



Knowledge Discovery in Data: naar performante én begrijpelijke modellen van bedrijfsintelligentie

BART BAESENS, CHRISTOPHE MUES
& JAN VANTHIENEN

ABSTRACT

BEDRIJVEN HEBBEN GEDURENDE DE LAATSTE DECENNIA MASSALE HOEVEELHEDEN GEGEVENS VERZAMELD. MET DE TOENAME IN HARDWARE-REKENCAPACITEIT EN DE OPKOMST VAN GEAVANCEERDE DATA-MINING TECHNIKEN CREËERT DIT NIEUWE OPPORTUNITEITEN: HOE KUNNEN WE UIT DEZE DATA KENNIS ONTGINNEN EN DE DAARUIT RESULTERENDE BESLISSINGS-MODELLEN SUCCESVOL AANWENDEN ALS HULPINSTRUMENT VOOR EEN VERBETERDE BEDRIJFSVOERING? BIJ DE ONTWIKKELING VAN DERGELIJKE MODELLEN MOETEN MEERDERE KWALITEITSCRITERIA IN OGENSCHOUW GENOMEN WORDEN. IN DIT ARTIKEL GAAN WE DIEPER IN OP HET BELANG VAN DE TRADE-OFF TUSSEN PERFORMANTIE EN INTERPRETEERBAARHEID, EN WE ILLUSTREREN DEZE AAN DE HAND VAN EEN VOORBEELDTOEPASSING IN DE FINANCIËLE SECTOR: DE ONTWIKKELING VAN BESLISSINGS-ONDERSTEUNENDE SYSTEMEN VOOR KREDIETVERLENING.

BUSINESS INTELLIGENCE EN KNOWLEDGE DISCOVERY IN DATA

Business Intelligence (BI) omsluit een brede categorie van ICT-applicaties en -technologieën voor het ver-

zamelen, analyseren en verspreiden van bedrijfsinformatie, die de bedoeling hebben om de bedrijfsvoering te ondersteunen of te optimaliseren. Daarbij duiken doorgaans begrippen op als data warehousing, data mining en knowledge management. Met name het distilleren van bruikbare patronen uit de almaar groeiende stroom van ruwe data vormt een sleuteluitdaging binnen dit geheel. De geautomatiseerde ontginning van kennis, en het daarmee geassocieerde traject, wordt doorgaans onder de noemer *Knowledge Discovery in Data* (KDD) gevat. Het is een iteratief proces dat ruwweg uitgesplitst kan worden in drie fasen: (1) *sampling en data preprocessing*; (2) data mining; (3) de ontwikkeling van beslissings-ondersteunende systemen.

In de sampling en data-preprocessing fase wordt ondermeer beslist welke populatie gebruikt zal worden voor verdere analyse. Verder worden extreme observaties geïdentificeerd en ontbrekende waarden opgevangen. Op de opgeschoonde dataset wordt, in de daaropvolgende data-mining fase, een leeralgoritme losgelaten, dat de geëxtraheerde kennispatronen voorstelt in de vorm van, bijvoorbeeld, een neurale netwerk, beslissingsboom, regelverzameling of een statistisch classificatiemodel. In de laatste fase wordt het resulterende model getoetst aan de al bestaande expertise. Een cruciaal vraagstuk hierbij betreft de integratie van nieuw ontgonnen en bestaande kennis tot één coherent beslissingsondersteunend systeem. Toepassingen van KDD bestaan in bijna alle functionele domeinen waar voldoende data voorhanden zijn. Enkele frequente voorbeelden vinden we in marketing – we denken hierbij aan *market basket analyse*, waar het de bedoeling is patronen in het aankoopgedrag van klanten op te sporen, of bijvoorbeeld het voorspellen van klantverloop (churn prediction) –, in financieuzen (bijvoorbeeld *stock picking*), fraudedetectie, en zo meer. In wat volgt illustreren we enkele typische kenmerken, uitdagingen en mogelijke struikelblokken daarbij aan de hand van een concrete voorbeeldtoepassing: het gebruik van KDD bij de beoordeling van kredietaanvragen.

IN DIT NUMMER

PAG. 1 EN 4

KNOWLEDGE DISCOVERY IN DATA:
NAAR PERFORMANTE ÉN BEGRIPPELIJKE
MODELLEN VAN BEDRIJFSINTELLIGENTIE
Bart Baesens, Christophe Mues & Jan Vanthienen

PAG. 2-3

BAGGING VAN STATISTISCHE
CLASSIFICATIETEGELLEN
Christophe Croux en Aurélie Lemmens

EEN VOORBEELD TOEPASSING VAN KDD: CREDIT SCORING

In een kredietverleningscontext kan KDD toegepast worden voor de opstelling van modellen die de kredietwaardigheid van toekomstige klanten voorspellen. Gebaseerd op de kenmerken en het terugbetalingsgedrag van klanten uit het verleden tracht men hierbij modellen te schatten die de kans op succesvolle terugbetaling van nieuwe potentiële klanten zo nauwkeurig mogelijk berekenen (ook wel *credit scoring* genoemd). Op basis hiervan kan dan een beslissing genomen worden om de kredietaanvraag te aanvaarden dan wel te verwerpen. Het spreekt vanzelf dat we hier met een klassiek binair classificatieprobleem te maken hebben: is de klant een wanbetaler of niet, gegeven zijn inkomen, spaarmiddelen, huwelijksstatus, etc.? Een brede waaier van classificatietechnieken kunnen op dit probleem toegepast worden (Baesens et al. 2003b). Voorbeelden zijn statistische discriminant-analyse, beslissingsbomen, neurale netwerken, support vector machines, k-nearest neighbour, Bayesiaanse netwerken, fuzzy classificatoren, genetische algoritmen, tot zelfs zogeheten *'ant colony algorithms'*, gebaseerd op het gedrag van mierenkolonies. Met deze proliferatie aan (almaar complexere) technieken bestaat het gevaar dat men door de bomen het bos niet meer ziet en niet weet op welke basis een keuze te maken.

ACCURAAATHEID ALS KWALITEITSMATSTAF VOOR CREDIT-SCORING MODELLEN

Een eerste voor de hand liggend keuzecriterium is de discriminerende kracht of nauwkeurigheid van de ontwikkelde modellen. Hoewel nauwkeurigheid een intuïtief prestatiecriteria lijkt, dient erop gewezen te worden dat een ondubbelzinnige kwantificering ervan
(Vervolg op pag. 4)

ABSTRACT: CLASSIFICATIETECHNIKEN WORDEN GEBRUIKT OM TE KUNNEN VOORSPELLEN TOT WELKE GROEP EEN BEPAALD INDIVIDU ZAL BEHOREN, EN DIT OP BASIS VAN EEN REEKS OBSERVEERBARE KARAKTERISTIEKEN VAN HET INDIVIDU. ZO KAN MEN TRACHTEN TE VOORSPELLEN OF EEN PERSOON EEN AANGEBODEN PRODUCT ZAL KOPEN OF NIET, GEBRUIK MAKENDE VAN DIRECT OBSERVEERBARE VARIABELEN ZOALS LEEFTIJD, GESLACHT, SOCIAALDEMOGRAFISCHE CRITERIA, EN VROEGER KOOPGEDRAG. ZULK EEN PREDICTIE MAAKT HET VOOR EEN BEDRIJF MOGELIJK OM HAAR POTENTIËLE KLANTEN IN VERSCHILLENDE CATEGORIEËN IN TE DELEN EN HUN PUBLICITEITSCAMPAGNES EN PRODUCTAANBIEDINGEN GERICHTER

TE MAKEN. STANDAARDMETHODEN VOOR CLASSIFICATIE ZIJN LOGISTISCHE REGRESSIE EN CLASSIFICATIEBOMEN. IN DIT ARTIKEL WILLEN WE AANTONEN HOE EEN RECENTE STATISTISCHE METHODE, BAGGING GENAAMD, GEBRUIKT KAN WORDEN OM DE PERFORMANTIE VAN ZULKE CLASSIFICATIEMETHODEN TE VERBETEREN. DEZE TECHNIKEN VRAGEN GEEN EXTRA INFORMATIE OVER DE INDIVIDUËN, MAAR ZIJN WEL ERG REKENINTENSIEF. GEBRUIKMAKEND VAN EEN MODERNE COMPUTER IS HET ECHTER EENVOUDIG EN SNEL OM EEN BAGGING VERBETERING VAN EEN CLASSIFICATIETEGEL TE BEKOMEN.

Bagging van statistische

CHRISTOPHE CROUX EN AURÉLIE LEMMENS

1. OPSTELLEN VAN EEN CLASSIFICATIETEGEL

Onderstel dat een individu tot twee verschillende groepen kan behoren, die gecodeerd worden als groep 1 en 0. We noteren $y = 1$, respectievelijk $y = 0$, om aan te duiden tot welke groep een individu behoort. Voor dit individu observeren we dan de waarden van enkele verklarende variabelen, die we noteren met x_1, x_2, \dots, x_p . Men beschikt nu over een steekproef van individuen waarvoor we zowel de waarden van de verklarende variabelen als de waarde van y kennen. Deze steekproef noemt men de *training sample*, die men veelal kan samenstellen op basis van beschikbare gegevensbestanden. Beschouw als illustratief voorbeeld een steekproef van 1000 mensen die een kredietaanvraag deden, en waarvan men optekende of ze goede ($y = 1$), dan wel slechte ($y = 0$) terugbetalers zijn. Tevens gekend zijn de verklarende variabelen leeftijd (x_1 in jaren), stand van de zichtrekening ($x_2 = 0$ indien laag, $x_2 = 1$ indien hoog), afbetalingsbedrag als percentage van het beschikbare inkomen (x_3), en het bestaan van andere kredieten ($x_4 = 0$ indien ja, $x_4 = 1$ indien nee). Op basis van deze steekproef zal men dan de classificatieregels opstellen die toelaat om y te voorspellen voor nieuwe kredietaanvragers, waarvan men nog niet weet of ze hun lening zullen terugbetalen of niet. Dit probleem is gekend onder de naam *credit scoring*. We bespreken nu kort twee manieren om zo een classificatieregels op te stellen.

1a) Logistische regressie:

Hier maakt men gebruik van een statistisch model wat de kans dat men tot groep 1 behoort, gegeven de karakteristieken x_1, x_2, \dots, x_p , modelleert als $P(y = 1 | x_1, x_2, \dots, x_p) = 1 / (1 + \exp(b_1 x_1 + b_2 x_2 + \dots + b_p x_p))$. Op basis van de training sample worden de ongekende parameters b_1, b_2, \dots, b_p van dit model geschat. Voor ons voorbeeld bekomt men zo $b_1 = 0.019$, $b_2 = 0.089$, $b_3 = -0.18$ en $b_4 = -0.098$. Voor een nieuwe

observatie berekent men dan de lineaire combinatie $s = b_1 x_1 + b_2 x_2 + \dots + b_p x_p$, wat men ook de score noemt van deze nieuwe "kredietaanvragers". Op basis van deze score en bovenstaande formule bekomt men dan onmiddellijk de kans dat $y = 1$, in ons voorbeeld dus de kans dat het krediet terugbetaald wordt. Indien deze kans groter is dan 50%, zal men voorspellen dat de lening zonder problemen terugbetaald zal worden.

1b) Classificatiebomen:

Logistische regressie is een lineaire methode in die zin dat alle informatie in de verklarende variabelen in één enkele lineaire combinatie, de score genaamd, samengevat

variabelen opsplijst in meerdere rechthoeken (meer precies hyper-rechthoeken). Dit gebeurt volgens een slim opgesteld algoritme, wat tot een optimale verdeling van de ruimte in verschillende rechthoeken leidt. Elke rechthoek correspondeert dan met een voorspelde waarde van y . Deze bekomen verdeling in rechthoeken kan visueel voorgesteld worden door een boomstructuur, zoals we zien in Figuur 1. Meerdere statistische softwarepakketten laten toe zulke classificatiebomen op te stellen. Wanneer we nu een nieuwe observatie hebben, zal deze de boom binnenkomen aan de wortel (bovenaan de boom), en afhankelijk van de waarden van de verklarende variabelen de verschillende knopen van de boom doorlopen. Uiteindelijk belandt men bij een laatste knoop, wat ook

een bladknoop genoemd wordt. Met elke bladknoop is een getal geassocieerd, gelijk aan de relatieve frequentie van het aantal observaties die in deze bladknoop belanden waarvoor $y = 1$. Dit getal is dan ook een schatting voor de kans om tot de eerste groep te behoren. Het aantal bladknopen bepaalt dan de complexiteit van de boom. Het is nu een eenvoudige oefening om te zien dat een observatie waarvoor $x_1 = 20$, $x_2 = \text{hoog}$, $x_3 = 2$, $x_4 = \text{ja}$, zal belanden in een bladknoop waarvoor de kans dat $y = 1$ geschat wordt op $8/24$. Gezien deze waarde kleiner is dan 0,5, voorspelt men dan dat dit individu tot de groep van de slechte betalers zal behoren.

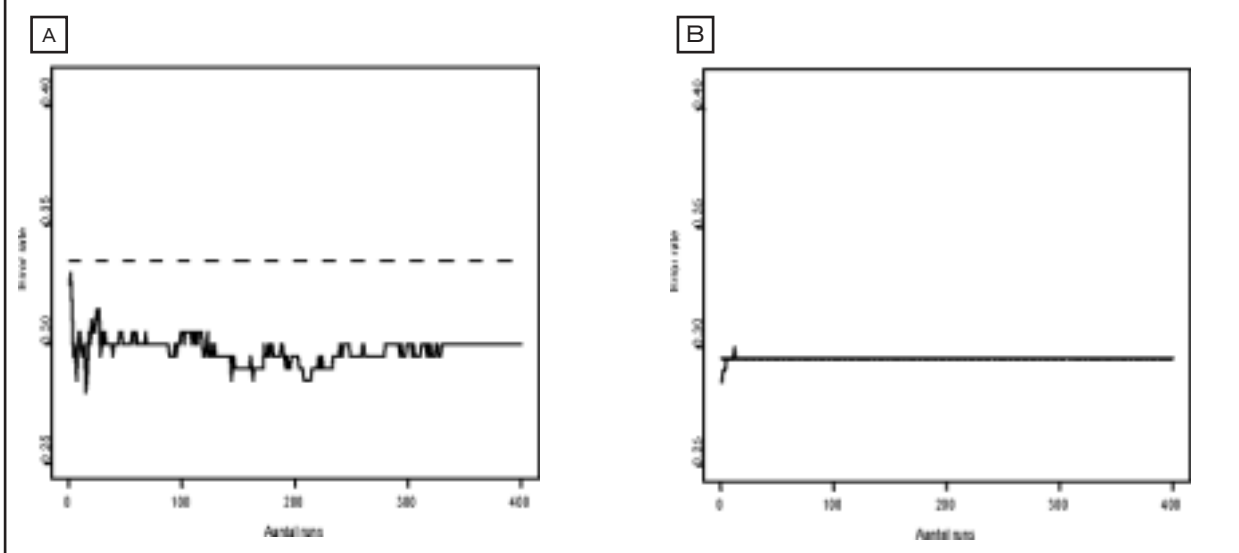
De performantie van verschillende classificatieregels kan gemeten worden met verschillende criteria. Het eenvoudigste criterium is het percentage goed geklasseerde observaties: dit zijn observaties waarvan de voorspelde en geobserveerde waarde van y aan elkaar gelijk zijn. Het is wel belangrijk dat we hiervoor observaties nemen die niet tot de training sample behoren, om te vermijden dat er interferentie ontstaat tussen het opstellen van de classificatieregels en het valideren ervan. Men kan bijvoorbeeld 20% van de observaties van de training sample afsplitsen en voorbehouden om de classificatieregels te valideren. Het percentage goed geklasseerde observaties in een *validation sample* noemen we daarom de (geschatte) *error rate*. In het eenvoudige voorbeeld van de credit scoring data, bekomt men een error rate van 29,5% voor de

FIGUUR 1: CLASSIFICATIEBOOM VOOR HET CREDIT SCORING VOORBEELD. DE KNOPEN VAN DE BOOM VOORSPELLEN OF MEN EEN GOEDE OF EEN SLECHTE TERUGBETALER ZAL ZIJN.

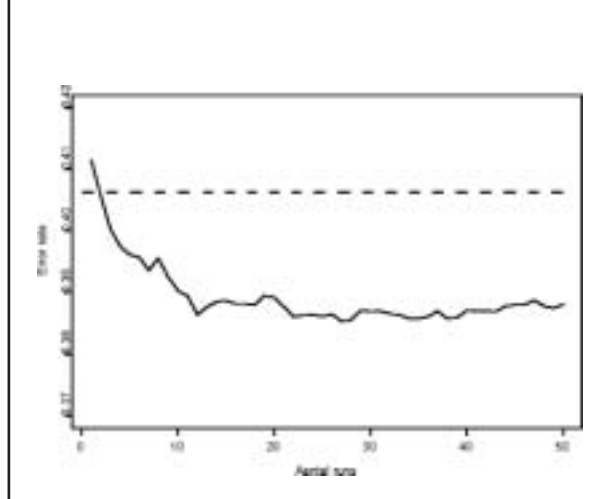


wordt. De ruimte der verklarende variabelen wordt opgesplitst in twee half-vlakken (meer precies half-hyper-vlakken), elk corresponderend met één waarde van de te voorspellen binaire variabele. Classificatiebomen zijn nu een niet-lineaire techniek, die de ruimte van de verklarende

FIGUUR 2: ERROR RATE VOOR DE BAGGED (A) CLASSIFICATIEBOOM (B) LOGISTISCHE REGRESSIE METHODE VOOR HET CREDIT SCORING VOORBEELD, IN FUNCTIE VAN HET AANTAL RUNS.



FIGUUR 3: ERROR RATE VOOR DE BAGGED CLASSIFICATIEBOOM VOOR HET CHURNING VOORBEELD, IN FUNCTIE VAN HET AANTAL RUNS



che classificatieregels

methode van logistische regressie, en 33,5% voor de beslissingsboom.

Het is echter zeker niet algemeen waar dat logistische regressie betere resultaten geeft dan classificatiebomen. Het blijkt dat voor meer complexe problemen, met een groot aantal verklarende variabelen, classificatiebomen beter werken. In [3] wordt een voorbeeld uitgewerkt uit de telecommunicatiesector waarin men moet trachten te voorspellen of een klant trouw blijft aan een bepaalde telecomoperator, wat men in de marketing een *churning* probleem noemt. Het aantal variabelen, zelfs na een zinvolle variabelenselectie, bedraagt nog steeds een veertigtal variabelen. Enkele van die variabelen zijn bijvoorbeeld het gemiddelde gebruik van de operator in minuten over het laatste jaar (een gedragsvariabele), abonnementsprijs (een monetaire bedrijfsspecifieke variabele), aantal promotieaanbiedingen (een klant-bedrijf interactievariabele), inkomen (een socio-demografische variabele), ... Voor dit churning voorbeeld bekomen we een error rate van 43,6% voor logistische regressie, en voor de classificatieboom 40,6%.

2. BAGGING VAN EEN CLASSIFICATIEREGEL

De Bootstrap AGGREGATING methode om een classificatieregel te verbeteren werd geïntroduceerd door [1], en is geïnspireerd door het meer gekende bootstrap principe. Noteer de geobserveerde steekproef met Z en stel door x de waarden voor van de verklarende variabelen voor een nieuwe observatie waarvoor we een voorspelling willen maken. Het idee is om op een artificiële manier nieuwe training samples Z^1, Z^2, \dots, Z^B te creëren, waarvoor men dan telkens opnieuw de kans om tot groep 1 te behoren berekent. Dit geeft ons dan als voorspelde kansen $p^1(x), p^2(x), \dots, p^B(x)$ waar B het aantal runs genoemd wordt. De uiteindelijke voorspelde kans is dan simpelweg het gemiddelde van al deze kansen:

$$P_{bag}(x) = \frac{1}{B} (p^1(x) + p^2(x) + \dots + p^B(x))$$

De nieuwe steekproeven Z^1, Z^2, \dots, Z^B hebben dezelfde steekproefgrootte als Z , en worden bekomen door op een willekeurige manier observaties uit de oude steekproef Z te trekken, maar met teruglegging. Deze artificiële steek-

proeven bevatten dus elementen van Z , waarvan er sommige meerdere keren kunnen voorkomen, en andere geen enkele keer. Het verassende is nu dat dit bootstrap idee de error rate van classificatiemethoden kan verbeteren.

In Figuur 2a toont de volle lijn, voor het voorbeeld van de credit scoring data, hoe de error rate van de bagging techniek toegepast op de classificatieboom verandert in functie van het aantal runs B . Men ziet dat de error rate eerst daalt met toenemend aantal runs, en voor B groter dan 50 vrij stabiel wordt. De grootte B bepaalt de hoeveelheid rekenwerk, en zou hier dus gelijk aan 50 gekozen kunnen worden. De error rate van de oorspronkelijke classificatieboom was, zoals weergegeven door de stippellijn in de grafiek 33,5%, en kan door de bagging techniek gereduceerd worden tot 30%. Een gelijkaardige figuur vindt men we in Figuur 2b, maar dan voor logistische regressie. Merk op dat er hier helemaal geen reductie is van de error rate. Men kan immers aantonen dat voor classificatiemethoden die "te lineair" zijn, zoals logistische regressie, bagging geen verbetering geeft.

Ook voor het meer complexe *churning* probleem, waar we tientallen variabelen en duizenden observaties hebben, kan men zonder problemen de bagging techniek toepassen. In Figuur 3 ziet men een duidelijke daling van de error rate die zich stabiliseert vanaf $B = 20$. Voor een bagging met 50 runs bekomt men een error rate van 38,8%, wat een relatieve verbetering is met bijna 5%. Dit lijkt misschien niet erg veel, maar men moet beseffen dat deze verbetering komt zonder enige supplementaire gegevensinformatie. De reductie in error rate is enkel te wijten aan de statistische bagging techniek, en kost een weinig meer rekentijd. In een bedrijfscontext kunnen enkel percenten meer correct geklasseerde observaties wel degelijk het verschil maken. Onderstaand voorbeeld toonde aan dat classificatiebomen kunnen toegepast worden op zeer grote complexe datasets, die vele variabelen mogen bevatten. Hun performantie kan verbeterd worden door het toepassen van de *bagging* techniek. Zoals in [2] kan men dan ook besluiten: classificatiebomen gecombineerd met *bagging* is een van de beste algemeen en dadelijk toepasbare classificatiemethoden. Er is echter nog heel wat onderzoek te verrichten om te weten in welke gevallen deze techniek een significant voordeel oplevert, en welke varianten het meeste renderen in welke omstandigheden.

CHRISTOPHE CROUX is hoofddocent aan het departement Toegepaste Economische Wetenschappen van de K.U. Leuven. Hij doceert cursussen econometrie en statistiek.

Hij onderzoekt en ontwikkelt statistische methodes, zowel vanuit een fundamenteel wiskundig als vanuit een toegepast oogpunt. Focus is op robuuste en multivariate technieken.

Email: christophe.croux@econ.kuleuven.ac.be
URL: www.econ.kuleuven.ac.be/christophe.croux



AURÉLIE LEMMENS studeerde handelsingenieur, optie Marketing, aan de Solvay Business School te Brussel. Zij bereidt momenteel een doctoraat voor bij de vakgroep Kwantitatieve methoden aan het departement Toegepaste Economische Wetenschappen van de K.U. Leuven.

Email: aurelie.lemmens@econ.kuleuven.ac.be



REFERENTIES:

- [1] Breiman, L. (1996), Bagging Predictors, *Machine Learning*, 26, 123-140.
- [2] Hastie, T., Tibshirani, R. and Friedman, J. (2001), *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, Springer Verlag: New York.
- [3] Lemmens, A., and Croux, C. (2003), *Bagging and Boosting Classification Rules for Predictive Market Segmentation*, manuscript.

niet altijd evident is. Als eerste naïeve benadering zou men kunnen streven naar het maximaliseren van het aantal correct geclassificeerde klanten op een onafhankelijke testset. Hoewel dit zeker geen slecht criterium is, vertoont het toch een aantal tekortkomingen. Slechts een klein aantal klanten zal wanbetaler zijn, wat ervoor zorgt dat een weinig informatieve regel zoals 'elke klant is een goede klant' reeds een heel goede prestatie oplevert. Men moet, met andere woorden, dus ook andere aspecten beschouwen, zoals de misclassificatiekosten van vals negatieven versus die van vals positieven. Deze zijn echter moeilijk te kwantificeren, aangezien de kosten typisch zullen variëren van klant tot klant (afhankelijk van het bedrag van de lening, interestvoet, en dergelijke) en bovendien ook nog eens over de tijd. Het meten van de accuraatheid van een classificatiemodel voor kredietverlening is dus al helemaal geen triviale oefening. Hetzelfde geldt trouwens voor verscheidene andere typische KDD-toepassingen. Bovendien is het zeker niet het enige criterium van belang.

DE ROL VAN OCCAM'S RAZOR VERTAALD NAAR CREDIT SCORING

William van Ockham, een bekende 14de-eeuwse filosoof, benadrukte dat modellen behalve accuraat ook begrijpelijk en eenvoudig moeten zijn (Occam's razor). Zo zal een eenvoudig model sneller en beter in de bedrijfscontext geïntegreerd kunnen worden dan een complex, sterk geparametriseerd black-box model. Deze keuze houdt doorgaans een trade-off in, aangezien complexe modellen vaak ook beter presteren inzake nauwkeurigheid.

Neem bijvoorbeeld neurale netwerken. Doordat deze laatste universele approximators zijn, leidt hun toepassing vaak tot zeer goed presterende modellen (Baesens et al. 2003b). Echter, een belangrijk nadeel naar de bedrijfsbesluitvorming toe is hun beperkte verklarende kracht: hoewel zij het mogelijk maken erg accurate uitspraken of predicties te doen, is het pijnpunt vaak dat de precieze wijze waarop zij dergelijke beslissingen afleiden niet pasklaar beschikbaar of eenvoudig interpreteerbaar is. Figuur 1 toont een voorbeeld van een neurale netwerk dat getraind werd voor het schatten van de kredietwaardigheid van klanten van een financiële instelling in de Benelux: krachtig maar moeilijk interpreteerbaar.

FIGUUR 2. 'ALS-DAN'-REGELS GEËXTRAHEERD UIT HET NEURALE NETWERK VAN FIGUUR 1

Als Looptijd > 12 maanden **en** Doel = cash provisie **en** Spaarmiddelen <= 12.40 Euro **en** Aantal jaren klant <= 3 **dan** Klant = slecht

Als Looptijd > 12 maanden **en** Doel = cash provisie **en** Eigendom onroerend goed = Neen **en** Spaarmiddelen <= 12.40 Euro **dan** Klant = slecht

Als Doel = cash provisie **en** Inkomen > 719 Euro **en** Eigendom onroerend goed = Neen **en** Spaarmiddelen <= 12.40 Euro **en** Aantal jaren klant <= 3 **dan** Klant = slecht

Als Doel = tweedehandswagen **en** Inkomen > 719 Euro **en** Eigendom onroerend goed = Neen **en** Spaarmiddelen <= 12.40 Euro **en** Aantal jaren klant <= 3 **dan** Klant = slecht

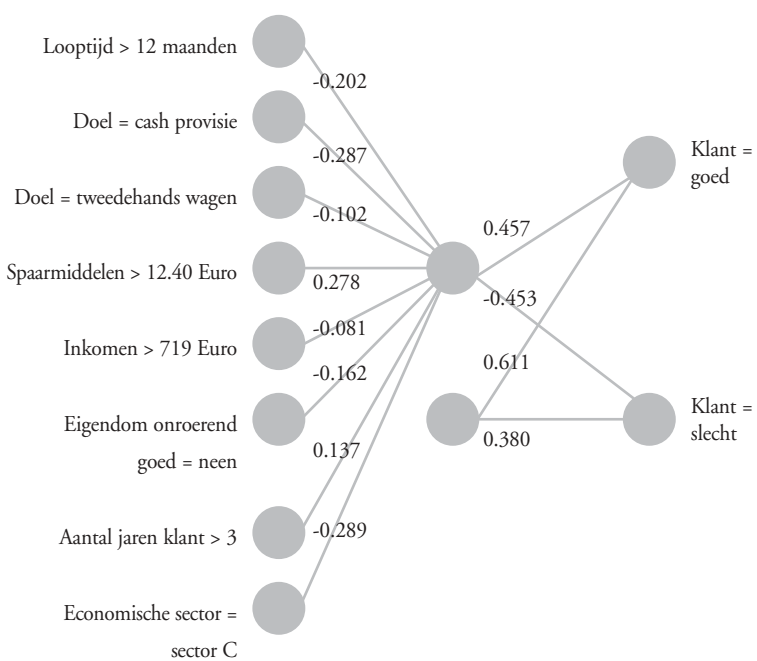
Als Spaarmiddelen <= 12.40 Euro **en** Economische sector = Sector C **dan** Klant = slecht

Default klasse: Klant = goed

FIGUUR 3. BESLISSINGSTABEL VOOR DE REGELS VAN FIGUUR 2

1. Spaarmiddelen (Euro)			≤ 12.40						> 12.40					
2. Economische sector	Sector C		andere						-					
3. Doel	-		cash provisie				tweedehandswagen		ander				-	
4. Looptijd	-		≤ 12 maanden			> 12 maanden			-		-		-	
5. Aantal jaren klant	-		≤ 3		> 3		≤ 3		> 3		-		-	
6. Eigendom onroerend goed	-		ja		nee		ja		nee		ja		nee	
7. Inkomen (Euro)	-		≤ 719		> 719		≤ 719		> 719		-		-	
1. Klant = goed	-	x	x	-	x	-	x	-	x	x	-	x	x	x
2. Klant = slecht	x	-	-	x	-	x	-	x	-	-	x	-	-	-
	1	2	3	4	5	6	7	8	9	10	11	12	13	14

FIGUUR 1. NEURALE NETWERK VOOR HET VOORSPELLEN VAN KREDIETWAARDIGHEID GETRAIND OP BASIS VAN 3123 OBSERVATIES



In de financiële sector (en tevens in andere contexten) bestaat dan ook een sterke terughoudendheid tegenover het gebruik van neurale netwerken, precies omwille van hun intransparantie. In Baesens et al. (2003a) werd daarom een benadering voorgesteld waarin de neurale netwerk black-box wordt geopend met behulp van regelextractiemethoden. Zonder zulke bijkomende, in dit geval regelgebaseerde, voorstellingswijze is de kans immers groot dat de organisatie zelf onvoldoende vertrouwen zou hebben in de correcte werking van het model. Bovendien bestaat er in sommige landen een wettelijke verplichting inzake de openbaarheid van het gehanteerde model. Figuur 2 bevat de 'als-dan'-regels die uit het netwerk van Figuur 1 werden geëxtraheerd. Deze regels zijn eenvoudig te interpreteren en bovendien krachtig: zij blijken namelijk even accuraat als het netwerk uit Figuur 1. De regels kunnen vervolgens op een gebruiksvriendelijke en efficiënt hanteerbare manier gevisualiseerd worden met behulp van beslissingstabellen (zie Figuur 3) (Mues 2002).

BESLUIT

Bij het gebruik van KDD voor de ontwikkeling van intelligente beslissingsondersteunende systemen spelen tal van aspecten een rol. In dit artikel benadrukten we dat de geëxtraheerde modellen idealiter zowel accuraat als begrijpelijk zijn. Wat het eerste betreft, argumenteerden we dat het meten van de accuraatheid geen triviale oefening is, en zeker nog ruimte biedt voor toekomstig onderzoek, zowel binnen de specifieke voorbeeldcontext van credit scoring als in andere toepassingen. Ter verbetering van hun interpreteerbaarheid, stelden we vervolgens voor om uit een krachtig getraind neurale netwerk een 'als-dan'-regelset te extraheren en te visualiseren in de vorm van een beslissingstabel. Ook andere representatievormen echter (zoals bijvoorbeeld de recent voorgestelde Bayesiaanse netwerken) kunnen mogelijkwijs intuïtieve en accurate modellen opleveren en vormen dan ook een interessant topic voor een vervolgstudie.

BART BAESENS
is doctoraal student aan het Departement Toegepaste Economische Wetenschappen van de K.U.Leuven, vakgroep beleidsinformatica.



E-mail: Bart.Baesens@econ.kuleuven.ac.be

CHRISTOPHE MUES
is postdoctoraal onderzoeker aan het Departement Toegepaste Economische Wetenschappen van de K.U.Leuven, vakgroep beleidsinformatica.



E-mail: Christophe.Mues@econ.kuleuven.ac.be

JAN VANTHIENEN
is gewoon hoogleraar aan het Departement Toegepaste Economische Wetenschappen van de K.U.Leuven, vakgroep beleidsinformatica.



E-mail: Jan.Vanthienen@econ.kuleuven.ac.be

REFERENTIES:

- BAESENS B., SETIONO R., MUES C., VANTHIENEN J., Using Neural Network Rule Extraction and Decision Tables for Credit-Risk Evaluation, Management Science, Volume 49, Issue 3, forthcoming, March, 2003a.
- BAESENS B., VAN GESTEL T., VIAENE S., STEPANOVA M., SUYKENS J., VANTHIENEN J., Benchmarking State of the Art Classification Algorithms for Credit Scoring, Journal of the Operational Research Society, forthcoming, 2003b.
- MUES C., On the Use of Decision Tables and Diagrams in Knowledge Modeling and Verification, PhD dissertation, K.U. Leuven, Dept. of Applied Economic Sciences, 223 pp., 2002.

CENTRUM VOOR TOEGEPAST ECONOMISCH ONDERZOEK

Voor informatie over onderzoek (groepen, seminars, jaarverslag), bezoek de website van het Centrum voor Toegepast Economisch Onderzoek: <http://www.econ.kuleuven.ac.be/cteo/>

Een lijst van onderzoeksrapporten met abstract is beschikbaar op: <http://www.econ.kuleuven.ac.be/cteo/reports/>

Reacties op Business IN-zicht zijn altijd welkom bij Linda Van de Gucht

(Linda.Vandegucht@econ.kuleuven.ac.be)

Voor een gratis abonnement op Business IN-zicht contacteer:

Elke.Tweepenninckx@econ.kuleuven.ac.be



KATHOLIEKE UNIVERSITEIT
LEUVEN