zien of het de legitieme gebruiker van die computer was. Onlangs zijn deze technieken ook gebruikt om te bepalen of de gebruiker van een mobiele telefoon de legitieme bezitter ervan is.

Andere profielen hebben betrekking op afbetaalgedrag, het gebruik van/toegang tot grote informatiebanken in de sociale zekerheid, componenten waaruit grote datastreams bestaan, enzovoorts. Veel van het werk is vertrouwelijk, maar overduidelijk is de wijze waarop wiskunde en machine learning deze activiteiten ondersteunt.

Ict-doorbraakprojecten 2003

Het Ministerie van Economische Zaken (Senter) wijst jaarlijks subsidiegelden toe aan projecten die zich kenmerken als 'doorbraken' en die van groot belang zijn voor zowel de maatschappij als de wetenschap. Op de eerste plaats van de lijst van ict-doorbraakprojecten 2003 staat het zogenaamde Diana-project. Het hoofddoel van het Diana-project is het ontwikkelen van nieuwe technologie voor adaptieve systemen die omvangrijke datastromen onderscheppen, ze in real-time analyseren en nuttige feedback geven. Deze technologie kan worden gebruikt om nieuwe generatiesystemen te bouwen voor fraudedetectie van elektronische betalingen, het detecteren van oneigenlijk gebruik van complexe en omvangrijke informatiesystemen, het detecteren van het binnendringen in netwerken, het verzamelen van business intelligence, enzovoorts. In feite wordt er voortgebouwd op werk zoals verricht door Kowalczyk (1997, 1998).

In dit project hebben zich vijf partners verzameld: de Vrije Universiteit Amsterdam (de Computational Intelligence Group), Moniforce, BKWI (Ministerie van Sociale Zaken), Interpay en Robeco. Moniforce – een jong en snel groeiend bedrijf uit Almere – zal de technologie leveren die de data moet onderscheppen uit de grote datastromen. De Vrije Universiteit zal de technieken en methodologie inbrengen voor datamining, voorspellende modellering en profiling. De drie andere partners zullen verschillende real-life cases inbrengen waarop de ontwikkelde technologie zal worden getest. Het is duidelijk dat dit een spannend en fascinerend project is dat door velen nauwkeurig zal worden gevolgd. Het project zal vier jaar duren.

Combineren van technieken en methoden

Er zijn methoden die al lang bekend staan om hun nuttige bijdragen bij profiling en – wat daar vlakbij ligt – scoring. Ook deze methoden worden door de grotere kracht van computers en de beschikbaarheid van data steeds vaker toegepast, met een toenemende geavanceerdheid. Zij

komen uit het statistische domein en – enkele – uit het domein van operation research. Voorbeelden zijn regressiemodellen (enkelvoudige, multi-pele, lineaire, logistische), principale componenten analyse, factor-analyse, multidimensionele schaling, discriminant analyse (univariaat en multivariaat), CHAID en CART. De lezer die hier meer over wil weten, zij verwezen naar standaard-literatuur over statistische data-analyse.

Belangrijk nieuw fenomeen is dat de kracht van deze technieken sterk kan worden verhoogd door ze te combineren met methoden uit de machine learning. Zonder hier dieper op de specifieke kenmerken van elk in te gaan, willen we er toch een aantal expliciet noemen: rough data models, naïeve baysiaanse methoden, beslisbomen, neurale netwerken, baysiaanse netwerken, support vector machines, boosting, bagging en stacking. De Engelse terminologie verwijst al naar het ontstaansdomein van deze technieken, namelijk het terrein van machine learning, datamining en kunstmatige intelligentie. De verwachting is dat binnenkort vanuit het terrein van evolutionaire computing (zie o.a. Eiben en Smith, 2003) nieuwe bijdragen aan het veld van profiling zullen worden geleverd. Bij evolutionaire computing gebruikt men populaties en struggle-for-life om optimale combinaties en kenmerken te vinden. Succesvolle toepassingen vindt men nu bij human resource planning, roostervraagstukken, volgordeproblemen en routeproblemen.

De ervaring leert dat het kiezen voor één of twee technieken niet tot het meeste succes leidt: het meest succesvol is wanneer men verschillende technieken tegelijkertijd hun bijdrage laat leveren. Dit laat zich als volgt illustreren:

In de dagelijkse praktijk waar zich deze detectievraagstukken afspelen, beschikt men doorgaans over veel expertkennis. Fraudecoördinatoren, fraudedeskundigen en opsporingspersoneel hebben in de loop der jaren een enorme kennis opgebouwd over afwijkende patronen en gedrag. Deze menselijke kennis moet worden benut én behouden. De benutting ervan vindt plaats middels het operationaliseren van expertregels in een zogenaamde rule-engine. De menselijke kennis wordt behouden door de aanwezigheid van de menselijke factor bij het beoordelen van verdachte cases en patronen (output van het fraudedetectiemodel). We zullen hier later op ingaan.

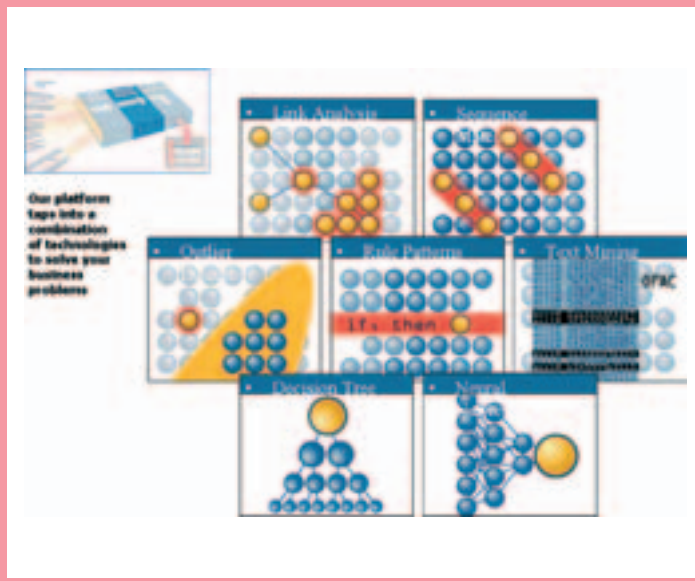
De rule-engine bevat de regels, principes, ervaringen en vermoedens van de menselijke experts. Dit kunnen harde

regels zijn (in de trant van ‘bij drie overboekingen via internet binnen dertig minuten, blokkeer de vierde’) maar ook zachte regels die ertoe leiden dat bepaalde cases niet worden geblokkeerd maar wel in de schijnwerpers komen te staan. De rule-engine wordt ook gevuld met regels die uit andere componenten van het fraudedetectiemodel afkomstig zijn, zoals uit statistische technieken. Kenmerkend voor rule-engine is dat deze flexibel moet zijn, snel aan te passen is en bovenal transparant is voor de gebruiker. In Figuur 2 wordt een schematisch overzicht gegeven van de interactie en combinatie van de verschillende technieken en inputs. Naast de rule-engine zijn de modules van statistische technieken, de modules van neurale netwerken, machine learning en patroon-

Figuur 2. Hybride modellen voor fraudedetectie.



Figuur 3. Tweecomponentenmodel waarbij KP het klantprofiel en TP het transactieprofiel is.



herkenning zichtbaar. Het is interessant te zien hoe bij het opsporen van witwaspraktijken dezelfde technieken worden gebruikt. In Figuur 3 (pag. 15) wordt een overzicht gegeven van de 'binnenkant' van detectiesoftware van Mantas. Mantas is één van de toonaangevende internationale bedrijven voor software voor het detecteren van witwaspraktijken. Belangrijk is de grote overeenkomst in gebruikte methoden. De toekomst van goede detectiesoftware ligt in het uitbreiden van het assortiment van methoden en technieken aan de binnenzijde, met moderne en krachtige algoritmen die de grote stroom data voortdurend en alert monitoren. Veel van deze technieken komen uit machine learning en uit een combinatie van verschillende wetenschappelijke vakgebieden.

Een uitstapje naar componentenmodellen

Bij de bespreking van modellen voor fraudedetectie is het nuttig een onderscheid te maken naar de complexiteit ervan. We kenmerken de modellen naar het aantal hoofdactoren die een rol spelen. Merk op dat het erbij deze indeling niet toe doet welke technieken concreet gebruikt worden om afwijkingen te signaleren.

Het meest eenvoudige model is het zogenoemde tweecomponentenmodel waarbij twee hoofdactoren met elkaar in verband worden gebracht. Figuur 4 geeft hiervan een

Figuur 4. Tweecomponentenmodel waarbij KP het klantprofiel en TP het transactieprofiel is.



voorbeeld voor transactieverkeer, in het bijzonder betalingsverkeer.

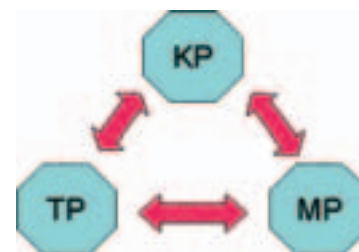
In dit model bestaat het detectiemodel uit de confrontatie van twee profielen. Elk van de profielen is opgebouwd uit analyses van relevante data van klant respectievelijk transactie. De confrontatie van deze twee profielen met elkaar leidt tot een score die de waarschijnlijk-/onwaar-

schijnlijkheid van de combinatie weergeeft. Er kan een kritische drempel worden overschreden waardoor de transactie wordt tegengehouden.

Dit model kan goed worden uitgebreid met gegevens die de tegenpartij van de transactie betreffen. We krijgen dan te maken met een driecomponentenmodel.

Dit model kan goede diensten verlenen bij bijvoorbeeld creditcardtransacties, internettransacties en het opsporen van fraude bij verzekeringen. In dat laatste geval betreft het

Figuur 5. Driecomponentenmodel waarbij KP het klantprofiel, TP het transactieprofiel en MP het merchantprofiel is.

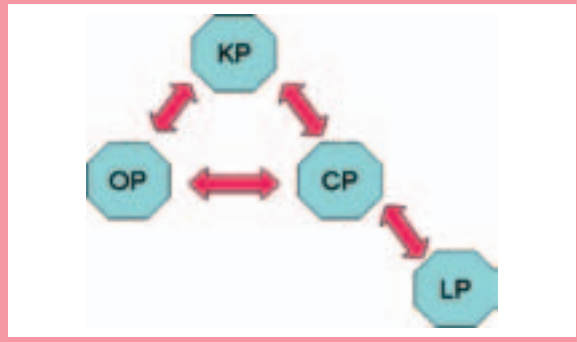


transactieprofiel (TP) de declaratie en is MP de partij die het geneesmiddel of de behandeling geeft. Het zal duidelijk zijn dat – mits de drie profielen goed getuned zijn – de detectiekwaliteit beter is dan bij een tweecomponentenmodel. De score in dit driecomponentenmodel is opgebouwd uit bijdragen van elk van de drie actoren en daarmee wordt een bredere range van fraudebronnen bestreken. Met dit in het achterhoofd kan men soms beter de berichten in de media begrijpen wanneer transacties worden onderschept of kaarten worden geblokkeerd².

De noodzaak om over juiste en actuele profielen te beschikken is natuurlijk voor alle componentenmodellen aanwezig. In de praktijk zal het maken van meer profielen meer werk kosten, maar men verkrijgt een groter onderscheidingsvermogen.

Viercomponentenmodellen gaan nog een stap verder. In Figuur 6 (pag. 17) wordt daar een voorbeeld van gegeven: vier profielen worden gematcht: klantprofiel, objectprofiel, claimprofiel en een leveranciersprofiel. Toepassingen vindt men bij autoschade, zorg, onroerend goed, et cetera. Bij autoschade bevat het objectprofiel gegevens en (vuist)regels over de auto, het claimprofiel betreft de ingediende declaratie en de regels en benchmarks voor reparatie, en het leveranciersprofiel bevat kenmerken van

Figuur 6. Viercomponentenmodel waarbij KP het klantprofiel, OP het objectprofiel, CP het declaratieprofiel en LP het leveranciersprofiel is.



het bedrijf dat de reparatie verricht. Een generalisatie naar andere toepassingsgebieden is eenvoudig te maken.

In de praktijk blijkt men met drie- en viercomponentenmodellen te kunnen volstaan. Dit is natuurlijk sterk afhankelijk van de aard van de fraude en onregelmatigheden, de frequentie en dynamiek ervan, de beschikbaarheid van data en last-but-not-least, de urgentie waarmee men deze detectie goed wil uitvoeren.

De juistheid van een detectie heeft twee aspecten. Net als iedere uitgevoerde voorspelling, kan men twee soorten fouten onderscheiden: een fout van de eerste soort en een fout van de tweede soort. Onder een fout van de eerste soort verstaat men het verschijnsel dat – hoewel het niet om een frauduleuze transactie gaat – het model dit toch aangeeft, met andere woorden iets wordt ten onrechte als frauduleus aangemerkt. De gevolgen van een dergelijke fout zijn totaal anders dan bij een fout van de tweede soort: een frauduleuze transactie wordt door het model aangemerkt als een goede transactie. Men noemt dit verschijnsel kortweg het probleem van de asymmetrische kosten.

De ernst van dit probleem is sterk context-gebonden: soms is het ernstig bijvoorbeeld bij het ten onrechte mailen of benaderen van een persoon voor een aankoop, en soms draagt het zelfs bij aan het imago van een veilig product wanneer bij bonafide gebruikers gecontroleerd wordt of zij inderdaad de gebruikers in kwestie zijn (bijvoorbeeld bij creditcards).

Profielen

De profielen zijn gebouwd op achterliggende gegevens en worden (voortdurend) bijgewerkt. De frequentie van het actualiseren is sterk afhankelijk van de aard van het

verschijnsel: bij een hoge dynamiek en mobiliteit van de fraude moet ook het actualiseren van de profielen bijtijds gebeuren.

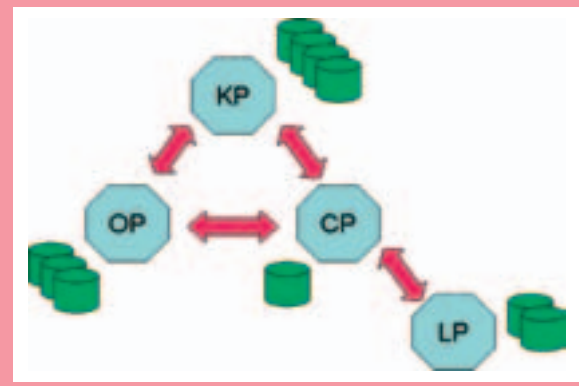
De inhoud van de profielen kan zowel uit harde gegevens bestaan als uit 'zachte' gegevens. We onderscheiden hierbij verder brongegevens (rechtstreeks uit de grote datastromen genomen) en metagegevens: afgeleide variabelen die gedrag en optreden van verschijnselen beschrijven. Het bedrag van een transactie en het tijdstip vallen in de eerste categorie, terwijl leeftijdscategorie van de klant en frequentie van zakendoen metavariabelen zijn. De slimheid en vaardigheid van medewerkers om de juiste effectieve profielen samen te stellen is één van de succesfactoren bij het detecteren van fraude.

In de profielen kunnen zich, behalve combinaties en voorkeuren, ook directe verwijzingen en links bevinden. Men kan zich voorstellen dat bij high-risk goederen alleen al het aanwezig zijn van een verwijzing naar een high-risk klant een effectieve regel kan opleveren die een keten van extra detecties in werking stelt.

Voorbeelden van toepassingen

Toepassingen liggen in alle processen waarbij datastromen worden gemonitord op mogelijke verstoringen en afwijkingen. Dit kunnen processen zijn die de

Figuur 7. Achterliggende datasets in een viercomponentenmodel.



toegang tot centrale computersystemen bewaken of het gebruik van elektronische devices zoals mobiele telefoons, PDA's en lokale computers. Verder alle processen waarbij centraal gegevens worden bijgehouden. Betalingsverkeer valt onder deze laatste categorie, maar men kan ook denken aan fraude met creditcards, debetcards, internetbetalingen en witwaspraktijken.

Literatuur:

Bunt, S. en M. van der Aalst, Risicosturing bijstandsfraude, 2003, Research voor Beleid bv, in opdracht van StimulanSZ.

Eiben, A.E. en J.E. Smith, 2003, Introduction to Evolutionary Computing, Springer, ISBN 3-540-40184-9.

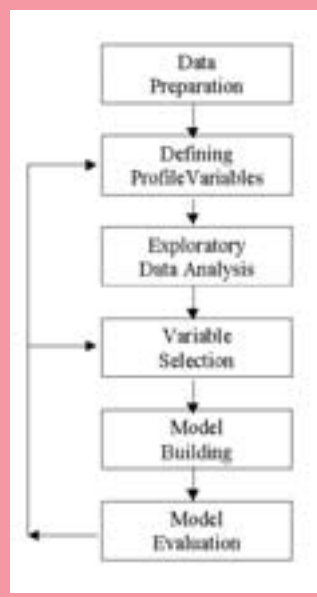
Hoeven, van der, G.J., D. Ruimschotel, R. van den Sigtenhorst en A.J.M. Verkoren, 2003, Kwetsbaarheid van de zorgsector voor georganiseerde fraude (in opdracht van het Ministerie van Justitie, november 2003), CMC/T11 Company, Amsterdam.

Kowalczyk, W., 1998, Rough Data Modeling: A new technique for analyzing data. In: L. Polkowski and A. Skowron (eds.) Rough Sets in Knowledge Discovery, pp. 400-421, Physica-Verlag, 1998.

Kowalczyk, W. and Piasta, Z., 1998, Rough sets-inspired approach to knowledge discovery in business databases. In Proceedings of The Second Pacific-Asia Conference on Knowledge Discovery and Data Mining, PAKDD-98, Melbourne, Lecture Notes in Artificial Intelligence, vol. 1394, Springer-Verlag, 186-197.

Kowalczyk, W. and Slisser, F., 1997, Analyzing customer retention with rough data models. In Proceedings of the 1st European Symposium on Principles of Data Mining and Knowledge Discovery, PKDD'97, Trondheim, Norway, Lecture Notes in AI 1263, Springer-Verlag, pp. 4-13.

Figuur 8. Overzicht van de te nemen stappen bij fraude-detectie.



Dezelfde technieken kan men overigens toepassen bij direct mailings, eventdriven marketing, het opsporen van ziekten en het opsporen van market opportunities.

Hoe doe je zoiets?

Het alert volgen en detecteren van (mogelijke) fraude vraagt om de beschikbaarheid van de juiste

data én om de mogelijkheid om snel variabelen te combineren, herdefiniëren, een andere gewichtsfactor te geven, enzovoorts. Het volgende schema (Figuur 8) geeft een beeld van de stapsgewijze benadering die wordt gevolgd bij het opbouwen van de profielen.

Dit schema is ontleend aan een van de projecten die op dit terrein heeft plaatsgevonden. Het voortdurend doorlopen van deze stappen en lussen moet er voor zorgen



dat het tempo van de fraudeurs wordt bijgehouden en dat nieuwe regels worden opgesteld en oude regels eventueel uitgeschakeld.

Mix

We staan aan de vooravond van het grootschalig gebruik van intelligente en adaptieve systemen voor het opsporen van fraude. Reeds nu al bewijzen deze technieken hun waarde. Succesvolle technieken zijn een mix van slimme wiskundige methoden én methoden die komen uit machine learning en kunstmatige intelligentie. En nu kijken wie er slimmer en sneller zijn: fraudeurs of het personeel van financiële instellingen. •

Bert Kersten is prinipal consultant bij LogicaCMG en tevens hoogleraar Bedrijfswiskunde aan de Vrije Universiteit van Amsterdam. Hij werkt op dit terrein veel samen met dr. Wojtek Kowalczyk, die werkzaam is aan de Vrije Universiteit Amsterdam en een aantal prijzen heeft gewonnen op het terrein van patroonherkenning en machine learning. Laatstgenoemde is de trekker van het Diana-project, nummer één van de ict-doorbraakprojecten 2003.

- 1 Volgens een andere bron zou de mogelijke totale fraude met verzekeringen in Nederland alleen 800 miljoen Euro bedragen
- 2 Zie bijvoorbeeld de Volkskrant, dd 15 juli 2003 en 5 november 2003