



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 neuraal 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)

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.

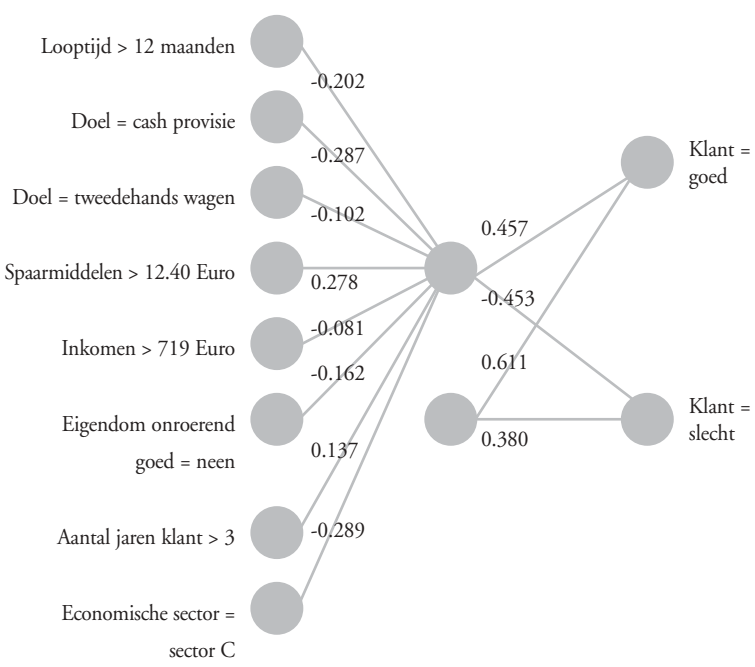
FIGUUR 2. 'ALS-DAN'-REGELS GEËXTRAHEERD UIT HET NEURALE NETWERK VAN FIGUUR 1

- 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	≤ 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	
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

