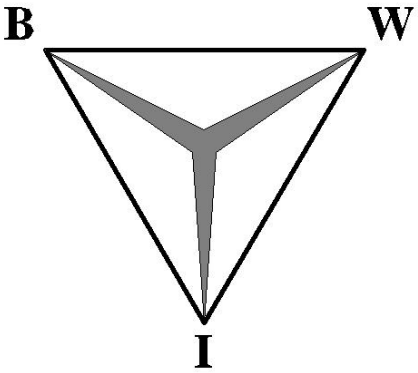


Datamining Services

Een andere manier van onderzoeken

(een business case voor besluitvorming)

Door: T. Witteveen

Zoetermeer, 29 juli 2003

Inhoudsopgave

1 Investeringssamenvatting

Moet EIM investeren in Datamining Services[1]? Is het opzetten van deze nieuwe dienst rendabel? Dat zijn de kernvragen in dit investeringsplan.
Datamining is een manier om informatie uit grote hoeveelheden gegevens te halen. Het is een nieuwe, andere manier van onderzoek doen, die goed aansluit bij de kernactiviteit van EIM.
Volgens onderzoek ontstaat er steeds meer behoefte aan Datamining Services[2]. Hiervoor zijn de volgende redenen aan te geven:

- door de automatisering zijn er steeds meer gegevensbronnen aanwezig
- de capaciteit om gegevens op te slaan is sterk vergroot en de kosten om gegevens op te slaan zijn sterk verminderd
- de softwareprogramma's worden steeds beter en gebruiksvriendelijker
- de gegevens bevat informatie die essentieel is in de huidige competitieve markt
- de bedrijven / instellingen hebben zelf geen kennis en kunde om datamining uit te kunnen voeren

Datamining Services kan antwoord geven op vragen als:

- Veranderen patronen van klanten veranderen door bepaalde advertenties?
- Gaan ondernemers na een lastenverlichting meer investeren?

Kortom, datamining kan antwoord geven op vragen die de klanten van EIM beantwoord willen hebben. Datamining is geen triviale bezigheid. Het hele proces dat doorlopen moet worden vereist veel kennis over de gegevens en men moet het proces goed begeleiden om tot het juiste antwoord te komen.
Aangestoken door het succes van datamining bij banken en creditcard maatschappijen zullen ook andere bedrijven van deze techniek willen profiteren. Deze bedrijven zijn zelf niet in staat om datamining uit te voeren. Ze missen de kennis en kunde, en de aanschaf van de software is voor deze bedrijven te duur. Daar gaat een markt ontstaan voor bedrijven als EIM om Datamining Services op te zetten. Het nadeel is dat de vraag naar de techniek op dit moment er nog niet is. Men probeert de voordelen van datamining aan de markt kenbaar te maken. Er zijn ideeën te over voor het toepassen van datamining in de markt van zowel beleidsinstanties als bedrijven. Deze ideeën zijn groot in aantal, maar concrete toepassingen ontbreken nog.
Daarnaast zijn er financiële risico's verbonden aan het aanschaffen van de datamining software. Voor Clementine van SPSS zijn de kosten € 63.100 in het eerste en € 47.140 in de opvolgende jaren. Bij de aanschaf van enterprise miner van SAS zijn de kosten nog hoger, namelijk ongeveer € 120.000 in het eerste en ongeveer € 80.000 in de opvolgende jaren.
Een break-even scenario om de kosten van de aanschaf van dataminingsoftware en opleidings- en promotiekosten terug te verdienen volgt hierbij. Het zijn de noodzakelijke opbrengsten van Datamining Services waardoor de gemaakte kosten in 4 jaar terugverdiend kunnen worden. De klant zal een bedrag van € 1000 (clementine) of € 2000 (enterprise miner) per dag extra moeten betalen voor het gebruik van de datamining software. Onder realistische aannames zullen deze opbrengsten niet haalbaar zijn.

ROI Datamining SPSS	0,82%

[1] In het document zal Datamining Services als aanduiding van de nieuwe dienst gebruikt worden; de term datamining staat voor de combinatie van datamining methoden en technieken

[2] Elliot K. Scionti, R. Page, M. (2003) Two Rivers, The confluence of data mining and market research for smarter CRM, zie bijlage VI

	Jaar 0	Jaar 1	Jaar 2	Jaar 3
Kosten	€ 113.520	€ 65.568	€ 65.568	€ 65.568
Opbrengsten	€ 20.000	€ 48.000	€ 100.000	€ 160.000
Verschil	-€ 93.520	-€ 17.568	€ 34.432	€ 94.432
Netto contante waarde tegen 5% rekenrente	-€ 93.520	-€ 16.731	€ 31.231	€ 81.574

ROI Datamining SAS	0,37%			
	Jaar 0	Jaar 1	Jaar 2	Jaar 3
Kosten	€ 181.800	€ 105.000	€ 105.000	€ 105.000
Opbrengsten	€ 31.875	€ 76.500	€ 159.375	€ 255.000
Verschil	-€ 149.925	-€ 28.500	€ 54.375	€ 150.000
Netto contante waarde tegen 5% rekenrente	-€ 149.925	-€ 27.143	€ 49.320	€ 129.576

Ondanks dat de combinatie EIM en datamining veel voordelen biedt, zoals

- datamining Services sluit goed aan bij de kernactiviteit van EIM
- er zal behoefte voor Datamining Services ontstaan
- Datamining Services staat dicht bij de huidige kernactiviteit.
- EIM heeft de mensen en de middelen, behalve de software, al in huis om deze dienst in de markt te introduceren.
- de naam van EIM is bekend en vertrouwd in de markt

zijn de nadelen op dit moment groter, namelijk

- De behoefte aan Datamining Services groeit, maar is nog niet volop aanwezig
- De financiële risico's zijn groot.

Het grootste struikelblok is de kosten van de software, dit probleem zou ondervangen kunnen worden door een strategisch partnership met een universiteit of softwareleverancier op het gebied van datamining. Een samenwerking met een strategische partner dient nader onderzocht te worden.

Het is mijn mening dat EIM op dit moment niet zou moeten investeren in Datamining. EIM zou wel, mede gezien de huidige dienstverlening, de markt op het gebied van datamining in de gaten moeten blijven houden. Want in de toekomst zijn er kansen voor EIM op het gebied van Datamining Services. De vraag is op welk moment EIM hiervoor stappen moet zetten. De positieve punten van de combinatie van EIM en datamining zijn onmiskenbaar.

2 Inleiding

Onderzoek doen op basis van gegevens is dé kwaliteit van EIM. In de loop der jaren is het bedrijf er groot mee geworden. Door de groei van de automatisering van bedrijven wordt er steeds meer gegevens verzameld. Veel gegevens van allerlei aard worden geregistreerd en opgeslagen in databases. Deze databases worden steeds groter en onoverzichtelijker, mede door de samenvoeging van verschillende databases in zogenaamde datawarehouses. Databases met honderdduizenden records met tientallen variabelen zijn geen uitzondering. Deze gegevens bevat informatie. Klassieke methoden op basis van statistische data analyse door pakketten als SPSS zijn niet meer toereikend om de informatie te ontdekken. Daarvoor zijn de gegevens te ongestructureerd, ontbreken gegevens en zijn de bestanden te groot. Ook kunnen deze klassieke methoden geen antwoord op sommige vragen geven. Vragen over individuele klanten zijn met klassieke technieken niet te beantwoorden. Er zijn nieuwe methoden nodig om de gegevens te analyseren en vragen te beantwoorden. De methode die daarvoor gebruikt wordt is datamining: op een geautomatiseerde manier gegevens analyseren, verbanden en (gedrags)patronen ontdekken in de gegevens met het doel nuttige informatie te verschaffen aan de gebruiker. Deze technieken maken gebruik van de meest recente ontwikkelingen op het gebied van kunstmatige intelligentie (artificial intelligence). Datamining methoden kunnen informatie uit gegevens halen die men zelf niet zou vermoeden. In onderstaand plaatje is te zien hoe datamining (dm), naast marktonderzoek (mr) in een overzicht van bedrijfsactiviteiten gevat kan worden.

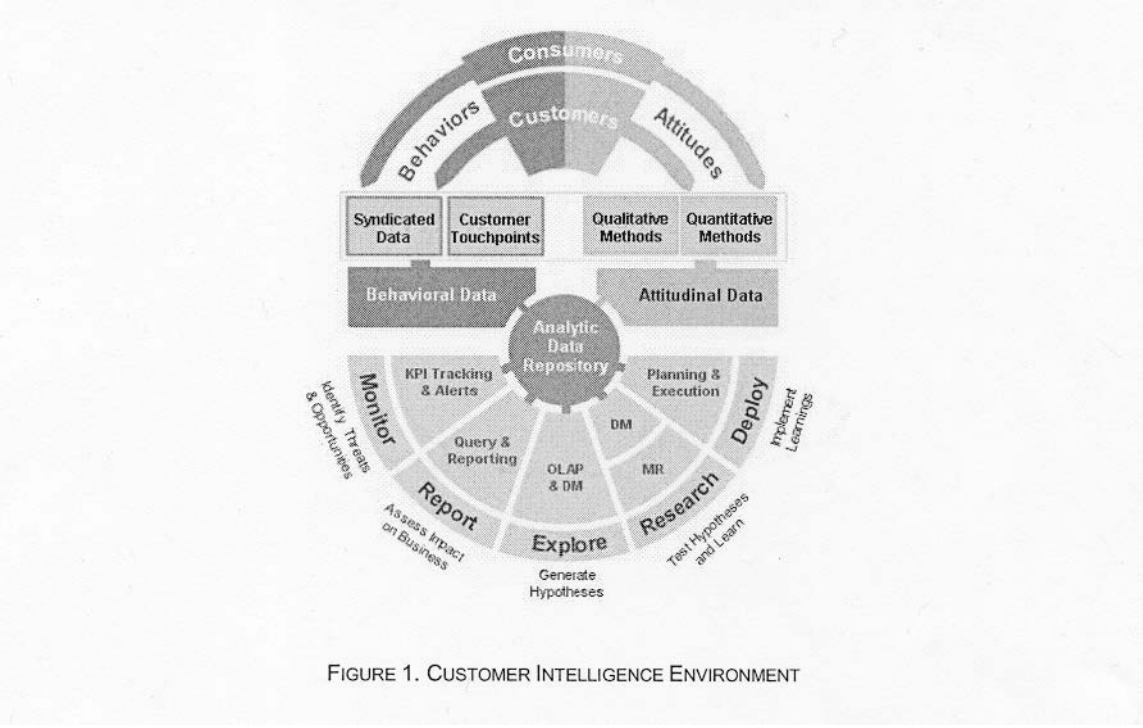


FIGURE 1. CUSTOMER INTELLIGENCE ENVIRONMENT

Datamining is een andere manier van informatie te verkrijgen. Sterke punten van datamining zijn onder andere dat er geen aannames vooraf gedaan hoeven te worden over de onderzoeken vraag en het toepassen van niet-lineaire methodes. Datamining technieken vinden de informatie, op welke manier is voor de klant af en toe lastig te begrijpen. Datamining

is meer dan de gegevens aan een programma aanbieden, er gaat een heel traject aan voorraf. Dit traject is lastig en vereist kennis over de gegevens en begrip van datamining methoden, datamining is een vak apart. In de markt lijkt de vraag naar dit specialisme toe te nemen. SPSS organiseert rond september / oktober een voorlichtingsbijeenkomst over datamining speciaal voor ondernemers.

Nog niet veel bedrijven en onderzoeksbureaus beheersen datamining en bieden een datamining service aan. EIM BV[1] en met name de unit DataWarehousing heeft ruime ervaring met gegevens en beschikt ook over benodigde kennis op economisch-wiskundig gebied om de methoden te kunnen leren. Het is deze unieke combinatie van kennis en ervaring die het mogelijk maakt om de een aantal medewerkers van DataWarehousing in de markt te zetten als datamining experts. Het is een nieuwe manier van "data voor beleid".

Is datamining een nuttige aanvulling van de diensten van EIM? En is het opzetten van deze service rendabel? Op deze vragen moet dit investeringsplan een antwoord geven. Om deze vraag goed en volledig te beantwoorden zullen eerst onderstaande vragen beantwoord worden:

- Wat zijn datamining technieken?
- Hoe ziet de markt eruit?
- Sluit deze dienst aan op de huidige strategie van EIM?
- Welke investeringen, in mensen in hard- en software, zijn nodig om bedrijven op datamining gebied te kunnen adviseren?
- Welke gevolgen en risico's heeft het opzetten van deze dienst voor EIM en de unit DataWarehousing?

De beantwoording van deze vragen wordt door verschillende analyses, zowel kwalitatief als kwantitatief ondersteund. Ook zal er een stappenplan, roadmap, voor het opzetten van deze dienst opgesteld worden. Tot slot zullen deze analyses leiden tot een conclusie en zullen verdere aanbevelingen gedaan worden.

[1] in het vervolg wordt EIM BV aangeduid als EIM

3 Wat is datamining

3.1 Wat is datamining? [1]

Datamining is het proces van het ontdekken van interessante informatie uit grote hoeveelheden gegevens die opgeslagen zijn in databases, datawarehouses of andere opslagmogelijkheden. Ook wordt vaak de term knowledge discovery in databases (kdd) gebruikt. Datamining analyseert de gegevens en ontdekt voor de gebruiker nuttige en relevante informatie. Op basis van deze gegevens kunnen trends ontdekt worden, verbanden gevonden worden en beslissingen genomen worden. De basisvragen bij datamining zijn:

1 kunnen we gebruik makend van de historische kennis nu beter, efficiënter beleid voeren
2 Kunnen we het toekomstige gedrag voorspellen als we de opgeslagen gegevens analyseren
3 Welke onbekende informatie is te vinden in de gegevens die wij in de loop der jaren verzameld hebben?

Bij datamining wordt gedacht aan het samenstellen van profielen of het herkennen van bepaalde patronen. Als voorbeeld kan de bonuskaart van de Albert Heijn dienen. Deze kaart registreert alle aankopen van de klant. Op basis van deze gegevens wordt ontdekt dat welke artikelen klanten vaak tegelijk kopen (market basket analyse). Van sommige producten mag dat verwacht worden: bier en chips of beleg en kaas. Er zijn ook minder voor de hand liggende combinaties mogelijk, zoals bier en luiers. Op basis van deze analyses kan men winkels anders inrichten met het doel meer producten te verkopen. Door deze klantengegevens te koppelen aan andere kennis over die periode, feestdagen of informatie over het weer, kunnen nieuwe ontdekkingen gedaan worden. Die bevindingen kunnen bijvoorbeeld leiden tot aanpassing van de voorraden.
Sommige datamining methoden sporen verbanden op zonder vooronderstellingen. Dit kan leiden tot verrassende conclusies. De gegevens zijn meestal ruwe onoverzichtelijke gegevens; er zijn nog geen gegevens samengevat of verwijderd. Datamining kan grote aantallen gedetailleerde gegevens analyseren en daar informatie uit halen.
Datamining proces is niet alleen het toepassen van de datamining technieken, het voorbereidende traject is ook belangrijk. Dit traject wordt beheerst door de keuzes die de gebruiker moet maken om de relevante patronen te vinden.
Het datamining proces bestaat uit de volgende stappen:

1) opschonen van gegevens
2) integreren van gegevens
3) selecteren van gegevens
4) transformeren van gegevens
5) datamining
6) evalueren van gevonden patronen
7) presenteren van informatie

Ad 1) Het opschonen van de gegevens houdt in het consistent maken van de gegevens. Vaak zijn er in een database verschillende velden die dezelfde gegevens zouden moeten bevatten. Door tikfouten of problemen bij het samenvoegen van verschillende databases klopt dat niet altijd. Ook wordt dan gekeken of de gegevens wel

[1] Han & Kamber (2001), Datamining concepts and techniques, Academic Press

geschikt is om datamining op uit te voeren, dan wordt gekeken naar de kwaliteit van de gegevens.

Ad 2) Bij de fase van de integratie van gegevens wordt een andere vraag gesteld. Bevatten de gegevens wel het antwoord op de vraag die gesteld is? Moeten externe gegevens aan de huidige gegevens toegevoegd worden? Moeten twee databases aan elkaar gekoppeld worden om de benodigde informatie te kunnen vinden?

Ad 3) Bij de selectie van de gegevens wordt gekeken welke gegevens er precies in de analyse meegenomen moeten worden. Het is mogelijk dat een huisnummer van grote invloed is op de uitkomsten. Hiermee wordt voorkomen dat het datamining proces ongewenste informatie oplevert. De software is gemaakt om veel gegevens te analyseren maar het blijft zo dat "garbage in = garbage out".

Ad 4) De transformatie van de gegevens. Sommige methoden, bijvoorbeeld neurale netwerken, werken met de fout tussen de gewenste waarde en de voorspelde waarde. Als een fout in absolute zin groot, maar in relatieve zin klein is, kan het proces negatief beïnvloed worden en gevolgen hebben voor de kwaliteit van de uitkomsten. Daarom transformeert men de gegevens zodat aan elke variabele een bepaald belang toegekend wordt.

Ad 5) Het toepassen van de verschillende datamining methoden. Deze methoden worden door de computer uitgevoerd. Het is belangrijk te begrijpen hoe het algoritme aan de uitkomst komt. Op deze methoden wordt hierna uitgebreider ingegaan.

Ad 6) Daarna komt de stap die de gevonden patronen en resultaten moet evalueren. Zijn er interessante gegevens gevonden en zijn de gevonden gegevens dan ook zinnig? Komt er alleen maar triviale informatie uit de voorschijn? Dit is misschien wel de moeilijkste stap in het proces. Vaak is het lastig om te zien hoe de verbanden gevonden zijn dan is de vraag wat de betekenis is van de gevonden verbanden. Kennis van en inzicht van de verschillende datamining methoden en van de gegevens is bij deze stap een must. Vaak wordt deze stap gedaan in combinatie met de klant.

Ad 7) Tot slot moet de gevonden informatie nog zo opgeschreven worden dat de personen die de beslissingen moeten nemen de resultaten begrijpen. De gevonden informatie moet wel in duidelijke taal opgeschreven worden. Een statement als "X1 leidt tot X4" is natuurlijk niet te begrijpen, "roken leidt tot hart- en vaatziekten" wel.

Om dit proces goed uit te kunnen voeren wordt veel gevraagd van de creativiteit van de gebruiker. De gebruiker moet echt 'spelen' met de gegevens om de gewenste informatie uit de gegevens te halen. Dit hele proces moet doorlopen worden om tot een goed resultaat te komen. Datamining is geen simpel proces, datamining vereist meer kennis dan waaraan de meeste mensen bij het horen van de kreet datamining aan denken, zoals blijkt uit onderstaande citaten[1]:

'DE VOORBEREIDINGSFASE IS HET MOEILIJKSTE GEDEELTE VOOR MENSEN DIE NOG GEEN OF WEINIG ERVARING MET DATAMINING HEBBEN'

'DE KEUZE VAN DE DATA DIE BIJ DE ANALYSE WORDT BETROKKEN, HEEFT EEN DUIDELIJKE INVLOED OP WELK MODEL GEVONDEN KÁN WORDEN'

'HET DATAMININGPROCES IS NIET PER SE AFGESLOTEN, WANNEER HET MININGRESULTAAT IS BEREIKT'

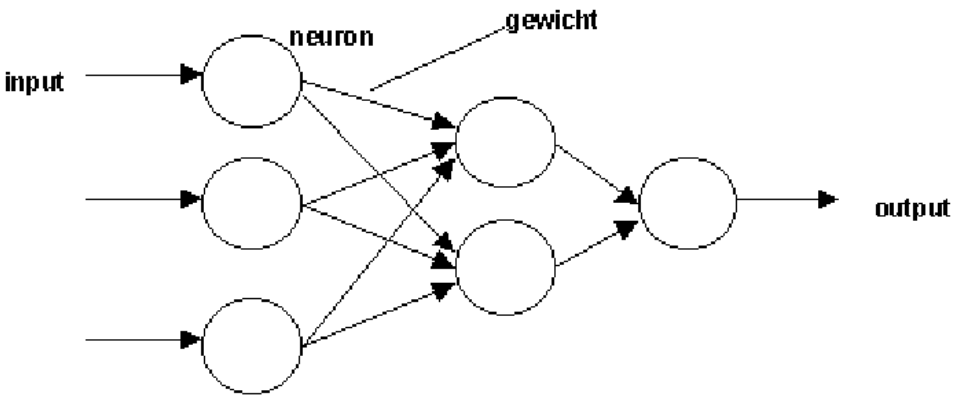
[1] http://www.beyondmagazine.nl, zie bijlage X

3.2 Datamining methoden[1]

Datamining maakt gebruik van verschillende recent ontwikkelde methoden. In dit hoofdstuk wil ik de meest gebruikte methoden nader toelichten. Voor andere methoden en overwegingen die in het datamining proces een rol spelen verwijs ik naar de literatuur.
In de praktijk worden vaak meerdere methoden op de gegevens toegepast en dan wordt de methode met de beste resultaten gekozen. Het is namelijk lastig van te voren te voorspellen welke methode het beste resultaat zal geven. Het beste resultaat zal in een elk geval verschillend zijn. Soms is het de methode met het beste scoringspercentage, soms is het de methode die het duidelijkst inzicht geeft in de vraag die gesteld is. Maar altijd zal het resultaat antwoord moeten geven op de vraag die door de klant gesteld is.

3.2.1 *Neurale netwerken*

Neurale netwerken, een van de populairste datamining methoden, proberen de werking van het menselijke brein na te bootsen. Een netwerk wordt getraind door de gegevens in een netwerk in te voeren. Van de historische gegevens kent men de resultaten. De input en de uitkomst zijn bekend. Bij elke input voorspelt het netwerk ook een uitkomst. Als het netwerk de uitkomst niet goed voorspelt, wordt dit netwerk op een bepaalde manier aangepast zodat een volgend geval wel goed geclassificeerd wordt. De training gegevens worden een aantal keer aan het netwerk aangeboden, totdat het netwerk de gewenste uitkomsten van de training set voldoende goed voorspeldt (voldoende is meestal dat 80-90% van de waarnemingen correct voorspelt wordt).
Het trainen van het netwerk is de eerste stap. Daarna controleert men het netwerk door middel van een testset. Te kort trainen leidt tot een netwerk dat zijn taak niet goed genoeg uitvoert. Te lang trainen leidt tot een netwerk dat teveel gericht is op de specifieke training gegevens. Daarom moet er altijd een test uitgevoerd worden op bekende gegevens die het netwerk nog niet eerder gezien heeft. Als het netwerk deze gegevens ook genoeg goed classificeert, mag aangenomen worden dat het netwerk ook nieuwe onbekende gegevens goed zal classificeren. Dan kan het netwerk in de praktijk toegepast worden.



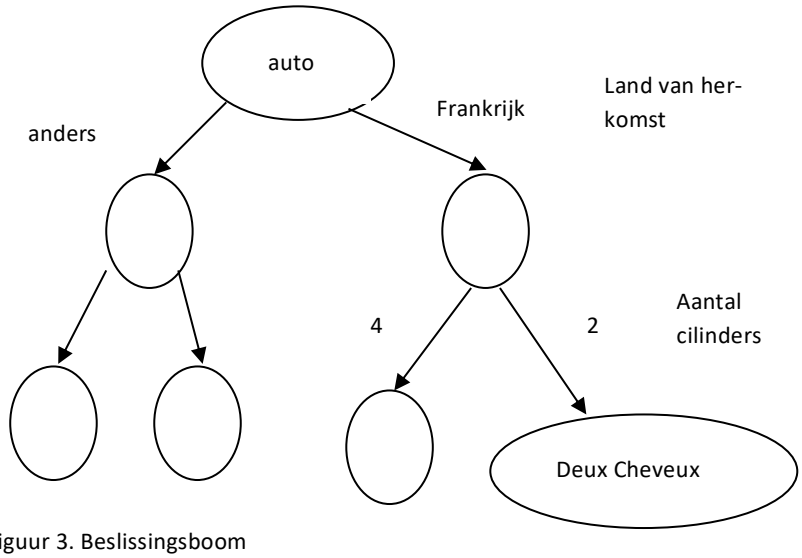
Figuur 2. Neuraal netwerk

[1] Hoeksema, E. (1999), Datamining klaar voor de massa?,
http://**www.ez.nl**/home.asp?page=/technieuws/technws9/tn9909/9909vs.htm

Een neuraal netwerk richt zich met name op het herkennen van patronen in de gegevens. Op basis van de gevonden patronen zal het nieuwe gevallen van dezelfde gegevens goed kunnen voorspellen.
De sterke punten van Neurale Netwerken zijn dat je er heel complexe problemen mee kunt oplossen, omdat het werkt met non-lineaire functies (de berekening in het neuron). Ook kan het voor veel verschillende problemen worden toegepast, gericht en ongericht, voor taken als classificatie en voorspelling en voor categorische en continue variabelen. Het nadeel is echter dat het weinig inzicht geeft, in hoe het resultaat is behaald. Belangrijker is dát het zo is, dan te weten waarom het zo is.
Een toepassing van een neuraal netwerk is het indelen van individuen in verschillende klassen. Hebben mensen met bepaalde kenmerken een grotere kans om een bepaalde ziekte te krijgen? Welke individuen uit de totale populatie schrijf je aan om mee te doen aan een preventief onderzoek? De uitkomst kan een lijst met namen zijn van mensen die aangeschreven moeten worden.

3.2.2 *Beslissingsbomen*

Een methode die juist wel een verklaring geeft voor de resultaten is de methode die gebruik maakt van beslissingsbomen. Beslissingsbomen zijn makkelijk af te lezen, omdat ze regels representeren als: ALS land = Frankrijk EN #cilinders = 2 DAN is het een grote kans dat auto = Citroën 2CV. Vanaf het begin van de boom wordt bij elke splitsing het op dat moment meest onderscheidende attribuut bepaald, net zolang tot alle attributen zijn gebruikt of totdat de opsplitsing in verschillende groepen geen toegevoegde waarde meer heeft. Een record is dan te classificeren door het pad in de boom van boven naar beneden te doorlopen.



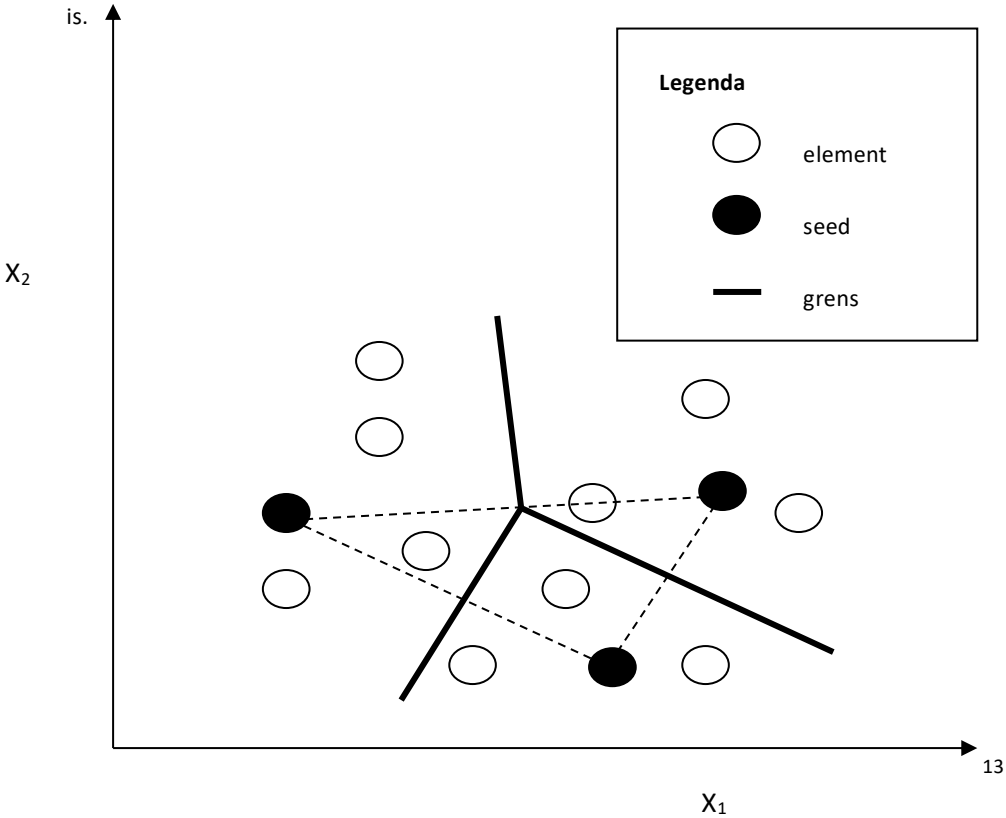
Figuur 3. Beslissingsboom

Net als Neurale Netwerken kunnen beslissingbomen continue en categorische variabelen aan. Verder is de methode geschikt voor classificatie én voorspelling. De benodigde rekenkracht gering en geeft ze een duidelijk inzicht in de manier waarop de resultaten zijn bereikt (welke attributen zijn de belangrijkste).
Een voorbeeld is het herkennen van succesvolle ondernemingen. Welke kenmerken maken dat een onderneming het goed doet? Door op bepaalde kenmerken in de boom een andere tak in te gaan kan inzicht verkregen worden in het succes van startende ondernemingen. Door de gegevens met behulp van een beslissingsboom te analyseren kan meer inzicht verkregen worden in het succesvol zijn van een enkele onderneming in plaats van het geheel.

3.2.3 *Clustering*

Clustering is een vorm van ongerichte datamining. Aan de hand van de overeenkomsten tussen records worden deze ingedeeld in clusters. Deze clusters maken het mogelijk groepen van klanten te onderscheiden. Het voordeel van deze methode is dat het een makkelijk toe te passen methode is, die met verschillende vormen van gegevens kan werken. Het kan alleen soms moeilijk zijn om de gevonden resultaten te interpreteren. Ook is de uitkomst erg afhankelijk van de functie die wordt gekozen om de gelijkheid te berekenen. Het wordt dan ook vaak gebruikt in combinatie met andere methoden, zoals beslissingsbomen, om een verklaring te vinden voor een cluster.

Clustering werkt door de afstanden, bepaald door de hoek tussen vectoren, tussen de elementen (records) te meten. Eerst worden een paar elementen, zogenaamde zaden die het aantal clusters weergeven, in een dimensie afgebeeld. De andere elementen worden vervolgens ingedeeld bij de cluster dat de kleinste onderlinge afstand heeft. Hierna worden de clusters afgebakend door lijnen te trekken die haaks staan op de rechte lijn tussen twee zaden. Vervolgens worden de gemiddelden van elk cluster berekend, worden alle elementen opnieuw ingedeeld en worden de nieuwe grenzen van de clusters bepaald. Als de grenzen niet meer veranderen stopt het proces. In het onderstaande voorbeeld zijn er twee dimensies waarop wordt geclusterd. Als voorbeeld kan het indelen van klanten in goede en slechte klanten genomen worden. Op de assen staan dan het aantal keren te laat betalen en de huidige schuldpositie. Na invoering van alle gegevens en na aangegeven te hebben welke 4 klanten de initiële seeds zijn, komt er voor elke klant uit of deze een goede of slechte klant is.



Figuur 4. Clusteren

Een andere manier om deze methode toe te passen is elke keer een nieuw geval toe te voegen en zo de gegevens op basis waarvan de beslissingen genomen worden steeds aan te passen. Een toepassing van een dergelijk nearest neighbour netwerk is het al dan niet accepteren van te laat betalen door klanten. Op basis van gegevens van een klant mag de betalingsdatum overschreden worden. Bij andere klanten wordt meteen een aanmaning gestuurd. Op de assen zou dan bijvoorbeeld het aantal keren te laat betalen, de hoogte van het te innen bedrag en de relatie met de klant kunnen staan. De gegevens worden daarna natuurlijk aangepast. Een tweede keer wordt de beslissing ten aanzien van dezelfde klant op basis van nieuwe gegevens genomen.

3.2.4 *Market Basket analysis*

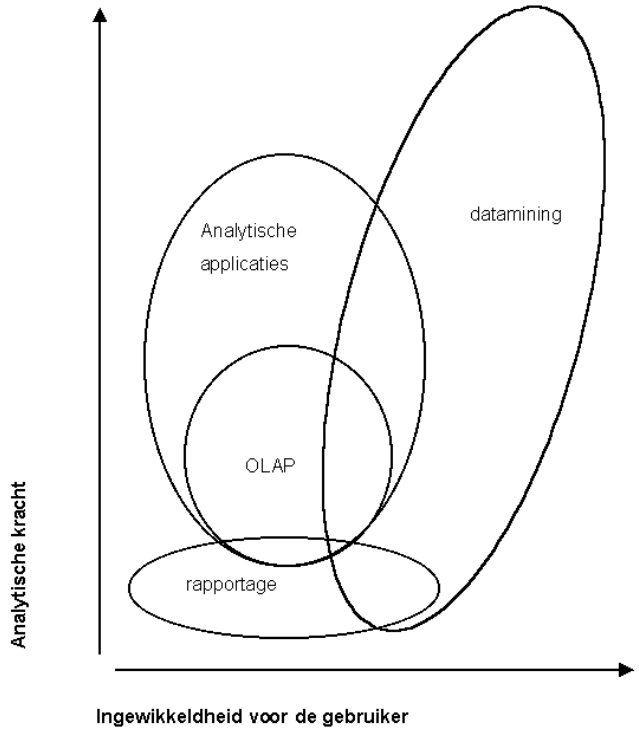
Een andere methode die ook vaak als startpunt wordt gebruikt is Market Basket analysis, zeg maar 'boodschappen' analyse. De methode gaat na welke artikelen in combinatie met elkaar worden gekocht, zodat het inzicht oplevert in koopgedrag en betere aanbiedingen kunnen worden gedaan. Eerst wordt er een multidimensionale matrix opgesteld van producten die met elkaar gekocht worden. Er worden aan de hand van deze matrix regels gemaakt, bijvoorbeeld ALS A en B dan C. Vervolgens wordt van elke regel de ondersteuning (percentage van de transacties waarin die producten in combinatie met elkaar voorkomen) en de betrouwbaarheid (als het ene product wordt gekocht, wat is dan de kans dat het andere product wordt gekocht uit de regel) uitgerekend en worden de kansen geëvalueerd. Een voorbeeld van een regel die hiermee werd ontdekt is dat luiers en bier veel met elkaar worden gekocht, blijkbaar door mannen die er door hun vrouw op worden uitgestuurd om luiers te halen.
De voordelen van deze methode zijn dat het met variabele lengtes van records kan werken, dat het duidelijke resultaten oplevert en dat de berekeningen vrij simpel zijn. De methode wordt vooral gebruikt door grote supermarkten en 'drugstores'. Deze methode kan doordat transacties steeds minder anoniem zijn, door de bonus en credit cards, nuttige informatie opleveren. Ook internet bezoek kan aanleiding geven voor een dergelijke analyse.

figuur 5 Market basket Analysis

3.3 Datamining vs. statistiek[1]

Op het eerste gezicht lijken datamining en statistiek misschien op elkaar, beide methoden vinden relevante informatie uit gegevens. Als men zich er wat dieper in op de verschillende methoden in gaat blijkt dat is niet het geval is.
Het grote verschil tussen datamining en statistiek is wel dat datamining gebruik maakt van niet-lineaire technieken. Juist door gebruik te maken van technieken die niet uitgaan van het gemiddelde en variantie, kunnen er nieuwe verassende inzichten gevonden worden.
Ten tweede kan datamining supervised en un-supervised gegevens analyseren. In het eerste geval wordt heel gericht naar bepaalde informatie gezocht. In het geval van un-supervised learning laat men het programma, de software zelf naar interessante patronen en informatie zoeken. Het doel is om met deze patronen 'het werk' beter te kunnen doen. Bijvoorbeeld om een productaanbod beter op een bepaald klantprofiel af te stemmen. Zo kan men beter in schatten welke klant in wat voor soort product interesse heeft. Datamining speurt vanuit alle aanwezige gegevens in databases naar (verborgen) patronen.
Ten derde kan datamining uitspraken doen over het (toekomstige) gedrag van individuele personen. Op basis van de kenmerken van een individu wordt een voorspelling over dat individu gedaan en niet over de groep in het algemeen.
Ten vierde is datamining beter in het behandelen van grote bestanden. Onder grote bestanden wordt verstaan dat de gegevens veel variabelen bevatten en minimaal 10 keer zoveel waarnemingen.
Tot slot kan datamining iets beter omgaan met missing values dan analytische methoden.
Een voorbeeld om een verschil tussen datamining en statistiek aan te tonen.
Bij datamining zou de vraag kunnen zijn: welke factoren zijn bepalend voor de hoogte van salarissen? Alle gegevens worden dan verzameld, zoals werkervaring, leeftijd, opleiding, etc. Vervolgens wordt gekeken welke gegevens het salaris beïnvloedt. Daaruit kan blijken dat tot 25 jaar de leeftijd van invloed is, vanaf 25 jaar de werkervaring.
Statistische methoden zijn meer theoretisch en toetsend van aard. Veronderstelt men een zeker verband, dan toetst men en kijkt men of dat verband wel of niet bestaat. Bijvoorbeeld een verband tussen leeftijd en salaris: men zoekt de benodigde gegevens erbij, om die daarna te toetsen. Dan volgt een representatieve steekproef voor de gehele populatie, en de uitkomsten van de analyses generaliseert men voor de gehele populatie. Statistiek gaat dus vanuit een bewust - op basis van een veronderstelling - geselecteerd deel van gegevens naar algemeen geldende regels. Datamining vindt zo ook verbanden waar de onderzoeker misschien zelf niet direct aan had gedacht. Datamining vindt 'zelf' op basis van de gegevens de informatie. Bij statistiek moet de onderzoeker zelf de verbanden veronderstellen en dan kan statistiek aantonen of dit verband al dan niet bestaat.
Datamining is wezenlijk anders dan statistiek en beide methoden kunnen elkaar niet vervangen.

[1] Hulzebos, G, Statistiek en datamining: verschillen en overeenkomsten, http://www.spss-mining.nl/SPSS/Nieuws/no05_statistiek.htm



figuur 6 vergelijking verschillende analytische technieken

4 EIM

4.1 De EIM Groep

In 2001 is uit de EIM stichting de EIM groep ontstaan. De EIM groep bestaat uit een zestal afzonderlijke BV's,

- EIM Consult BV
- EIM Stratus BV
- EIM Survey BV
- EIM BV
- IOO BV
- EAYS BV

De EIM Groep BV is een houdstermaatschappij (holding in de vorm van een BV) die wordt bestuurd door een holdingdirectie. De EIM groep is een overkoepelend orgaan dat onder andere toezicht houdt op en verantwoordelijk is voor de afstemming tussen de verschillende BV's.
De missie van de EIM groep:[1]
"Wij creëren kennis over het bedrijfsleven en bedrijfsprocessen ten behoeve van beslissingen in bedrijf en beleid. Wij doen dit met behulp van onze unieke kennisstructuur. Wij zijn het MKB-kenniscentrum in Europa."

4.2 EIM BV

EIM is een toonaangevende speler op de beleidsonderzoekmarkt. Het is de grootste BV binnen de EIM groep met ongeveer 90 medewerkers. Het bedrijf bestaat sinds 1930 en heeft zich als zodanig een vaste positie in de beleidsonderzoekmarkt verworven. Grote klanten van EIM zijn verschillende ministeries (met name het ministerie van economische zaken) en verschillende overkoepelende brancheorganisaties. EIM is gespecialiseerd in onderzoek over en advies voor het midden- en kleinbedrijf. EIM is ambitieus als het gaat om het vergroten van de omzet en marktposities. Dit blijkt uit het jaarverslag 1999 van EIM "een terugblik vanuit de toekomst":
"EIM groeit in de eerste jaren van de 21ste eeuw uit tot een volwaardige commerciële onderneming die haar sporen heeft verdiend met een breed scala aan onderzoeks- en adviesactiviteiten. De link met de denker maakt duidelijk dat EIM goed onderbouwd te werk wil blijven gaan. De onderzoeksorganisatie die de 21ste eeuw betrad kan met de volgende termen worden getypeerd: human capital en ICT, beleids- en marktonderzoek, contract research, programma MKB en ondernemerschap, internationaal consultancy en massamaatwerk."

De missie van EIM BV is:
"Wij creëren kennis over het bedrijfsleven en bedrijfsprocessen ten behoeve van beleidsbeslissingen. Wij doen dit met behulp van onze unieke kennisstructuur. Wij zijn het MKB-kenniscentrum in Europa."[2]

[1] Strategienota EIM 1999-2000+

[2] Strategie nota 2002-2006 "We gaan voor allround beleidsonderzoek"

4.3 Unit DataWarehousing

DataWarehousing is een unit (afdeling) binnen EIM. DataWarehousing heeft bijzondere onderzoekstaken op het gebied van data en beleid. De oorspronkelijke gegevens moet geanalyseerd en opgeschoond worden om deze voor onderzoek te kunnen gebruiken. Dit proces is van groot belang voor het onderzoek dat EIM doet.
Onder de unit DataWarehousing is het account data en informatiesystemen gepositioneerd. Met de invoering van het account data en informatiesystemen heeft men duidelijk gemaakt dat de unit DataWarehousing ook een positie op de markt heeft. Het account heeft tot doel het binnenhalen van opdrachten op het gebied van data en informatiesystemen. Data en informatiesystemen is een bloeiend account binnen EIM. De diensten die ze aanbieden zijn het "clearinghouse" concept, het maken van informatiesystemen en het ontsluiten van gegevens via het internet.
Binnen het account ziet men mogelijkheden op het gebied van datamining, omdat deze activiteit zeer goed aansluit bij de competenties van de unit. Deze interesse blijkt uit het accountplan "Data voor beleid" dat in 2001 is goedgekeurd door de directie van EIM. Onderstaand fragment komt uit dit accountplan en vormt de aanleiding voor dit investeringsplan.

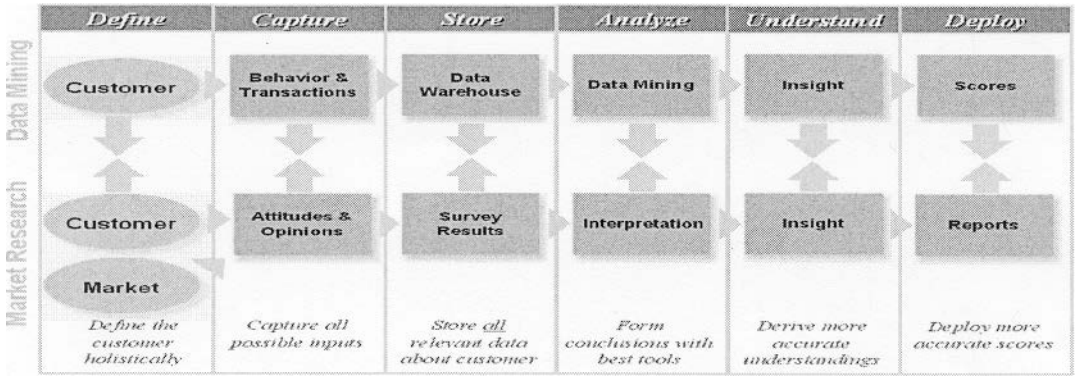
Daarnaast ontstaat er een terrein waarbij vanuit transactiegeoriënteerde data systemen allerlei primaire informatie over koopgedrag via complexe statistische analyse systemen gebruikt wordt voor specifiek CRM (customer relationship management). Hierbij is vaak sprake van de volautomatische koppeling van gebruiks- en verkoopgegevens aan klantgegevens en de opslag daarvan in zeer grote databanken. Deze systemen worden bijvoorbeeld aangetroffen bij dienstverlenende bedrijven met een hoog ICT-gehalte zoals internet service providers, elektronische media providers, waar onder Pay-per-view en betaal-TV, maar ook bij winkelketens, franchise en inkoopverenigingen (loyalty programma's zoals clubcard, klantenpas of Airmiles).

Kenmerkend voor de statistische analyse van deze omvangrijke datastroom is de automatische generatie van gebruikers- en klantenprofielen van allerlei aard op basis van datamining met behulp van neurale netwerk concepten. De technische implementatie van dergelijke systemen zal vooralsnog in handen blijven van gerenommeerde systeemhuizen en softwareleveranciers, maar bedrijven die behoefte hebben aan dergelijke systemen hebben meer en meer ook de inhoudelijke kennis nodig, die nodig is om tot bruikbare resultaten te komen. Hierbij ligt het accent met name op kennis van economische processen en kennis van statistische methoden en technieken, en minder op kennis van databases en datawarehousing.

4.4 Datamining en EIM

In alle voorafgaande hoofdstukken staat het adviseren van het nemen van beslissingen centraal. Men wil een allround beleidsonderzoeksbureau zijn en het MKB kenniscentrum van Europa. Datamining is een manier om kennis te verkrijgen uit gegevens. Sluit er een activiteit beter aan bij de activiteiten van EIM dan datamining? Dat er beleidsonderzoek gedaan kan worden door middel van datamining blijkt wel uit het artikel De mogelijkheden van *da-*

tamining voor de Nederlandse politie[1]. Daarnaast heeft datamining veel weg van onderzoek doen zoals blijkt uit onderstaande figuur.



figuur 7. vergelijking onderzoekprocessen

Om een gedeelte van de definitie van datamining te herhalen: Datamining is het proces van het ontdekken van interessante informatie uit grote hoeveelheden gegevens. Deze activiteit sluit aan bij de doelstellingen van EIM, kennis vergaren over en voor beslissingen. Het is een nieuwe manier van "kennis voor beleid".

4.4.1 *Datamining Services*

Datamining services is een naam van een nieuw onderdeel van DataWarehousing, dat alle projecten met betrekking tot datamining moet uitvoeren. De taken van dit onderdeel van DataWarehousing is het modelmatig analyseren van gegevens met behulp van datamining. Dit onderdeel zal het complete datamining proces moeten beheersen en voor de klanten elke dataset binnen 10 werkdagen in informatie om moeten kunnen zetten. "Datamining Services adviseert bedrijven en instellingen bij het nemen van beleidsbeslissingen door gebruik te maken van datamining" zo zou de missie van Datamining Services kunnen luiden.
Datamining vereist kennis op het gebied van datamining technieken en gegevens, een zekere mate van creativiteit en goed statistisch inzicht.
Een service op dit gebied lijkt voor EIM een logische ontwikkeling op weg naar een allround beleidsonderzoekbureau. Bij de klanten van EIM bestaat weinig ervaring op het gebied van datamining. Daarnaast ontbreekt bij klanten de tijd en kennis om datamining goed te leren en te beheersen. DataWarehousing heeft de mogelijkheden om de kennis en de techniek in huis te halen om de bedrijven van advies te kunnen dienen. Ook ligt onderzoek doen op basis van datamining dicht bij de kernactiviteit van EIM, dat blijkt uit de verschillende processtappen die in hoofdstuk 3.1 beschreven zijn.

[1] Lopez, M.J.J., De mogelijkheden van *datamining* voor de Nederlandse politie,
Het Tijdschrift voor de Politie, nr. 6, juni 2000, pag. 26-29, zie bijlage V

5 Markt

5.1 Inleiding

Slechts zeven procent van de aanwezige data wordt omgezet in informatie[1].

"Er wordt tot nu toe onvoldoende gebruikt gemaakt van aanwezige data in een bedrijf. Veel bedrijven beseffen nog steeds niet wat er allemaal mogelijk is met alle gegevens die her en der worden opgeslagen. In theorie kunnen grote efficiencyvoordelen bereikt worden. In de praktijk willen de efficiencyvoordelen nog wel eens ondergeschikt zijn aan de effectiviteit-voordelen: misschien is het nog wel het belangrijkste dat 'de gebruikers' meer plezier in het werk krijgen."[2]

In de toekomst zullen meer beslissingen genomen worden op basis van gegevens[3]. Vele soorten gegevens en vooral veel gegevens. Alle mensen met wie ik over datamining en de bijbehorende markt gesproken heb maken een aantal opmerkingen.

- Datamining is een techniek die mensen lastig zullen begrijpen. Er is hoogwaardige technische, economische en wiskundige kennis en creativiteit vereist. Goede datamining is echt een specialisme.
- Datamining lijkt steeds populairder te worden bij het bedrijfsleven. Ten grondslag hieraan ligt het succes van datamining bij grotere bedrijven overal in de wereld.
- De behoefte aan datamining groeit nog steeds sterk. Hiervoor zijn de volgende redenen aan te voeren[4]:
  - Er zijn meer informatiebronnen beschikbaar dan ooit tevoren, en de hoeveelheid aan informatie groeit exponentieel;
  - De opslagmogelijkheden worden steeds groter en goedkoper, daardoor meer gegevens opgeslagen;
  - De datamining programma's worden steeds krachtiger en makkelijker te gebruiken;
  - Kennis over klanten wordt steeds belangrijker in een competitieve markt.

Voor het toepassen van datamining zijn er verschillende mogelijkheden. Ten eerste op het gebied van customer relationship management. De klant is in beweging. Het voorspellen van het gedrag van de klant wordt steeds belangrijker. Ten tweede kan er door middel van datamining ook onderzoek naar het gevoerde beleid gedaan worden. Hoe reageren ondernemingen en burgers (de klanten van de overheid) op bepaalde maatregelen. Ten derde is er analyse mogelijk van allerlei gegevens die in het verleden verzameld en opgeslagen is. Datamining biedt de mogelijkheid om deze gegevens te vertalen in beleidsinformatie.

5.1.1 *Twee voorbeelden:*

Voorbeeld 1

[1] volgens BNT-Holland

[2] B. Heida, Blauw research

[3] Eiben, A.E. Het gebruik van data voor beslissingsondersteuning, zie bijlage XII

[4] Elliot K. Scionti, R. Page, M. (2003) Two Rivers, The confluence of data mining and market research for smarter CRM, zie bijlage VI

Een voorbeeld dat tijdens mijn stage op tafel kwam in gesprek met Peter Brouwer, accountmanager Arbeid en Sociale Zekerheid, het onderzoeken van het HRM beleid bij verschillende bedrijven.
Bij verschillende bedrijven moet de effectiviteit van waardering van het HRM beleid bepaald worden. Er worden drie groepen mensen gevraagd om mee te werken aan het onderzoek, de bedrijfsleiding de HRM medewerkers en het personeel. Zo krijgt men een groot aantal gegevens vanuit verschillende hoeken. Door deze bestanden te koppelen en te analyseren krijgt men een ander inzicht in de verschillende maatregelen bij bedrijven en hun gevolgen. Door deze manier van onderzoeken kunnen nieuwe inzichten verkregen worden.

Voorbeeld 2
Een bedrijf wil een mailing naar zijn klanten versturen. Datamining Services helpt het bedrijf om klanten te selecteren die een grote kans hebben om op de mailing te reageren. Het bedrijf levert zijn klantengegevens en op basis van die gegevens worden klanten aan wie de mailing verzonden moet worden geselecteerd. Het bedrijf bespaart zo duizenden euro's aan promotiekosten.

5.2 Klanten

In de gesprekken die ik met verschillende mensen gevoerd heb, bleken genoeg ideeën te zijn over het toepassen van datamining.[1] Men is het er over eens dat er mogelijkheden op het gebied van Datamining Services zijn bij bedrijven en instellingen. Het blijven echter ideeën, nog geen concrete mogelijkheden. In de markt bestaat er nog geen vraag naar datamining toepassingen. Men vermoedt dat de bedrijven op middellange termijn interesse krijgen in het nut van datamining. Op dit moment zijn voor de bedrijven de mogelijkheden van datamining te onbekend en het te behalen voordeel te onzeker. Ondernemers hebben last van koudwatervrees[2]. Er zijn veel ongebruikte gegevens opgeslagen bij de verschillende bedrijven. Maar welk bedrijf wil deze gegevens om zetten in informatie? Welke ondernemer gelooft dat daar voor hem winst te behalen is? Er zijn genoeg toepassingen te verzinnen, maar de ondernemers geloven (nog) niet in de meerwaarde van datamining. Zeker niet na de beloftes die de afgelopen jaren op IT gebied gemaakt zijn.
Om Datamining Services op te zetten zal men de ondernemers moeten overtuigen dat er meerwaarde zit in het analyseren van hun gegevens door middel van datamining. Maar deze markt zal men zelf moeten creëren, de dienst zal de markt in gepusht moeten worden. Op basis van verschillende gesprekken heb ik een analyse van mogelijkheden tot het toepassen van datamining bij verschillende klanten gemaakt. Deze (beperkte) analyse dient inzicht te geven in het marktpotentieel van Datamining Services. De (on)mogelijkheden tot het toepassen van datamining verschillen per klantengroep, daarom zijn de klanten in een aantal groepen opgedeeld.
Ten eerste de bedrijven die zelf al aan datamining doen en daar een aparte afdeling voor opgezet hebben. Op een bijeenkomst van SPSS over predictive marketing analysis waren onder meer verzekeringsmaatschappijen, de gouden gids, banken en vergelijkbare bedrijven aanwezig. Datamining is bij deze bedrijven wel een succes.
Ten tweede de bedrijven die een te kleine klantenkring hebben om datamining nuttig te laten zijn, de kleine zelfstandigen. Voor deze bedrijven zullen de kosten niet opwegen tegen de opbrengsten, bovendien zal een datamining hen niets nieuws vertellen.

[1] gesprekverslagen, zie bijlage III

[2] Van der Putten, P. Broodnodige intelligentie voor CRM, zie bijlage VII

Ten derde de beleidsinstanties. Deze instanties beschikken over het algemeen niet over de gegevens die direct met datamining geassocieerd worden. Bij het CBS zijn in de loop der jaren wel zeer veel gegevens verzameld. Door deze verschillende gegevens te koppelen ontstaan gegevens waar door middel van datamining nuttige informatie uit te halen is. Het kan tot verrassende conclusies leiden.
Ook bij MKB Nederland en bij TNO-TPD liggen veel gegevens opgeslagen volgens de accountmanagers. Het zou de kunst zijn om die gegevens boven tafel te krijgen en te kunnen onderzoeken. Wat was in het verleden het effect een bepaalde beleidsmaatregel? Zo kan het mogelijke effect van een nieuwe vergelijkbare maatregel berekend worden. Men zou het effect van bepaalde beleidsmaatregelen over verschillende jaren kunnen afleiden door middel van datamining.
Ook brancheorganisaties beschikken op dit moment niet over gegevens die geschikt zijn voor datamining. Deze organisaties hebben relatief weinig gegevens van de bij hun aangesloten bedrijven. Bedrijven helpen hun eigen concurrenten eenmaal niet graag aan informatie om het beter te doen. Maar bij brancheorganisaties liggen wel degelijk mogelijkheden voor een benchmarkonderzoek door middel van datamining. Stel ieder aangesloten bedrijf geeft (via internet) een (groot) aantal gegevens prijs. Deze worden dan anoniem bij elkaar in een groot bestand gestopt. Op dit grote bestand kan dan datamining toegepast worden, met twee resultaten: ten eerste een algemene trend als resultaat voor de brancheorganisatie en ten tweede voor ieder bedrijf een aparte analyse van hun positie ten opzichte van de concurrentie. Zo hebben de afzonderlijke bedrijven en de brancheorganisaties profijt van deze analyse door middel van datamining.
Ten vierde hebben individuele bedrijven gegevens om datamining op toe te passen alleen doen ze het nog niet. De gegevens worden opgeslagen maar er wordt niets mee gedaan.
De organisaties zijn ook onder te verdelen. Ten eerste zijn er de met de overheid samenwerkende instellingen. Volgens verschillende accountmanagers, P. Vroonhof, A. Muizer en P. Brouwer, zijn er verschillende mogelijkheden tot het doen van datamining voor bestaande klanten. Met name het CWI wordt genoemd als een mogelijke klant. (het monitoren van het baan vinden op basis van een werknemerprofiel) Nadeel hierbij is de vertrouwelijkheid van de gegevens. Het CWI zou in dit geval zelf tot aanschaf van de software over moeten gaan. EIM zou dan ter plekke assistentie kunnen verlenen. Daarnaast is ook Syntens met de innovatiemonitor een goede mogelijkheid om datamining technieken te benutten. Bij het account “sociale zekerheid en arbeid” komen offerteaanvragen binnen die met datamining beantwoord kunnen worden. Er zullen in de toekomst nog meer onderzoeken aangeboden worden.
Er zijn ook mogelijkheden bij de franchise organisaties. Bij franchise organisaties zijn ook mogelijkheden voor het toepassen van datamining. Waarom loopt een aanbieding bij het ene filiaal wel en bij het andere niet? Welke aanbiedingen moeten we bij welk filiaal doen? Op dit soort vragen kan door middel van datamining antwoord gegeven worden, bijvoorbeeld door analyse van de kassastromen.
Tot slot zijn er de individuele bedrijven. De grotere MKB-bedrijven bieden het grootste potentieel op het gebied van datamining. Ook volgens SPSS is dit een markt die behoefte heeft aan Datamining Services. Een voorbeeld van dergelijke ondernemingen zijn de bedrijven waarbij de informatiestroom op dit moment, mede onder invloed van internet, aan het veranderen is. De klant gaat niet meer langs alle bedrijven om te kijken naar de verschillende producten, of loopt een winkel binnen om zich voor te laten lichten over het aan te schaffen object. De klant kijkt vanuit de luie stoel op de internetsite en laat zich on line voorlichten. Het herkennen van patronen van die klanten op internet wordt steeds belangrijker. Bedrijven zullen steeds meer beslissingen moeten nemen op basis van internetdata. Steeds meer bedrijven krijgen portals op hun site en monitoren hun klanten op basis van hun IP nummer. De informatie uit die gegevens zal steeds belangrijker worden. In deze gevallen kan datamining een oplossing zijn om de juiste informatie te achterhalen. Waar deze ontwikkelingen

vooral veel gevolgen hebben zijn de bedrijven waarbij óf de backoffice óf de frontoffice geautomatiseerd wordt. Voorbeeld hiervan zijn de reisorganisaties. Er zal er veel veranderen in de nabije toekomst bij deze winkels. Juist in deze vechtmarkt is een goede kennis van de klanten en daarmee een goed on line visite kaartje naar de klant noodzakelijk. Ook hier kan Datamining Services uitkomst bieden en op basis van analyses kunnen beslissingen genomen worden.
Maar is dit de markt van EIM? Volgens de missie van EIM richt EIM zich vrijwel uitsluitend op beleidsinstanties. Wil men ook de markt van individuele ondernemers betreden en hoe komt men die markt binnen? Is dit niet het gebied van Stratus en Consult?
Uit gesprekken blijkt dat de grenzen tussen de verschillende bedrijven binnen de EIM groep klein zijn. Het zal geen probleem zijn voor DataWarehousing om de individuele bedrijven te benaderen. Maar het verdient zeker overweging en aanbeveling om hierin samen te werken met Stratus en Consult, zeker als het om de contacten met de individuele bedrijven en de mogelijkheden van datamining en internet gaat.
Om op deze markt te gaan opereren zullen nieuwe contacten bij bedrijven gezocht moeten worden. Het huidige netwerk biedt daar uitstekende mogelijkheden voor. Dit kan bijvoorbeeld ook via MKB-Nederland. Brancheorganisaties kunnen deze dienst aanbieden aan hun leden. Het netwerk van EIM zal nog niet direct op een datamining service door EIM zitten te wachten, de bedrijven achter die organisaties wel. EIM heeft een uniek netwerk om deze groep bedrijven te kunnen bereiken.
Kortom er zijn voldoende mogelijkheden te bedenken om datamining toe te passen, maar in de praktijk wordt er weinig mee gedaan en is er weinig animo voor om er wat mee te doen. Misschien heeft men wel genoeg aan de huidige onderzoeken of heeft men nog geen zin om geld uit te geven aan onderzoek op deze manier.
Nu is er blijkbaar nog geen behoefte aan Datamining Services in de markt, maar over een paar jaar zou dat wel eens anders kunnen zijn. Het verdient ook aanbeveling om deze ontwikkelingen scherp in de gaten te houden.

5.3 Concurrenten

Op dit moment zijn er geen bedrijven die zich puur richten op het analytische datamining advies. De mogelijke concurrenten zijn op te splitsen in twee groepen. Ten eerste de softwareleverancier die naast het verkopen van de software ook datamining advies gaat geven aan klanten. Ten tweede andere (beleids)onderzoekbureaus.

Software leveranciers

De grotere softwareleveranciers op het gebied van datamining zijn: SAS, SPSS, datadistilleries, megacomputer, BNT-Holland en IBM. De voornaamste taak van deze groep bedrijven is het maken van goede programma's. Sommige software bedrijven houden zich ook bezig met het adviseren van bedrijven en het helpen analyseren van gegevens. Deze groep ziet EIM meer als klant dan als concurrent. De software leveranciers zijn geen concurrent van EIM op het gebied van Datamining Services. Natuurlijk kunnen de software bedrijven deze service opzetten maar het analyseren van gegevens is niet hun kernactiviteit. Daarom zien bijvoorbeeld SPSS en SAS brood in deze nieuwe service van EIM. Ze hopen door middel van EIM datamining in een nieuw markt te kunnen aanboren. SPSS en SAS zouden EIM willen helpen bij het ontwikkelen van kennis op datamining gebied en dat zou nuttig zijn.

Onderzoeksinstellingen

Onderzoeksinstellingen zijn om voor de hand liggende redenen concurrenten. Bedrijven als GfK, NIPO-Consult, Nielsen, Scanmar en Sentient Machine Research en andere zijn bezig met datamining, maar meer voor intern gebruik. Ook andere onderzoeksinstellingen zouden

advies kunnen gaan geven op het gebied van datamining. Potentiële concurrenten die nog niet aan datamining doen maar het wel zouden kunnen doen zijn: Blauw research, Ecorys-Nei, Research voor Beleid en ESI. Vele bedrijven lijken wel een datamining service aan te bieden maar dat is meer het gebruik van de kreet datamining dan het echt op datamining lijkt.

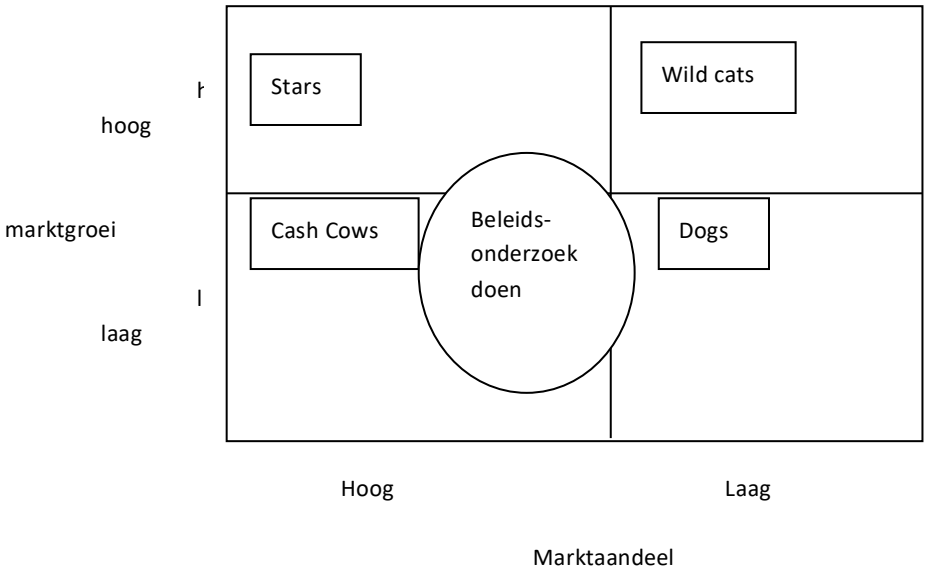
Welk onderzoeksbureau gaat als eerste de markt op met Datamining Services? Wie weet als eerste de klanten te overtuigen van het nut van datamining? De koudwatervrees bij bedrijven en de prijs van de software zijn het grootste struikelblokken op weg naar succes voor Datamining Services. Op dit moment is nog geen van de bedrijven een vorm van datamining services aan het promoten. Is iedereen op elkaar aan het wachten? Wie gaat de eerste stap zetten? Het is een kwestie van tijd totdat het eerste bedrijf datamining service zal aanbieden. Beleidsonderzoek en datamining liggen dicht bij elkaar, qua opzet en eindresultaten. Onderzoek houdt zich bezig met geaggregeerde gegevens en gegevens verkregen door middel van onderzoek. De resultaten kunnen door middel van de klassieke analyse, door middel van toetsen, berekend worden. Voor datamining is een andere kennis noodzakelijk. Kennis van neurale netwerken, beslissingsbomen en clustering netwerken en daarnaast gevoel voor gegevens, creativiteit en beleidsmatig inzicht. Ook vereist datamining een zekere mate van wiskundige kennis om de algoritmes te kunnen begrijpen. De gewenste informatie ligt echter dicht bij elkaar het is een andere weg die gezocht wordt om tot informatie te komen. Op het moment dat de markt zou ontstaan blijft het een goede stap voor EIM om zich op dit gebied te bekwamen. Op dit moment beschikt geen onderzoeksbureau over voldoende mogelijkheden om datamining services te kunnen verlenen, maar dat zal zeker veranderen.

5.4 Portfolio analyse EIM[1]

Een goede bedrijfsportfolio is in alle vier de kwadranten gevuld. Er zijn activiteiten die geld kosten (Wild cats) omdat men deze nieuwe producten en diensten in de markt wil zetten. De wild cats hebben een laag marktaandeel maar een hoge marktgroei. Dit zijn mogelijke nieuwe producten die het bedrijf in de markt wil zetten. De stars leveren geen geld op maar kosten ook geen geld meer. Ze hebben zich wel bewezen in de markt en hebben ook een zekere marktpositie veroverd. De markt groeit en het marktaandeel is hoog. De Cash Cows leveren het geld op, dat zijn producten waarin men niet meer hoeft te investeren. Ze hebben een hoog marktaandeel in een markt die aan het verzadigen is. En de Dogs kosten geen geld maar leveren ook geen geld meer op; deze producten zijn aan het eind van hun levenscyclus en worden uit het aanbod van producten gehaald. Voorbeelden van deze ontwikkelingen zijn bij veel producten te zien.
Binnen EIM heeft men eigenlijk maar 1 grote activiteit en dat is beleidsonderzoek doen. Iedere account is daar op zijn manier mee bezig en probeert opdrachten binnen te halen. Beleidsonderzoek doen is de cash cow en star van EIM. Er worden immers langdurige steeds terugkerende projecten uitgevoerd maar ook worden initiatieven ondernomen om bij nieuwe bedrijven een onderzoek te kunnen doen. Dit zorgt echter voor een grote afhankelijkheid van deze ene dienst. Mocht de markt voor dit onderzoek in elkaar klappen dan heeft EIM een probleem. Het zou nuttig zijn voor EIM om ook op andere gebieden diensten in de markt te zetten om minder afhankelijk te zijn van die ene service. Natuurlijk moet de nieuwe dienstverlening wel passen bij de kernactiviteit van EIM, beleidsonderzoek doen.

Op dit moment ziet de portfoliomatrix van EIM er zo uit:

[1] Keuning en Eppink; 2000, Organisatie en leiding; pag. 491



figuur 8: portfoliomatrix EIM BV

6 Voorstel

6.1 Voorstel

"EIM BV investeert 30 mandagen (2 x 15) plus de kosten van een datamining cursus (2 x € 2000) en de kosten voor de aanschaf van de datamining software programma, met als doel een datamining service op te zetten"

6.2 Product Service Kennis

6.2.1 *EIM datamining services*

EIM Datamining Services wil zich in de markt positioneren als een expert op het gebied van datamining. Dit betekent dat de dienst zich bezighoudt met het ontdekken van informatie uit gegevens. Datamining Services levert een complete dienstverlening voor klanten. U levert ons uw gegevens, wij leveren u de gegevens met informatie en advies terug.
De kennis zal op het gebied van datamining opgebouwd moeten worden. Kennis van het totale traject van datamining en kennis van de algoritmes, zoals neurale netwerken en beslissingsbomen in het bijzonder.
Het is aan te bevelen om datamining via een cursus te leren. PATO, een instituut voor post academisch technisch onderwijs, organiseert een goede cursus[1]. Daar worden in 5 dagen de basis principes van datamining uitgelegd en de algoritmes behandeld. De veronderstelde voorkennis bij de deelnemers aan de cursus: *academisch/HBO niveau óf een door ervaring verkregen gelijkwaardig kennisniveau. Er wordt uitgegaan van basiskennis van de wiskunde en statistiek.* Deze kennis zou eventueel ook van het internet gehaald kunnen worden of geleerd door het bestuderen van literatuur. Daarnaast bestaat er de mogelijkheid om het softwareprogramma te leren door middel van een cursus. De cursus is niet gericht op de achtergrond van de verschillende algoritmes. Een cursus kost ongeveer € 2000. Welke cursus er gekozen wordt hangt af van de medewerker in kwestie.
Twee medewerkers volgen een cursus en worden opgeleid tot datamining expert. Zij kunnen elkaar dan ondersteunen en elkaar vervangen bij vakantie en ziekte. Om gelijk 3 mensen in te zetten lijkt overbodig, aangezien dat hoge extra kosten met zich meebrengt. Bovendien kunnen nieuwe geïnteresseerden de kunst van datamining intern onder de knie krijgen. Deze medewerkers zullen creatief met gegevens om moeten kunnen gaan en op basis van goede analyses duidelijke keuzes met betrekking tot de gegevens moeten maken.

6.2.2 *Verwachtingen*

Datamining bewijst zich steeds vaker in de praktijk als een bruikbaar IT instrument met aantoonbare strategische waarde. De technieken en algoritmen achter datamining vormen niet het meest complexe onderdeel. De crux ligt bij het op- en doorstarten van het datamining proces, om stap voor stap te komen tot een efficiënte, kennisverwervende en dus lerende organisatie. Zodat u als ondernemer kan doen wat het meest telt vandaag de dag: klanten en concurrenten een stap voor zijn.[2]

[1] technieken voor datamining, zie bijlage XI

[2] Van der Putten, P. broodnodige intelligentie voor CRM, zie bijlage VII

De laatste jaren is er een soort van hype op het gebied van datamining ontstaan. Door slechte kennis van de techniek bleven de verwachte resultaten uit. Nu die hype voorbij is, is het tijd om aan te tonen wat datamining werkelijk kan betekenen. Men zal het gedeeltelijk geschonden vertrouwen van de markt in deze techniek moeten heroveren. De mogelijkheden van datamining blijven namelijk onverminderd groot. De moeilijkheid bij datamining ligt zoals uit het citaat blijkt niet zozeer in het toepassen van datamining technieken; het lastige van datamining is dat het hele proces zorgvuldig uitgevoerd moet worden. Dat is niet eenvoudig. Bedrijven hebben hulp nodig om tot een juiste oplossing te komen. Er zal in de toekomst zeker een markt op het gebied van Datamining Services gaan ontstaan.

6.3 Stappenplan

Mocht EIM Datamining Services willen opzetten dan is een tijdschema opgenomen. Dit schema gaat uit van een start in september 2003. Een aantal stappen zal op een ander tijdstip niet meer mogelijk zijn. De bijeenkomst van SPSS is waarschijnlijk eenmalig en men zal moeten onderzoeken wanneer men op cursus kan. Bij een software bedrijf zal dat makkelijker te regelen zijn dan bij een instelling voor post academisch technisch onderwijs.

Stappenplan opzetten Datamining Services	
september 2003	Spreken op bijeenkomst over datamining voor ondernemers van SPSS Voorlichting accountmanagers
november 2003	Promotiefolder of Cd-rom Datamining Services moet klaar zijn
februari 2004	Medewerkers op cursus
	Medewerkers oefenen met datamining op minimaal twee test datasets
april 2004	Datamining Services is operationeel. Het eerste datamining project wordt gestart (eventueel met ondersteuning door software leverancier)
januari 2005	Evaluatie project
januari 2006	Evaluatie project
januari 2007	Evaluatie project

In september 2003, begint EIM met het ontwikkelen van Datamining Services. Men zal moeten beginnen met het ontwikkelen van promotiemateriaal om de nieuwe service zowel intern bij de andere accountmanagers als extern onder de aandacht te brengen. In dit promotiemateriaal zal duidelijk moeten staan wat datamining is en op welke gegevens deze toegepast kan worden. Deze folder moet natuurlijk ook op de website van EIM komen te staan. Daarnaast moet ervoor gezorgd worden dat EIM op verschillende andere websites bekend is. Bijvoorbeeld bij startkabel.datamining.nl, SPSS en anderen. Ook zal gekeken moeten worden op welke plek EIM in de Google zoeklijst komt te staan. Dit is de huidige standaard van de zoekmachines. Als iemand iets op internet wil zoeken gebruikt men Google (www.google.nl). Daarnaast zal men op de datamining bijeenkomst voor ondernemers georganiseerd door SPSS moeten spreken.
Op basis van de opgedane informatie en de drukte op de genoemde bijeenkomst ligt er in november 2003 een beslissingsmoment. Gaat EIM door met Datamining Services of niet en zo ja welke software gaat men gebruiken. Mocht dit antwoord positief zijn dan komt de volgende fase.
In de tweede fase, april 2004, zal het ontwikkelde promotiemateriaal via een mailing aan diverse potentiële klanten gestuurd moeten worden. Daarnaast zullen alle accountmanagers op de hoogte van deze nieuwe service gebracht moeten zijn en zal ook via die kanalen de nieuwe service gepromoot moeten worden.
Om de kracht en mogelijkheden van Datamining Services mogelijk te maken en zou met een klant in 2003 een onderzoek op basis van datamining gedaan moeten worden. De beste mogelijkheid daarvoor lijkt me de een klein maar wel strategisch project, bijvoorbeeld de be-

leids- of starterpanels van MKB-Nederland. De bereidheid om deel te nemen is groot en juist op dit soort gegevens kan datamining goed tot zijn recht komen.
Minimaal een keer per jaar, of meer inden nodig, zal er een evaluatie van het project plaatsvinden. Dan kan besloten worden het project door te laten gaan of af te breken. Na vier jaar zal de nieuwe service winstgevend en de investeringen terugverdiend moeten zijn.

7 Plan

7.1 Plan Ontwikkeling

- De uitvoering van het investeringsplan valt onder de verantwoordelijkheid van de accountmanager van de afdeling DataWarehousing.
- Rapportage vindt jaarlijks plaats aan het MT van EIM BV

7.2 Planbeheer

Het product van de investering is ervaring met een nieuwe onderzoeksmethode. Kennis van deze methode moet zoveel mogelijk benut worden. Hoe meer onderzoeken door middel van datamining gedaan kunnen worden, hoe meer naambekendheid en ervaring EIM Datamining Services krijgt. Overal waar EIM kan moet ze uitdragen dat deze nieuwe dienstverlening bestaat. Het belangrijkste waarop gestuurd moet worden is het binnenhalen van projecten. Er zal door de accountmanagers het benodigde aantal projecten binnengehaald moeten worden. Met name door de afdeling datawarehousing maar ook de andere accounts zullen dit nieuwe product moeten promoten. Ook zal intern gekeken moeten worden of voor het eerste vluchtige onderzoek datamining niet gebruikt kan worden. Datamining kan een goede eerste indruk geven van welke richting op gezocht moet kunnen worden.

7.3 Risico's

Het grootste risico dat er is dat er mede door de recessie geen projecten op datamining gebied binnengehaald worden. Dit risico is wel degelijk aanwezig, zeker door de hoge prijs die voor Datamining Services gevraagd moet worden. Klanten zullen overtuigd moeten raken van het nut en de kracht van datamining. Daarvoor zullen aansprekende resultaten neergezet moeten worden. Op dit moment is er niet direct een klant te vinden die met EIM een datamining traject in wil gaan. Men vindt het wel interessant maar men echt de stap tot de uitvoering hiervan zetten is niet aan de orde. Ook bij EIM ziet men wel mogelijkheden maar echt concreet wordt het niet. Mocht men ooit Datamining Services op willen zetten dan is de promotie van deze service en het kenbaar maken van de mogelijkheden van datamining voor klanten van het grootste belang.

7.4 Showstoppers

Showstoppers zijn redenen waarom het project afgebroken moet worden.
De eerste reden daarvoor is een matige tot zeer geringe opkomst bij de bijeenkomst over datamining voor ondernemers van SPSS. Dit is het moment om te kijken hoe de animo voor datamining bij bedrijven is. Ook kunnen in de periode tot november 2003 accountmanagers de belangstelling voor deze dienst bij beleidsinstanties in gesprekken getest hebben. Als er geen belangstelling is uit de markt dan moet men niet tegen de stroom in doorgaan.
De tweede reden is op het moment dat men bij een van de halfjaarlijkse besprekingen meer dan 50% achterloopt op de verwachte noodzakelijke resultaten. Dan blijkt dat ondanks alle positieve signalen de verwachte opbrengsten niet gehaald zullen worden. Dan is het misschien beter tussentijds te stoppen dan tegen beter weten in door te blijven gaan.

De derde reden is dat de wetgeving op het gebied van privacy het verplaatsen van deze gegevens verhinderd, maar dat lijkt me op het eerste gezicht niet het geval, zeker niet omdat deze gegevens slechts voor intern gebruik geanalyseerd worden[1].
De vierde reden is dat klanten hun gegevens niet willen delen met een andere organisatie. Om deze drempel weg te nemen zal er met de grootste zorgvuldigheid met de gegevens omgegaan moeten worden.

[1] http://www.ivir.nl/wetten/privacy/nederland.html

8 Product

Wat is het uiteindelijke product dat EIM aan deze investering overhoudt? Wat zijn zowel de directe als indirecte opbrengsten?

- Het directe resultaat van de investering zijn 2 medewerkers van de afdeling Data-Warehousing die de verschillende datamining technieken beheersen.

Het indirecte resultaat is dat

- EIM een sterkere positie in de beleidsonderzoek markt krijgt doordat EIM meer mogelijkheden kan aanbieden aan bedrijven dan de concurrenten
- EIM de markt waarop men zich kan richten vergroot
- EIM vernieuwt en bezig is om een allround beleidsonderzoeksbureau te worden.
- EIM met Datamining Services een leidende marktpositie op datamining gebied inneemt.

9 ROI

9.1 Opbrengsten

De opbrengsten van Datamining Services is de waarde die EIM toevoegt aan een datamining project. De toegevoegde waarde van Datamining Service wordt uitgedrukt in de tijd die een medewerker bezig is met de verschillende onderdelen van datamining:

- Het preparen van de data
- Het integreren van de data
- De kennis en het uitvoeren van datamining technieken
- Het omzetten van de resultaten in nuttige informatie
- Het gebruik van de software
- De kennis van de software
- Het gebruik van de hardware

Deze posten dienen in de offerte aan de klant opgenomen te worden. Het gebruik van de hardware en de opleiding van de medewerkers is al in de standaard tarieven opgenomen. Hier volgt een voorbeeld offerte van Datamining Services en een korte toelichting op deze offerte.

Activiteit		dagen	tarief in euro's	projectkosten	
Data prepareren		3,5	1030	3605	
Data integreren		1,5	1030	1545	
Datamining					
	Data selecteren	0,5	1030	515	
	Data transformeren	0,5	1030	515	
	Datamining	1,5	1030	1545	
	Evalueren gevonden patronen	2	1030	2060	
Rapport offreren en bespreken		0,5	1030	515	
Totaal		10	1030		10300
Gebruik software		10	1000	10000	
Verkrijgen extra gegevens				5000	
Totaal					15000
Totale kosten					25300

De tijd die het prepareren van de data kost is afhankelijk van de kwaliteit van de data die onderzocht worden. Daarom zal de tijd die met deze stap kost zal in ieder project afzonderlijk berekenend worden. Het prepareren van de gegevens zal vaak het meeste tijd kosten. Ook de rest van de datamining stappen dient voor de duidelijkheid naar de klant toe uitgesplitst te worden. Met name de verschillende onderdelen van het datamining proces, zo wordt bij de klant inzicht in de kosten gegeven. Ervaring moet uiteindelijk leren hoeveel dagen men met iedere stap in het proces bezig is. Het is lastig om hier op voorhand een inschatting van te geven.

Tot slot zijn er nog twee aparte posten in de offerte opgenomen. De kosten van het verkrijgen van de eventuele extra gegevens, door telefonische enquête of het kopen van gegevens. Tot slot de kosten voor het gebruik van de software. Deze is hier als een toeslag per dag opgenomen maar er zijn ook andere mogelijkheden te bedenken, bijvoorbeeld een vast bedrag per datamining project.

9.2 Kosten

De kosten die voor de investering gemaakt moeten worden zijn:

1 software
2 opleiding
3 promotie
4 onvoorziene kosten.

Ad 1) Er zijn twee softwarepakketten die in aanmerking komen om aan te schaffen, clementine van SPSS en SAS enterprise miner. Andere onderdelen van deze software worden op dit moment al door EIM gebruikt. Het lijkt logisch om een keuze uit deze twee te maken, maar eventuele alternatieven zijn de software van datadistilleries of BNT-Holland. Zowel clementine als enterprise miner voldoen aan de eisen aan een goede datamining software. Op basis van twee demonstraties lijkt enterprise miner flexibeler en krachtiger in gebruik, maar ook duurder. Voor clementine zijn de kosten € 63.100 in het eerste en € 47.140 in de opvolgende jaren, bij twee licenties. Bij de aanschaf van enterprise miner zijn de kosten nog hoger, namelijk ongeveer € 120.000 in het eerste en ongeveer € 80.000 in de opvolgende jaren.

Ad 2) De kosten van de opleiding bestaan per medewerker uit het volgen van een cursus (€ 2000) en de kosten van de tijd dat de medewerker niet inzetbaar is (10 * € 1000). Er zullen twee medewerkers dat opgeleid moet worden. Twee medewerkers kunnen elkaar ondersteunen en zullen elkaar kunnen vervangen bij vakantie en ziekte. Drie medewerkers lijken op dit moment niet nodig. Deze kosten zijn eenmalig aangezien de medewerkers door het bezig zijn met datamining hun kennis zullen onderhouden.

Ad 3) De verwachte kosten van de promotie zijn € 7500 per jaar.

Ad 4) Voor onvoorziene kosten een percentage van 20% opgenomen. Dit lijkt erg hoog maar reëel gezien de onverwachte uitgaven die bij het opstarten van een nieuwe dienstverlening komen kijken.

9.3 ROI berekening

De return on investment berekening is gemaakt op basis van de volgende uitgangspunten:

1 alleen de opbrengsten voor het gebruik van de software dragen bij aan het terugverdienen van de kosten. Met deze opbrengsten zullen de totale kosten van de investering terug verdiend moeten worden
2 de opbrengsten die gegenereerd worden door het inzetten van de medewerkers kunnen niet direct aan Datamining Services worden toegeschreven. Immers er was zonder Datamining Services ook werk voor hen te doen geweest.

Welke opbrengsten precies aan Datamining Services toegeschreven moeten worden zal nader bepaald moeten worden. Moeten ook de opbrengsten van de positieve uitstraling en verkrijgen van nieuwe opdrachten aan Datamining Services toegeschreven worden?
Er zijn twee verschillende berekeningen gemaakt één uitgaande van de aanschaf van Clementine van SPSS en één uitgaande van de aanschaf van Enterprise miner van SAS. De berekening maakt duidelijk wat de netto opbrengsten van Datamining Services in de eerste vier jaar zouden moeten zijn om de kosten terug te verdienen. Er wordt gerekend met de netto

contante waarde tegen waarin een rente van 5% wordt gehanteerd. Bij het berekenen van de opbrengsten is het uitgangspunt een groeiende vraag naar datamining services geweest.

ROI Datamining SPSS	0,82%			
	Jaar 0	Jaar 1	Jaar 2	Jaar 3
Kosten	€ 113.520	€ 65.568	€ 65.568	€ 65.568
Opbrengsten	€ 20.000	€ 48.000	€ 100.000	€ 160.000
Verschil	-€ 93.520	-€ 17.568	€ 34.432	€ 94.432
Netto contante waarde tegen 5%	-€ 93.520	-€ 16.731	€ 31.231	€ 81.574

Bij deze bedragen is de return on investment 0,82%.

ROI Datamining SAS	0,37%			
	Jaar 0	Jaar 1	Jaar 2	Jaar 3
Kosten	€ 181.800	€ 105.000	€ 105.000	€ 105.000
Opbrengsten	€ 31.875	€ 76.500	€ 159.375	€ 255.000
Verschil	-€ 149.925	-€ 28.500	€ 54.375	€ 150.000
Netto contante waarde tegen 5%	-€ 149.925	-€ 27.143	€ 49.320	€ 129.576

Bij deze bedragen is de return on investment 0,37%

Op basis van bovenstaande berekeningen en na vaststelling van de toeslag per dag voor het gebruik van de software kan het aantal dagen dat aan datamining besteed moet worden berekend worden.
Stel dat de toeslag voor de software bij de aanschaf van Clementine van SPSS op € 1000 per dag gesteld wordt. Dan moet EIM in de eerste vier jaar 20, 48, 100 en 160 dagen met datamining projecten bezig zijn om de kosten terug te verdienen. Bij de aanschaf van SAS enterprise miner zullen de dagopbrengsten hoger moeten zijn. Ten eerste omdat de software duurder is en ten tweede omdat een hogere prijs de vraag naar Datamining Services zal verminderen. Als die toeslag op € 2000 bij aanschaf van Enterprise miner van SAS gesteld wordt. Dan moet EIM respectievelijk 16, 39, 80 en 128 dagen in de eerste vier jaar aan datamining projecten werken om de kosten terug te verdienen. Houdt men ook bij SAS een toeslag van € 1000 aan, dan zal er 32, 78, 160 en 255 dagen in de eerste 4 jaar aan datamining projecten gewerkt moeten worden. De geschetste scenario's lijken onder realistische omstandigheden op dit moment niet haalbaar.
Een manier om de kosten te verminderen is een strategisch partnership. Er zou met verschillende partners gesproken kunnen worden. SAS en SPSS kunnen interessante partners zijn, zij hebben er ook belang bij software te verkopen. Zijn er geen afspraken voor licenties per project te maken? Is er over de prijs te onderhandelen? Er zou ook met een universiteit gesproken kunnen worden. Daarmee zouden de kosten van de software wegvallen, de software heeft de universiteit reeds in bezit. Een tweede voordeel kan zijn dat partnership met een universiteit een wetenschappelijke uitstraling geeft, die het beeld van EIM als gedegen onderzoeksbureau kan versterken.

10 Conclusie

Er worden steeds meer gegevens opgeslagen en informatie uit die gegevens wordt steeds belangrijker. Datamining is uitgegroeid tot een succesvolle methode om gegevens te analyseren. Op dit moment wordt datamining op grote schaal en met succes bij vele bedrijven toegepast. Het zijn met name ondernemingen als banken, verzekeringen en creditcard maatschappijen die profiteren van de voordelen van datamining.
De steeds groeiende datahoeveelheid en behoefte aan informatie uit die data is de bron van datamining. Deze bron zal niet snel droog vallen in de huidige informatie en communicatie samenleving.

De centrale vraag in dit investeringsplan was of investeren in datamining rendabel is voor EIM? Datamining Services is een dienstverlening die erop gericht is de mogelijkheden van deze techniek bij zowel intern als extern aan te bieden. Bedrijven en instellingen kunnen zo profiteren van de mogelijkheden van datamining zonder zelf over de hardware, de kennis en de software te hoeven beschikken. Er zijn genoeg mogelijkheden en ideeën om datamining toe te passen, zowel bij beleidsinstanties als bij het individuele bedrijfsleven. Echter het blijven ideeën geen concrete toepassingen. Datamining kan op een andere manier inzicht onderzoek doen, inzicht verkrijgen en andere soorten vragen beantwoorden dan door middel van klassieke analystische methoden.

Er zitten voordelen aan het opzetten van de nieuwe dienst, Datamining Services.
- de uitstraling
- de completere dienstverlening
- de goede aansluiting aan bij de kernactiviteit van EIM, beleidsonderzoek doen.

Deze voordelen maken datamining zeker een gebied wat blijvend aandacht verdient. Op basis van deze informatie kan EIM kennis vermeerderen over het bedrijfsleven en bedrijfsprocessen ten behoeve van beleidsbeslissingen. Door het opnemen van deze service in het dienstenpakket wordt EIM een echt allround beleidsonderzoekbureau.

De voornaamste nadelen van het opzetten van Datamining Services zijn:
- de markt, die op het gebied van Datamining Services nog in ontwikkeling is. Bedrijven lijken op dit moment niet bereid te investeren in datamining, ze zijn er nog niet overtuigd van de meerwaarde van datamining
- de financiële berekeningen, die uitwijzen dat er op dit moment grote risico's kleven aan deze investering. De hoge kosten van de software lijken het project niet haalbaar te maken. Een mogelijkheid om dit probleem te beperken is een strategisch partnership met een softwarebureau of universiteit. Dit heeft het voordeel dat de kosten van de software verminderen, waardoor het opzetten van Datamining Services haalbaarder wordt. De mogelijkheid van een stratgisch partnership dient nader onderzocht te worden.

Het is mijn mening dat EIM op dit moment niet zou moeten investeren in Datamining.
EIM zou wel, mede gezien de huidige dienstverlening, de (software) markt op het gebied van datamining in de gaten moeten blijven houden. Want in de toekomst zijn er kansen voor EIM op het gebied van Datamining Services. De vraag is op welk moment EIM hiervoor stappen moet zetten. Er zijn genoeg verhalen die de potentie van datamining technieken bevestigen. Er zal interesse in datamining komen en men zal beseffen dat voor goede en degelijke datamining specialisten nodig zijn. In de toekomst kan Datamining Services een goede aanvulling van het dienstenpakket van EIM zijn.

Bijlage I SWOT DW

SWOT DW

Strength (kracht)
- Het omgaan en bewerken van data
- Combinatie inhoud en techniek
- Ervaring door EZ programma
- Technische kennis, goede opleiding van de afdeling

Weaknesses (zwaktes)
- Grote afhankelijkheid van enkele grote opdrachtgevers, met name ministeries
- Onbekendheid naar buiten toe

Opportunities
- Veel vraag naar structurering data en onderzoek daarnaar
- toepassing clearinghouse als generiek concept (!) maak onbekende service bekend
- (slogan: “maak Uw data consistent!”)
- profileren datamakelaar data-expert
- vanuit dataprojecten bijdragen aan onderzoek
- vanuit dataprojecten toename netwerk (datamining)

Treaths
- Stoppen EZ programma
- Veel concurrentie door malaise IT-sector

Bijlage II ROI berekeningen

ROI Datamining SPSS	0,82%			
SPSS	Jaar 0	Jaar 1	Jaar 2	Jaar 3
Kosten	€ 113.520	€ 65.568	€ 65.568	€ 65.568
Noodzakelijke opbrengsten	€ 20.000	€ 48.000	€ 100.000	€ 160.000
Verschil	-€ 93.520	-€ 17.568	€ 34.432	€ 94.432
NCW	-€ 93.520	-€ 16.731	€ 31.231	€ 81.574

Kosten jaar 0		
Aanschaf software	63100	
Opleiding medewerkers	4000	
Kosten out of billing	20000	
Kosten promotie	7500	
Onvoorzien (20%)	18920	
	Totaal	€ 113.520
Opbrengsten jaar 0		
Opbrengsten	20000	
	Totaal	€ 20.000
Kosten jaar 1		
Licentie software	47140	
Kosten promotie	7500	
Onvoorzien (20%)	10928	
	Totaal	€ 65.568
Opbrengsten jaar 1		
Opbrengsten	48000	
	Totaal	€ 48.000
Kosten jaar 2		
Licentie software	47140	
Kosten promotie	7500	
Onvoorzien (20%)	10928	
	Totaal	€ 65.568
Opbrengsten jaar 2		
Opbrengsten	100000	
	Totaal	€ 100.000
Kosten jaar 3		
Licentie software	47140	
Kosten promotie	7500	
Onvoorzien (20%)	10928	
	Totaal	€ 65.568
Opbrengsten jaar 3		
Opbrengsten	160000	
	Totaal	€ 160.000

ROI Datamining SAS	0,37%			
SAS	Jaar 0	Jaar 1	Jaar 2	Jaar 3
Kosten	€ 181.800	€ 105.000	€ 105.000	€ 105.000
Opbrengsten	€ 31.875	€ 76.500	€ 159.375	€ 255.000
Verschil	-€ 149.925	-€ 28.500	€ 54.375	€ 150.000
NCW	-€ 149.925	-€ 27.143	€ 49.320	€ 129.576

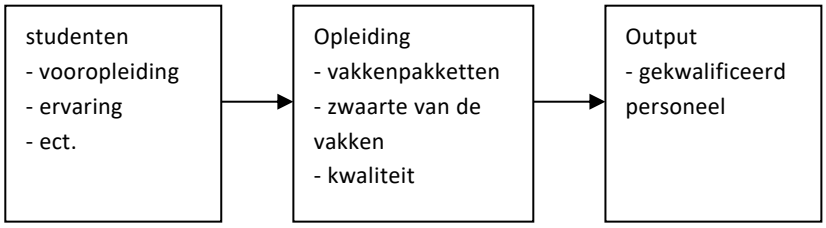
Kosten jaar 0			
Aanschaf software	120000		
Opleiding medewerkers	4000		
Kosten out of billing	20000		
Kosten promotie	7500		
Onvoorzien (20%)	30300		
	Totaal		€ 181.800
Opbrengsten jaar 0			
Opbrengsten	31875		
	Totaal		€ 31.875
Kosten jaar 1			
Licentie software	80000		
Kosten promotie	7500		
Onvoorzien (20%)	17500		
	Totaal		€ 105.000
Opbrengsten jaar 1			
Opbrengsten	76500		
	Totaal		€ 76.500
Kosten jaar 2			
Licentie software	80000		
Kosten promotie	7500		
Onvoorzien (20%)	17500		
	Totaal		€ 105.000
Opbrengsten jaar 2			
Opbrengsten	159375		
	Totaal		€ 159.375
Kosten jaar 3			
Licentie software	80000		
Kosten promotie	7500		
Onvoorzien (20%)	17500		
	Totaal		€ 105.000
Opbrengsten jaar 3			
Opbrengsten	255000		
	Totaal		€ 255.000

Bijlage III Gespreksverslagen

Gesprek met de heer F. Pleijster d.d. 24-04-2003

Na een korte introductie van datamining kwam de heer Pleijster met de volgende voorbeelden aanzetten waarbij datamining door EIM mogelijk zou zijn

Branche organisaties (UNETO)
Branche organisaties tellen een x aantal leden en het enige directe contact is het innen van de contributie, direct is er geen onderwerp voor datamining, er zijn echter andere indirecte mogelijkheden. Veel branche organisaties hebben eigen opleidingscentra.



Er is te sturen in de opleiding met als doel beter personeel te krijgen. Als nu op basis van de eigenschappen van student x een opleiding voor die persoon opgezet wordt zodat deze persoon maximaal rendement krijgt van zijn studie dan is dat beter voor de branche. De mogelijkheid om dit te financieren is er door middel van O&O fondsen (ontwikkeling en opleiding). Deze fondsen worden gesubsidieerd door en werknemers en werkgevers. Hier zou een mogelijkheid voor datamining services voor EIM liggen

commerciële organisaties (EUETO, INTRES)
Sommige bedrijven proberen op het logistieke vlak samen te werken. Om deze inkoop en andere deeltrajecten te optimaliseren willen deze organisaties steeds meer gedetailleerde informatie van de aangesloten bedrijven. Er zouden mogelijkheden zijn voor EIM om ook op deze gegevens datamining services aan te bieden

Schuldhulpverlening
Men is bezig om alle gegevens met betrekking tot de schuldhulpverlening te centraliseren. De partijen die aanwezig zijn bij deze dienst zijn: de persoon in kwestie, de sociaal maatschappelijk werken, gemeentelijke sociale dienst, de gemeentelijke kredietbank. Als al deze gegevens samengevoegd worden kan bekeken worden, welke methoden bij wat voor soort cliënten succes hebben.

Ziektemelding
Men wil per branche alle ziektemeldingen centraal gaan registreren. Met als deelnemende partijen, ARBO, verzekeringen, en de verschillende bedrijven.

Gespreksverslag met W. Verhoeven d.d. 24-04-2003

Na een korte introductie van datamining (hij heeft het toegestuurde memo niet gelezen) blijkt dat er in zijn account, ondernemerschap en marktontwikkeling eigenlijk geen directe datamining diensten aan te bieden zijn. De enige mogelijkheid zouden de starterpanels zijn een soort van monitor hoe het met de starters in bepaalde branches gaat.
Er wordt wat gesproken over de markt waarin EIM opereert, als grootste concurrenten worden Research voor beleid, NEI en Bartels genoemd, maar het blijkt dat ook de consultancybedrijven, PWC KPMG, zich in deze markt begeven omdat hun eigen markt door economische teruggang niet meer genoeg oplevert. Van de economische recessie heeft EIM niet zoveel last, wel van het feit dat het kabinet nog steeds demissionair is. Er wordt geen beleid meer gemaakt en dus ook geen onderzoeken meer om te kijken of het gevoerde beleid effectief is. Naast ministeries (EZ, BZK) zijn VNO-NCW en MKB Nederland grote opdrachtgevers voor EIM.

Gespreksverslag met de heer P. Vroonhof d.d. 28-04-2003

EIM is sterk om de inhoudelijke component (onderzoeksvragen etc.) te combineren met de technische kennis. Datamining zou dus heel goed een project van Consult ism data en informatiesystemen kunnen worden.
Er zijn behoeften op het gebied van sociale zekerheid bij zowel overheden als beleidsinstanties. Daarnaast moeten de grotere MKB-bedrijven ook niet uitgesloten worden.
Te denken valt aan de belastingdienst maar met name de intergratie bedrijven. Overheden willen weten hoe bepaalde bedrijven het doen, maar dat ligt ook aan de kenmerken van de werknemers. Reïntegratiebedrijven willen weten wat ze in hun offertes als score verwachting aan kunnen bieden. Hot item ivm omvorming WAO.
Daarnaast valt ook te denken aan klantentevredenheid en positioneringonderzoek. Waarin ben ik goed, waarin ben ik slecht hoe kan ik het verbeteren aan welke kenmerken ligt dat.
Ook is er de mogelijkheid om analyse te doen naar de verschillende paneldata. Meestal wordt nu de kenmerken van bijvoorbeeld sociale zekerheid gekoppeld aan de standaard gegevens van het bedrijf. Ook worden de milieugegevens gekoppeld aan deze standaard data. Echter de sociale zekerheid data wordt niet vergeleken met de milieugegevens in combinatie met de standaard gegevens. Zo zijn er meerdere sub-onderzoeken waardoor er veel verschillende data kenmerken van de bedrijven aanwezig is.
Een andere dataset waar misschien iets mee kan gebeuren is de WIW dataset
Daarnaast ligt er nog een hele leuke maar gevoelige dataset in de kluis van EIM een set waarbij de belasting aangifte gekoppeld is aan de ziekenfondsuitgaven van de persoon. Hot ivm basisverzekering.

Vragen die nodig beantwoord moeten worden
- Wat mag en kan met de data
- Hoe gevoelig is DM voor missing values

Het zou leuk zijn om een onderzoek te doen en aan de hand daarvan iets te publiceren en in het persbericht de kreet datamining een keer of wat te laten vallen.

Voorlopige conclusies
Er zijn mogelijkheden voor EIM op gebied van datamining
Grootste struikelblok is het bekend maken van de potentiële klanten met deze techniek en dan zelf als consultant binnen komen. Dit kan het beste in combinatie met EIM consult. Daardoor wordt ook gelijk de kracht van EIM duidelijk, juist de combinatie van kennis en techniek. De vraag van de manager kunnen vertalen en die vraag technisch goed en onderbouwd op kunnen lossen. Kees schijnt er goed in te zijn (quote P. Vroonhof) om mensen te inspireren en te overtuigen.
Misschien zelf eerst wat onderzoeken doen voor MKB Nederland op basis van het samenvoegen van de verschillende paneldata, daarvoor moet Joris Meijaard praten
Er zal dus dataminingsoftware aangeschaft moeten worden, wat zijn de kosten daarvan?
Is MKB te betalen voor de nieuwe monitor gegevens? (zo ja dan zijn we een eind op weg)
Vooral binnen arbeid en sociale zekerheid veel mogelijkheden
Ook bij branches in sectoren liggen mogelijkheden

De kunst wordt dus om klanten over te halen om software te kopen zodat EIM kan helpen bij de analyse van de uitkomsten. Daar komt wel de kracht van EIM om de hoek, combinatie kennis en techniek!!
Er moeten wel vragen beantwoord worden op gebied van

- privacy
- missing values (garbage in garbage out???)
- Wat kan en mag met de data

Gesprek met P. Brouwer, account manager Arbeid en Sociale Zekerheid

In het gesprek kwam duidelijk naar voren dat er interesse bestaat voor datamining, maar er bestonden nog veel vragen over de techniek en de soort van data waarop datamining gebruikt kan worden. De achtergrond hiervan was dat er een concrete offerte in de maak was, waarin misschien al datamining gebruikt zou kunnen worden. Het was inderdaad een interessante dataset maar het is wat prematuur om er al in een offerte te spreken van datamining. Het is nog in een fase waarin ik eerst moet inventariseren of het inderdaad nuttig is. Het blijkt dat er zeker bij het account A en SZ mogelijkheden liggen. Dit is in ieder geval een positieve conclusie.
Daarnaast hebben we gesproken over mogelijkheden voor datamining. Met name bij de CWI's (centra voor werk en inkomen) liggen eventueel mogelijkheden. Misschien is er wel een mogelijkheid om met iemand van het CWI eens te gaan praten.

Gespreksverslag met de heer R. Vogels EIM Stratus

De heer Vogels is goed bekend met datamining en de mogelijkheden van datamining. Ook de achtergrond en het soort data waarop deze techniek toegepast zou kunnen worden is duidelijk. Hij ziet niet waarom het nuttig voor EIM zou kunnen zijn om zich met datamining bezig te gaan houden, hiervoor voert hij een paar bezwaren aan.
EIM BV richt zich met name op beleidsinstanties en deze instanties beschikken niet over het soort data waarop datamining van toepassing zou kunnen zijn. De data van het CBS zijn geaggregeerd en andere klanten beschikken ook niet over het type data waarop het zinnig is datamining toe te passen. Dat is bezwaar één.
Ten tweede is het zo dat de bedrijven die wel de data voor datamining hebben, dit allang doen. Bedrijven als comfort card doen al aan datamining evenals de verschillende banken en de Air Miles organisaties. Bedrijven als Nielsen en GfK zijn al gevestigd in de datamining markt. Om daar als nieuwe speler nu in te komen wordt lastig want het is een hele competitieve markt.
Ten derde zijn de competenties die nodig zijn voor datamining niet aanwezig bij DW. Het vereist een bepaalde mate van creatief spelen met data. Constant op zoek zijn naar resultaten en niet genoeg nemen met tussenresultaten. Hij ziet niemand van de afdeling DW daar op dit moment toe in staat ook niet na een cursus. Mocht EIM iets aan datamining willen doen dan moet daar een nieuw persoon voor in dienst genomen worden.
De enige optie zou zijn om een soort van partnership met SPSS te sluiten en daar bij hun experts de kunst af te kijken. Immers op dit moment heeft EIM BV geen dataset waarmee men kan oefenen om de kunst onder de knie te krijgen. Dan moet EIM met die kennis de markt in gaan. De markt is in dit geval de bedrijven waar EIM contact mee heeft. Beleidsinstanties maar vooral de grotere bedrijven die eigenlijk geen tijd maar wel geld hebben en de data onderzocht willen hebben. Echter deze bedrijven zullen nieuwe contacten moeten zijn want in het huidige bestand van bedrijven zijn weinig bedrijven te vinden die aan de benodigde criteria voldoen.
Ook zou eens gekeken moeten worden welke software EIM al in bezit heeft. Is het echt nodig om Clementine® aan te schaffen of kan men het af met Neural connection en Answer Tree, twee modules behorende bij SPSS die al gebruikt kunnen worden.
Als enige mogelijkheden ziet de heer Vogels, het CWI en ARBO diensten alhoewel daar de privacy zeer gevoelig ligt.

Gespreksverslag met de heer P. Risseeuw

Datamining is een nuttige techniek, die echter de nodige wiskundige en economische kennis vereist. Bedrijven zullen deze techniek niet in huis hebben en zullen ook niet over deze kennis kunnen beschikken daarvoor zijn de mensen niet competent.
Daarna wordt de markt van EIM besproken
Beleidsinstanties zullen over het algemeen niet over de juiste data beschikken. De grote spelers in de datamining markt zullen het allemaal al wel zelf doen, denk aan banken en verzekeringsmaatschappijen. Alhoewel het af en toe ook erg klungelig gebeurt, met name het in de markt zetten van de bonus kaart wordt aangehaald. En wat doet Albert Heijn ermee? Hij merkt er niet veel van...
Wanneer zijn er dan wel mogelijkheden voor datamining. Met name in markten waarin het gedrag van klanten, mede door internet, aan het veranderen is. Men komt niet meer langs bij een reisbureau, men boekt zijn vakantie en vlucht via het internet. Men kijkt niet meer in de etalage van makelaars, men zoekt een nieuwe huis via het internet. Of assurantie maatschappijen die met alle verschillende verzekeringen contact hebben ook daar verandert er veel door de voortschrijdende informatisering. In dit soort markten ontstaan grote hoeveelheden data waaruit (klanten)profielen te halen zijn. Dit soort informatie zou interessant kunnen zijn voor bedrijven. Daar liggen de grotere mogelijkheden op het gebied van datamining binnen het midden en kleinbedrijf.

Gesprek met de heer De Koning, SPSS d.d. 01-07-2003

Het gesprek begon met het op tafel leggen van het plan van EIM om te verkennen of er een markt voor datamining zou zijn. En dat EIM bedrijven door middel van datamining een nieuwe service aan zou kunnen bieden. De reactie van de heer De Koning was: “je slaat de spijker op zijn kop”.
Zij hadden naar dezelfde dienstverlening gekeken en daar een white paper over geschreven. Zij zagen al voor het gesprek een rol voor EIM in deze markt. Wel is het zo dat deze markt zich vooral bij de retail bevindt, dus meer het gebied van Stratus en Consult. EIM zou de software en data moeten combineren tot nuttige informatie want duidelijk is dat de retail markt dat zelf niet gaat doen. Er is geen geld voor, er is niet voldoende kennis, ze hebben niet de achtergrond om het te kunnen en tot slot hebben ze er ook geen tijd voor. Dus er zullen datamining adviesbureaus moeten komen. Zij zien juist voor marktonderzoeksbureau een plaats in deze markt. (zie ook de white paper, two rivers)
Het is wel zo dat de taart nu verdeeld wordt, als men iets van deze taart mee wil krijgen dan zal men nu de stap moeten maken over 3 à 4 jaar is het te laat! De markt is zich nu aan het vormen, en publiciteit is dan ook stap nummer 1.
In september organiseert SPSS een bijeenkomst over ondernemers en datamining en daar zou EIM een van de sprekers kunnen zijn.
Daarna gingen er verschillende ideeën over toepassingen van datamining door EIM over tafel. Allemaal erg leuk en volgens mij ook haalbaar al zijn verschillende ideeën ook toekomst muziek. Het belangrijkste advies start klein en start strategisch. SPSS is bereid te helpen bij het opzetten van deze service. Het vreemde was realiseerde ik mij achteraf dat SPSS dan een soort consult aan EIM gaat aanbieden en dat EIM de uitvoerende instantie wordt. (eigenlijk hadden we in het begin van dit plan iets anders in gedachten)

Bijlage IV Van 'Ist' naar 'Soll'

Strategie	EIM groeit uit tot een allround beleidsonderzoekbureau		
	Ist →	Verandering →	Soll
Software	Geen datamining software	Aanschaf datamining programma	Mogelijkheid om datamining binnen EIM uit te voeren
Technische infrastructuur	Xeon dual processor	Geen	Xeon dual processor
Mensen	Interesse voor datamining	cursus volgen op gebied van datamining	beheersen van datamining
Organisatie	Afdeling DataWarehousing	Datamining Services opstarten	Datamining Services als onderdeel van DataWarehousing
Markt	Onderzoek voor beleidsinstanties	Promotie via huidige kanalen en aan klanten van huidige klanten	Marktuitbreiding via beleidsinstanties naar de individuele ondernemers
Product & Service	Alleen beleidsonderzoek op klassieke manier	Uitbreiding diensten met Datamining Services	Datamining Services groeit uit tot een succesvolle uitbreiding van de dienstportfolio van EIM

Software
Er zal software die datamining uit kan voeren aangeschaft moeten worden.

Technische infrastructuur
Op het gebied van de hardware zijn alle benodigde assets aanwezig.

Mensen
De mensen zullen moeten leren wat datamining inhoud. Niet alleen de mensen die de datamining problemen op moeten lossen maar zeker ook de accountmanagers, die het moeten verkopen. Daarnaast is het ook nuttig om alle medewerkers van deze nieuwe service op de hoogte te stellen zodat ze eventuele datamining mogelijkheden binnen hun onderzoek kunnen herkennen. Daartoe zal ook binnen EIM goede voorlichting over wat datamining is en inhoud moeten plaatsvinden.

Organisatie
Het organogram zal in beginsel niet aangepast worden. Datamining services valt eerst onder de afdeling DataWarehousing. Mocht het potentieel groot blijken dan zou overwogen kunnen worden er een apart account van te maken.

Markt

Er zal intern en extern veel reclame voor deze nieuwe dienst gemaakt moeten worden. EIM gaat een nieuwe service een nieuw product in de markt zetten en dat moet bekend worden. Om deze introductie goed te laten verlopen en het product in de markt te zetten zal er veel promotie rondom deze service gemaakt moeten worden. De dienst zal de markt in gepusht worden. Dit houdt in dat de markt zelf nog niet op een aanbod van Datamining Services vraagt, men zal de vraag zelf moeten creëren.

Product & Service
EIM breidt zijn dienstverlening aan de klant uit met Datamining Services. Deze service richt zich op het vinden van informatie uit grote bergen gegevens. Deze service zal vallen onder de afdeling DataWarehousing.

Bijlage V Datamining bij de Nederlandse politie

De mogelijkheden van *datamining* voor de Nederlandse politie

Het Tijdschrift voor de Politie, nr. 6, juni 2000, pag. 26-29
Door: Manuel J.J. López*

Binnen veel bedrijven en organisaties is datamining een middel om inzicht te krijgen in 'de markt' en in de kenmerken en wensen van 'de klant'. Telefonische marktonderzoeken, Air Miles en klantenkaarten zijn bekende middelen om aan informatie te komen. Multivariate analysetechnieken en neurale netwerken worden achter de schermen gebruikt om inzicht te krijgen in klantprofielen en veranderende marktomstandigheden. Ook voor de politie kan data mining nuttig zijn. Door gebruik te maken van dezelfde analysetechnieken als in de marketing kunnen profielen worden opgesteld van daders en slachtoffers van criminaliteit. Van wijken en slachtoffersegmenten kunnen risicoprofielen worden gemaakt die het mogelijk maken om preventieactiviteiten gericht in te zetten. Door patronen van modus operandi in kaart te brengen en te combineren met onze kennis over 'soorten' daders worden aanwijzingen gevonden die kunnen leiden tot aanhouding van een verdachte.

Dit artikel behandelt verschillende mogelijkheden van data mining voor de Nederlandse politie. Aan de hand van concrete praktijkvoorbeelden wordt hierbij uitgelegd welke analysetechnieken misdaadanalisten kunnen gebruiken om tot een verdiept inzicht over de criminaliteit te komen. Gezien het brede assortiment aan data mining mogelijkheden wordt hierbij een selectie gemaakt.

Slimmere inzet van blauw

Binnen de Nederlandse politie zien we een beweging waarin informatie en kennis een steeds belangrijker rol gaan spelen. De werkomgeving van de politie verandert immers snel. De toenemende verstedelijking, mobiliteit en individualisering van de maatschappij leiden tot een afbrokkeling van de handhavingmogelijkheden. Door het grote aantal delicten is de werkdruk bij de politie enorm gestegen. Ontwikkelingen binnen de technologie en gebouwde omgeving leiden tot continue veranderingen in modus operandi van daders. Het openstellen van de grenzen binnen de Europese Unie en de algemene internationalisering van de criminaliteit zorgen dat daders gemakkelijker de dans ontspringen dan twee of drie decennia geleden. Toenemende regelgeving en een steeds verdergaande inperking van politiebevoegdheden hebben het werk complexer en moeilijker gemaakt. Aangezien de grenzen van 'meer blauw' inmiddels duidelijk zijn bereikt, ligt de keuze voor een 'slimmere inzet van blauw' voor de hand. In dit zoeken naar mogelijkheden heeft de politie inmiddels ook informatie ontdekt als *'corporate resource'*. Informatie wordt hierbij gezien als middel dat net als arbeid, kapitaal en materiaal kan worden aangewend om doelstellingen te verwezenlijken. Voor het genereren van deze informatie kan in toenemende mate gebruik worden gemaakt van steeds krachtiger ICT-producten en -processen. Een voorbeeld van zo'n proces is data mining.

Wat is data mining?

Wanneer we een korte literatuurstudie zouden doen, zien we dat data mining een relatief nieuw en divers proces is waarin concepten uit de marketing, database management, statistiek en computer wetenschappen met elkaar worden gecombineerd. Door deze diversiteit zijn er eigenlijk evenveel definities van het begrip als er invalshoeken zijn. Als rode draad door deze definities lopen echter de woorden *zoeken, zinvolle verbanden* en *toepassingsgerichtheid*. Voor ons doel houden we daarom de definitie aan

* drs. Manuel J.J. López werkt als zelfstandig onderzoeker en adviseur bij Result Crime Management en is als zodanig ingehuurd door het Expertisecentrum Woningcriminaliteit.

die door het softwarebedrijf SPSS wordt gehanteerd. Deze definitie luidt *"data mining is het zoeken naar patronen en verbanden in uw gegevens om uw werk beter te kunnen doen"*.[1] Data mining maakt gebruik van technieken uit de statistiek en integreert deze met de moderne uitgangspunten van database management. In de traditionele statistiek stond met name het toetsen van hypothesen en theorieën centraal. Binnen de data mining ligt de nadruk veeleer op exploratie. De data miner kijkt met een open blik naar zijn gegevens in de hoop hierin onverwachte verbanden en verborgen patronen te ontdekken. Vanuit de marketing dankt de data mining haar focus op toepassingsgerichtheid. Immers: het vergaren van informatie is leuk, maar je moet er ook iets mee kunnen doen.

Omdat gebruik wordt gemaakt van de allernieuwste computertechnologieën wordt data mining door sommigen als een nieuw tovermiddel beschouwd; een gemakkelijke maar ondoorzichtige manier waarbij grote hoeveelheden data min-of-meer automatisch door 'smart technologies' worden omgevormd tot hapklare informatieblokken. Niets is echter minder waar. Data mining is op zichzelf geen oplossing; het is een *procedure* die we kunnen toepassen om oplossingen te vinden voor organisatorische vraagstellingen. Dit proces gaat niet automatisch. Het vereist harde arbeid, inzicht en planning.

Betekenis voor de politieorganisatie

Op dit moment loopt in het politiekorps Limburg-Zuid een pilot-project dat bekend staat onder de naam *Innovatieve Misdaadanalyse - Woninginbraak* (afgekort IMA). Het doel van dit project is om kennis over woninginbraak te genereren die verder gaat dan wat we nu weten.

Om dit te bereiken, wordt binnen het IMA-project een breed arsenaal van data miningtechnieken ingezet. Deze data mining technieken worden beproefd en op hun bruikbaarheid voor de Nederlandse politieorganisatie beoordeeld. Conventionele misdaadanalysetechnieken worden gebruikt en aangevuld met data mining-technieken die nog niet tot het conventionele pakket van de misdaadanalist behoren. Door gebruik te maken van beschikbare HKS-data ontstaat inzicht in de kwaliteit en kwantiteit van deze politiegegevens en wordt ervaring opgedaan met de inzet van innovatieve analysetechnieken bij de Nederlandse politie. De combinatie van gangbare en minder gangbare analysetechnieken leidt tot een gevalideerd analysemodel dat verder gaat dan wat men tot nu toe binnen de misdaadanalyse gewend is. Er ontstaat inzicht in de mogelijkheden en beperkingen van politiegegevens voor misdaadanalyse en er ontstaat een overzicht/informatiemodel van gegevens die verzameld moeten worden om het delict woninginbraak wel volledig en accuraat in beeld te kunnen brengen.

Onlangs is de eerste fase van het IMA-project afgesloten met de publicatie van een schriftelijke rapportage *Inbraakanalyse* en bijbehorende cd-rom. In deze rapportages kunnen we lezen dat de kwantiteit en kwaliteit van de huidige politiegegevens dermate beperkt zijn dat veel wezenlijke vragen over woninginbraak vooralsnog onbeantwoord moeten blijven.[2] Tegelijkertijd blijkt echter ook dat de politie veel meer informatie uit haar bestanden kan halen dan tot nu toe mogelijk werd gehouden. Door gebruik te maken van een eenvoudig stappenplan is het mogelijk om conventionele technieken voor misdaadanalyse te combineren met technieken die tot nu toe vooral tot het domein van marktonderzoekers en sociale wetenschappers behoren. Door dit te doen ontstaat niet alleen inzicht in de omvang en spreiding van delicten, maar ook in de samenhang tussen de verschillende aspecten van deze delicten.

Praktische voorbeelden

[1] Zie hiervoor de publicatie *Data Mining with Confidence*.

[2] Om dit te verbeteren is binnen het project *Innovatieve Misdaadanalyse - Woninginbraak (IMA)* een informatiemodel ontwikkeld dat verderop in dit tijdschrift door drs. B. Adriaens wordt besproken.

Om de bruikbaarheid van data mining voor de Nederlandse politie te demonstreren, volgt nu een drietal praktische voorbeelden. Deze voorbeelden bespreken respectievelijk de mogelijkheid om patronen te vinden in het (wederechtelijke) verzamelgedrag van de woninginbreker, een meer gedetailleerde vorm van profielanalyse en de mogelijkheid om risicoprofielen van wijken, woningen en/of slachtoffers te bepalen.

Market basket analyse

In marktonderzoek wordt soms gebruik gemaakt van *market basket analysis* en *customer profiling reporting*. Deze technieken maken het mogelijk om na te gaan wat winkelende klanten zoal in hun boodschappenmandje stoppen en of hier bepaalde patronen of combinaties in te vinden zijn. Door het toepassen van deze technieken is onder meer ontdekt dat sommige klanten (vooral jong volwassen mannen) vrijdagmiddag vlak voor sluitingstijd de supermarkt binnenlopen om luiers, chips en bier te kopen. Omdat deze aankoopcombinatie vaak voorkomt, spelen sommige supermarkten hier slim op in door daar selectief hun kortingsacties of assortiment op af te stemmen. Ook binnen misdaadanalyse zijn dergelijke analyses interessant. Het blijkt weliswaar niet haalbaar om iedere woninginbreker met een Air Miles of klantenkaart uit te rusten, maar we kunnen wel onderzoeken wat zij in hun 'boodschappenmandje' stoppen. Zijn er in ons bestand bepaalde buitcombinaties te ontdekken? En hoe zit het met veelvoorkomende combinaties in persoonskenmerken, wijzen van binnenkomst, gedragingen op de p.d. en gebruikte hulpmiddelen/werktuigen? Om dit soort combinaties te onderzoeken, maken marktonderzoekers gebruik van technieken als GRI[1] en factoranalyse. In ons onderzoek is voor de laatste analysetechniek gekozen.

In het gegevensbestand van woninginbraak in Limburg-Zuid vinden we zeven veelvoorkomende buitcombinaties. Soms lijken de combinaties logisch, maar in andere gevallen zijn ze ronduit opvallend. Het typerende hieraan is dat ze als buit vaak in combinatie weggenomen worden. De grootste groep bestaat uit diverse vormen van papieren zoals *identiteitspapieren*, *waardepapieren*, *kentekenbewijzen* maar ook een *portemonnee* (waar bijvoorbeeld de identiteitspapieren in kunnen zitten). In het oog springt ook de combinatie *cosmetica* en *kleding*. De vraag die dan meteen rijst is gericht op het geslacht van de dader. Zijn het voornamelijk vrouwen die deze buitsoort wegnemen? Nadere analyse leert dat relatief gezien meer vrouwen voorkeur hebben voor deze buitsoort. Het gevonden verband is echter niet significant en kan op toeval berusten. Het zou derhalve ook zo kunnen zijn dat de kleding en cosmetica met name door groepen inbrekers worden ontvreemd. Vrouwen plegen woninginbraken namelijk vaak in gezelschap van mannelijke mededaders.

Ook de buitcombinatie *personenvoertuig* en *sleutels* is opvallend. Het blijkt dus dat er daders zijn die ervoor kiezen om in een woning in te breken met als doel daar de autosleutels en (vervolgens) personenauto te ontvreemden. Om op dit fenomeen zicht te krijgen, is de market basket analysis op dit punt uitgebreid met een Chaid-analyse.[2] In ons bestand zitten 214 zaken waarbij de woninginbrekers het voertuig als buit hebben meegenomen. In 13,3% van deze gevallen (62 maal) werd dit voertuig meegenomen in combinatie met de sleutels. Binnen de woninginbraken waarbij zowel de sleutels als het voertuig zijn gestolen, valt verder op dat er een sterk verband bestaat met het gebruik van een valse sleutel of flipper. In 39 van de 62 gevallen (63%) heeft de dader een dergelijk hulpmiddel gebruikt om de woning binnen te komen. Bij zowel de daders die gebruik maken van een valse sleutel of flipper als bij de daders die op een andere wijze de woning zijn binnen gegaan, valt voorts op dat een groot deel van hen de te stelen auto gebruikt om hiermee tevens geluids- en/of beeldapparatuur mee te nemen. Bij de wo-

[1] GRI staat voor 'Generalized Rule Induction' en valt onder de categorie associatieve modeleertechnieken die ook wel vaak worden gebruikt in gecomputeriseerde kennissystemen en neurale netwerken.

[2] Chaid staat voor 'Chi-square Automatic Interaction Detector'.

ninginbrekers die wel het voertuig maar niet de sleutels meenemen, zien we dat een belangrijk gedeelte (16%) kleding, textiel en/of schoeisel meeneemt.

Profielanalyse
Binnen het HKS-systeem zijn nagenoeg geen gegevens aanwezig die het mogelijk maken om een uitgebreide situationele analyse van woninginbraak te maken. Om dit probleem te omzeilen, is het databestand binnen het IMA-project uitgebreid met sociaaldemografische wijkgegevens van het CBS. Hierdoor werd het mogelijk om situationele typologieën op wijkniveau te vinden.[1]

Om inzicht te krijgen in situationele typologieën, is binnen de analyse gebruik gemaakt van twee benaderingswijzen: een conventionele en een vernieuwende vorm van analyseren. Bij de conventionele benaderingswijze wordt gebruik gemaakt van kruistabellen en de Chi^2-toets. Bij de vernieuwende vorm van analyseren maken we een combinatie van *clusteranalyse*, *compare means* en *discriminantanalyse*. Wat betreft de uitkomst vertonen beide benaderingen veel gelijkenis. Toch zien we dat er bij de conventionele manier veel informatie verloren gaat, wanneer we enige significantie tussen variabelen willen ontdekken. Bij de analyse op de nieuwe manier wordt meer informatie behouden. Er is niet alleen zicht op het aandeel van de gegevens tot de vorming van de clusters, maar ook op de onderlinge verhouding daartussen.

Daar we op wijkniveau met 104 cases te maken hebben (104 wijken in de regio Limburg-Zuid) en we het aantal te vormen clusters niet op voorhand kennen, kiezen we bij de vernieuwende vorm van analyseren voor de zogenaamde hiërarchische clustertechniek. De variabelen "% allochtonen" en "% inkomensontvangers met uitkering" blijken in ons geval een negatieve invloed op de uitkomst uit te oefenen, omdat ze bij veel cases geen waarden vertonen. In dit geval wordt de analyse op slechts 45,5% van de casussen uitgevoerd. Als de analyse opnieuw uitgevoerd wordt, maar deze keer zonder deze variabelen, zien we een uitvoering op 97,1%. Uiteindelijk komen we tot de vorming van twee clusters, die respectievelijk 81% en 10% van de variantie verklaren. In totaal worden 98 van de 104 cases verwerkt binnen de analyse. Deze wijken krijgen een type toegewezen. Dit betekent dat 6 wijken niet meegenomen worden. Immers 3 wijken bleken in de clusteranalyse als 'missing' te worden aangemerkt en nog eens 3 wijken vielen in de clusteroplossing onder andere clusters.

Nu we bepaald hebben dat we twee typologieën verkrijgen, willen we weten hoe we deze inhoudelijk kunnen karakteriseren. Om deze vraag te beantwoorden zetten we de bewaarde clusteroplossing (de twee typologieën) af tegen de oorspronkelijke variabelen. De SPSS-functie *Compare Means* geeft voor de twee clusters de gemiddelde waarden van de betreffende variabelen. Aangezien we de variabelen "% allochtonen" en "% inkomensontvangers met uitkering" uit de clusteranalyse hebben gelaten, maar toch willen weten of deze onderscheidend werken op de verschillende clusters, nemen wij ook deze variabelen mee in de vergelijking. Daarnaast zijn we bij de profilering geïnteresseerd in de vraag of de situationele typologieën een onderscheid laten zien op de variabelen "inbraakrisico 1997" en "inbraakrisico 1998". Ook deze variabelen worden meegenomen in de vergelijking.

Net als bij de conventionele profielanalyse ligt het breekpunt op het aspect *stedelijkheid*. De marge tussen de beide profielen wordt echter meer nauwkeurig gelegd op de waarden 3 en 4 (respectievelijk *matig stedelijkheid* en *weinig stedelijkheid*). Zo ondervangt profiel 1 met een gemiddelde stedelijkheidswaarde van 4,57 de *weinig* en *niet stedelijke* gebieden. We zouden dit samen, net als bij de traditionele vorm, *niet stedelij-*

[1] Het ontbreken van bouwkundige gegevens in de gegevensbestanden van de politie en het CBS moet als een groot gemis worden ervaren dat een meer fundamenteel inzicht in met name de situationele kenmerken van woninginbraak in de weg staat. Door het invoeren van het door Bert Adriaens (elders in dit Tijdschrift) besproken *Informatiemodel Woninginbraak* hoopt de IMA-werkgroep dit gemis in de tweede fase van het project te compenseren.

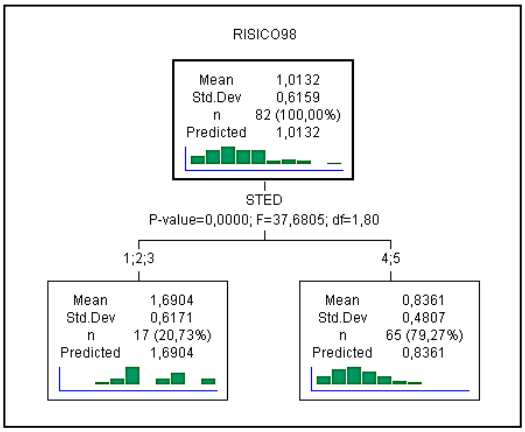
ke gebieden kunnen noemen. Daarentegen komt de gemiddelde waarde van profiel 2 uit op 2,70 hetgeen betekent dat de overige 3 gebieden hier vertegenwoordigd worden*:* de *zeer sterk, sterk* en *matig stedelijke* gebieden.

Reeds eerder is aangegeven dat 98 van de totaal 104 wijken in Limburg-Zuid binnen de analyse zijn verwerkt. Drie wijken bleken in de clusteranalyse als 'missing' te worden aangemerkt en drie andere wijken vielen op hun beurt in andere clusters dan de door ons geselecteerde. Om nu alle wijken in de gevonden clusteroplossing in te delen maken we gebruik van discriminantanalyse. Van de 98 wijken werden 88 gebieden gekaderd onder het eerste cluster en 10 wijken onder het tweede. We kunnen stellen dat deze twee type wijken sterk van elkaar verschillen (Wilks' Lambda= 0,000). Het model dat ontwikkeld is laat zien dat de voorspellingen voor 99% juist zijn. Slechts één case is 'misclassified'. Als we voorspelling laten toepassen op de zes overgebleven wijken dan zien we in profiel één een totaal van 93 wijken en in profiel twee een totaal van 11 wijken.

Risico-analyse

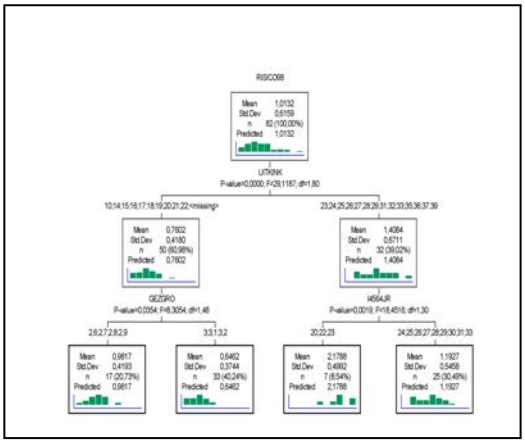
Naast het maken van typologieën is het ook mogelijk om op basis van situationele gegevens tot risicoprofielen van buurten te komen. Hiervoor maken we gebruik van segmentatietechnieken zoals Chaid. Deze technieken gaan actief op zoek naar díe variabelcombinaties die extreem hoge of lage scoren laten zien op een variabele als inbraakrisico. Natuurlijk is het mogelijk om hiervoor de gevonden situationele typologieën te gebruiken (deze vertonen immers extreme waarden op de inbraakrisico's van beide jaren), maar het is ook mogelijk om hier een minder geavanceerde techniek voor te kiezen. Deze minder geavanceerde techniek is weliswaar minder informatief, maar relatief eenvoudig door de gemiddelde misdaadanalist uit te voeren.

Figuur 1 Risicoprofiel en stedelijkheidsklassen. Bron CBS en HKS



Wanneer we al onze situationele wijkvariabelen als onafhankelijke variabelen (voorspellers) in een Chaid-model stoppen en het inbraakrisico 1998 als afhankelijke variabele (target) aanmerken, dan blijkt wederom dat de variabele stedelijkheid de meest sterke voorspeller is. Bij 82 wijken was het mogelijk om een inbraakrisico te berekenen. In 1998 was dit risico gemiddeld 1,01%. Wanneer we echter onderscheid maken tussen stedelijke en niet stedelijke gebieden (categorieën 1-3 versus 4-5) dan zien we dat in de stedelijke gebieden het inbraakrisico in 1998 1,69% bedroeg. In de niet stedelijke gebieden was deze daarentegen 0,83%. Deze uitkomst is geheel in de lijn met de uitkomst van de situationele typologieën. Hieruit blijkt dus dat deze typologieën goed gebruikt kunnen worden om risicoprofielen van wijken te bepalen.

Figuur 2 Risicoprofielen en sociaal-demografische factoren. Bron: CBS en HKS



Wat gebeurt er wanneer we de variabele 'stedelijkheid' buiten de analyse houden? Welke risicoprofielen krijgen we dan te zien? Het antwoord op deze vraag staat in het tweede plaatje weergegeven. Hieruit blijkt dat de variabele die het sterkst met inbraakpercentage samenhangt

‘% uitkeringsontvangers’ is. In wijken waar relatief weinig uitkeringsontvangers wonen (10-22%) is het gemiddelde inbraakrisico laag; namelijk 0,76%. In de wijken waar het percentage uitkeringsontvangers hoger is (23-39%) is het inbraakrisico bijna het dubbele; 1,41%. Binnen de wijken met relatief veel uitkeringsontvangers, zien we dat de leeftijdopbouw van de wijk een sterke samenhang vertoont met het inbraakrisico. In de wijken met weinig uitkeringsontvangers, geldt dit voor de gezinsgrootte. Het laagste inbraakrisico (0,65%) vinden we in de wijken met relatief weinig uitkeringsontvangers en een hoge gemiddelde gezinsgrootte. Het hoogste inbraakrisico (2,18%) vinden we in wijken waar relatief veel uitkeringsontvangers wonen en weinig personen in de leeftijd 45 tot en met 64 jaar. Naar verwachting zijn dit de meest anonieme woonwijken, waar nauwelijks sprake is van betrokkenheid en sociale cohesie.

Conclusie

Net als alle andere sectoren binnen de Nederlandse maatschappij is het werk van de Nederlandse politie de laatste decennia steeds complexer en moeilijker geworden. Om haar grip op de werkelijkheid te behouden, ziet de politie zich dan ook genoodzaakt om steeds ‘slimmer’ te werk te gaan. Goede en gedetailleerde informatie vormt hierbij een krachtig middel. Het geeft de mogelijkheid om de schaarse mensen en middelen gericht en efficiënt in te zetten.

Data mining kan de politie helpen om belangrijke informatie uit haar gegevens te halen. Uit het IMA-project is immers gebleken dat de gegevenssystemen van de politie veel meer informatie bevatten dan tot nu toe werd aangenomen. Deze informatie kan niet op de conventionele manier worden ontsloten. Het vereist een nieuwe benadering waarbij innovatieve technieken worden gecombineerd met moderne principes voor databeheer.

Bij het toepassen van data mining hoeft de politie niet zelf het wiel uit te vinden. Zij kan immers haar voordeel doen met de kennis en ervaring die binnen het bedrijfsleven voorradig zijn. Vooral binnen de sector marktonderzoek is de laatste decennia veel kennis opgedaan met data mining. De Nederlandse politie zou van deze kennis kunnen profiteren.

Literatuur

SPSS inc., *Data Mining with Confidence*, USA, 1999
López, M.J.J. & B. Adriaens, *Inbraakanalyse. De mogelijkheden om met behulp van innovatieve misdaadanalyse tot vernieuwende inzichten te komen*, NPI/LZ: 2000

*) Drs. Manuel J.J. López werkt bij onderzoeksbureau Result Crime Management en is als extern adviseur betrokken bij het Expertisecentrum Woningcriminaliteit. Voor meer informatie: m.lopez@RCM-advies.nl

Bijlage VI Two Rivers

Bijlage VII Broodnodige intelligentie voor CRM

Bijlage VIII Massa-individualisering in het MKB

Het volledige rapport is te lezen op http://www.minez.nl/publicaties/pdfs/25R12.pdf, tottaal 119 pagina's

Massa-individualisering in het MKB: de stand van zaken

drs. A.M. Jansen
EIM / Handel & Distributie
Zoetermeer, december 1999`

Samenvatting

Dit onderzoek is de eerste inventarisatie van de bekendheid van het concept massa-individualisering en bruikbaarheid voor het MKB. Ruim 1.600 MKB-bedrijven zijn telefonisch geïnterviewd.
Opvallend is dat de term massa-individualisering maar bekend is bij een derde van de MKB-bedrijven. Het begrip blijkt voor vele interpretaties vatbaar, abstract en moeilijk uit te spreken, en is daarom niet goed communiceerbaar. Daarentegen als gevraagd wordt of men aan maatwerk doet en inspeelt op individuele klantwensen, geeft 80% van de MKB-bedrijven aan hiermee bezig te zijn. Dat leidt tot de conclusie dat de term onbekend is, maar het concept dat hierachter steekt - het inspelen op individuele klantwensen - veel draagvlak heeft in het MKB. De meeste bedrijven stellen dat maatwerk zonder meerkosten mogelijk is. Ook verwacht 43% van de bedrijven een grote dynamiek in de markt en veel veranderingen op korte termijn. Terwijl 31% van de bedrijven al producten of diensten hebben met een korte levenscyclus. Opvallend is ook dat 60% van de ondervraagde bedrijven na 1990 maatwerk is gaan leveren. Verder verwacht bijna de helft van de maatwerk leverende bedrijven dat dit in de komende drie jaar zal toenemen.
Het is niet alleen een kwestie van zelf produceren, maar vooral hoe het proces op basis van de klantenwensen georganiseerd kan worden. Dit kan bijvoorbeeld door onderdelen van diensten en producten van anderen in te kopen. Daarnaast vereist het niet alleen samenwerking tussen afnemers, bedrijven en toeleveranciers, maar ook met complementaire
bedrijven en andere schakels in de ketennetwerken. Op dit punt scoort het MKB in de enquête niet zo hoog. Het blijkt lastig om goede samenwerkingspartners te vinden en de samenwerking goed te regelen.
Met uitzondering van de ICT-sector en dienstverlening worden computerverbindingen nog maar in beperkte mate ingezet voor de communicatie met leveranciers en klanten. De automatiseringsgraad, met name het gebruik van Internet, netwerken en EDI, blijkt sectorbreed nog onvoldoende ontwikkeld. De digitale interactie tussen verschillende partijen
in de ketens en één-op-één relaties met klanten blijven daardoor achter.
Typische maatwerkklanten willen een hoge kwaliteit, goed advies en een goede service. Behalve om kwaliteit gaat het om interactie met de klant. Juist dit laatste aspect krijgt een impuls door electronic commerce. Door Internet als handelskanaal te gebruiken wordt de individualisering van de klantrelaties en de identificering van de klantbehoefte op grote schaal mogelijk. Klanten zullen ook steeds meer hun individuele behoefte kenbaar maken. Bedrijven zullen daarop moeten anticiperen. De vertaling van klantenwensen

in producten en diensten zal complexer worden en een beroep doen op de creativiteit en het innovatievermogen. Door de persoonlijke advisering en de service om producten heen kan toegevoegde waarde gegenereerd worden. Daar liggen kansen voor het MKB. Opvallend is dat maatwerk voor veel MKB-bedrijven vereist dat er beter gekwalificeerd
personeel wordt ingezet, omdat juist de werknemers de kennis hebben die nodig is voor het vertalen van het concept naar de praktijk. De beschikbaarheid van dit gekwalificeerde personeel blijkt een knelpunt voor de bedrijven te vormen. Dit wordt vooral genoemd in detail- en groothandel non-food.
Naast gebrek aan kennis blijken kosten van maatwerk, complexiteit van het organiseren van het bedrijfsproces, en moeilijk om de juiste toeleveranciers te vinden bij 30% van de bedrijven een belemmering te vormen om maatwerk zonder meerkosten te leveren. In het leveren van maatwerk liggen nieuwe kansen voor toegevoegde waarde en klantenbinding. Maar dan moeten die kansen ingevuld en benut gaan worden door flexibilisering, een verhoging van de services rond producten, modulaire productie en precies leveren wat de klant wil (zowel consumenten als andere bedrijven). Ook samenwerken met andere bedrijven of netwerken van bedrijven wordt van toenemend belang. Inzet van informatie- en communicatietechnologie is bij dit alles onontbeerlijk, evenals gekwalificeerd personeel. De uitdaging om flexibel, innovatief en snel in te spelen op de individuele klantenwensen ligt bij de bedrijven.

Bijlage IX Schatgraven in databases

Datamining: schatgraven in databases

Door: Arno Siebes

Met behulp van datamining wordt kennis gedestilleerd uit een grote hoeveelheid gegevens. De belangstelling voor deze techniek neemt, met de toename van databanken enorm toe. Hoe werkt het?

Abstract gesproken is 'data mining' het induceren van een model uit een database.(1) Concreter kunnen we het omschrijven als het ontdekken van strategische informatie in -grote- databases. Voorbeelden in de marketing zijn, natuurlijk, het vinden van profielen van klanten die met een grote mate van waarschijnlijkheid reageren op een direct mailactie en het partitioneren (in stukken knippen) van een klantenbestand aan de hand van hun waarschijnlijkheid om een nieuw product af te nemen. Marketing is echter maar een van de vele gebieden waar data mining kan worden toegepast. Populaire toepassingen lopen uiteen van het analyseren van productieprocessen tot het schatten van de kredietwaardigheid van klanten. Van het analyseren van foto's van de sterrenhemel tot het analyseren van het koopgedrag van klanten.

Een manier om data mining te bespreken is dan ook het presenteren van een aantal uiteenlopende voorbeelden. Echter, de kans is dan groot dat een toepassing die juist interessant voor de lezer is niet aan bod komt. Bovendien levert data mining strategische informatie op. Het is daarom moeilijk om inzicht te krijgen waar data mining met succes is in gezet en waar niet. In plaats van een dergelijke opsomming schetst dit artikel dan ook de praktijk van het data minen, en wel aan de hand van een voorbeeld uit de marketingpraktijk. Als het proces duidelijk is, maakt de lezer zelf wel uit of data mining een oplossing is voor zijn problemen.

Data mining en zijn hele context wordt in de literatuur meestal Knowledge Discovery in Databases (kdd) genoemd, [1], in dit verhaal wordt deze gewoonte gevolgd. kdd is dus hele proces om kennis uit data te destilleren en omvat dan ook aspecten als het opschonen van bestanden en het verrijken van bestanden. Veel van deze aspecten zijn oude bekenden in de praktijk van de databasemarketing en andere statistische analyse toepassingen, er zijn echter twee belangrijke redenen om deze aspecten toch in dit verhaal op te nemen. De eerste is dat het hele kdd-proces noodzakelijk is om succesvol te minen. De tweede is dat de meeste fouten die bij datamining gemaakt worden, juist fouten in deze context zijn.

Het kdd proces omvat drie fasen; data warehousing, data mining en de interpretatie en het gebruik van de resultaten. Aan de hand van een voorbeeld zullen we de drie fasen bespreken. Voordat we dit voorbeeld introduceren nog een waarschuwing. Het feit dat we drie fasen onderscheiden wil nog niet zeggen dat kdd een lineair proces is wat je in een keer van het begin tot het eind doorloopt. Vaak zal men gedurende de analyse zien dat een iets andere vraag beter zou zijn, of dat er nog wat extra informatie nodig is. In al zulke gevallen moet men enkele stappen terug doen in het proces.

Big Bucks Credit Card
Een van de producten van de bank Big Bucks Inc. (bbi) is een credit card. Op dit moment gebruikt nog maar een klein gedeelte van bbi's klanten deze kaart. De bank vermoedt dat een veel groter percentage van haar klanten profijt van een credit card zou kunnen hebben. Een

uitgebreid informatiepakket naar alle klanten sturen is veel te duur en daarom wil bbi gericht die klanten aanschrijven waarvan zij verwacht dat zij in redelijke mate geïnteresseerd zijn in het bezit van een credit card. Het vinden van een profiel van zulke klanten is een typische datamining taak.

Data Warehousing

Om te minen heb je twee dingen nodig, een mining vraag en een mining bestand. De mining vraag beschrijft in wat voor soort strategische informatie je geïnteresseerd bent. In ons geval zijn dit profielen van potentiële credit card gebruikers. Het mining bestand is de database waaruit je deze informatie wilt destilleren. Als je weet wat je mining vraag is, kun je het mining bestand gaan samenstellen. De eerste vraag die je je moet stellen is: 'Welke informatie over mijn klanten zou relevant kunnen zijn voor het beantwoorden van mijn mining vraag?' In ons voorbeeld kennis over het inkomens- en uitgavenpatroon, kennis over de gezinssamenstelling en zo voort.

Als de analyse fase met behulp van traditionele statistische technieken wordt uitgevoerd, dan dient men zich te beperken in het aantal factoren wat men meeneemt; regressie werkt niet goed met 150 variabelen. Als men gebruikt maakt van data miningtechnieken hoeft dat niet, sterker nog, je moet je juist niet beperken. Ten eerste omdat factoren die geen rol blijken te spelen er vanzelf wel uitvallen. Ten tweede, en nog veel belangrijker, omdat je met data mining op zoek bent naar onbekende informatie.

Het kan heel goed zijn dat zowel het soort auto waarin men rijdt als de sport die beoefend wordt als geïsoleerd gegeven niets zeggen over het al dan niet hebben van een credit card. Maar dat wil nog niet zeggen dat de combinatie van auto en sport geen bepalende factor is. Als je op basis van een eerste (correlatie-)analyse zowel auto als sport verwijderd, kun je een dergelijke combinatie nooit meer vinden.

Nadat je bepaald hebt welke informatie relevant zou kunnen zijn, moet deze informatie bijeengezocht worden. Vaak zal het verspreid staan over verschillende (productie)databases en soms zal informatie van elders aangekocht moeten worden.[1] Dit betekent dat de data uit deze verschillende databases bij elkaar gebracht moet worden.

Dit bij elkaar brengen van informatie is beslist geen triviale taak, de twee belangrijkste aspecten zijn het opschonen van databases en het koppelen van databases. Het is een ervaringsfeit dat alle databases vervuild zijn, de hoofdoorzaak hiervan ligt in invoerfouten. Deze fouten variëren van eenvoudige typefouten tot niet-ingevulde velden. Sommige van deze fouten zijn makkelijk te herstellen. Typefouten volgen een standaard patroon, letters worden omgedraaid, er wordt met 'dikke vingers' getypt of de positie van de vinger op het toetsenbord is verschoven. Als er standaardlijsten bestaan (zoals voor straat- en plaatsnamen) zijn zulke fouten makkelijk te herstellen.

Voor fouten in namen ligt dat al anders. De namen Blauw en Blaauw komen naast elkaar voor en het is niet gemakkelijk te zien wat de goede spelling zou moeten zijn. Nog moeilijker ligt het bij niet ingevulde velden. Je zou deze met behulp van een statistische analyse kunnen proberen in te vullen, maar daar moet je voorzichtig mee zijn. Immers je gaat het resultaat gebruiken om een model te maken, door gebruik te maken van statistische technieken; op zo'n moment ben je bezig het eindresultaat een bepaalde richting op te sturen.

Nadat we de individuele databases zoveel mogelijk hebben opgeschoond, moeten we de bestanden gaan koppelen. Dit koppelen kan overigens helpen individuele fouten te herstellen. Bijvoorbeeld als we een Blauw en een Blaauw hebben die in alle (of veel) andere aspecten hetzelfde blijken te zijn kunnen we gokken dat het inderdaad om dezelfde persoon gaat. Op dezelfde manier kan koppelen helpen om niet ingevulde velden alsnog in te vullen.

Behalve deze plezierige aspecten heeft het koppelen van databases helaas ook minder plezierige kanten. Het is namelijk een moeilijke klus. 'Dezelfde velden' hebben vaak verschillende namen in verschillende bestanden en dan ook nog vaak een verschillende lay-out. Is de geboortedatum bijvoorbeeld één veld of zijn het er drie. Bovendien kunnen we dezelfde naam voor twee attributen in twee bestanden tegenkomen die volkomen andere aspecten van dezelfde klant beschrijven. Om databases te koppelen moet je dan ook eerst goed uitzoeken hoe de 'matching' tussen de databases ligt. Als je dan echt gaat koppelen gooit de vervuiling in de individuele bestanden nog weer roet in het eten. Immers, geen databaseprogramma ter wereld weet dat Mao Tse Tung en Mao Ze Dong dezelfde persoon zijn.

Het is zaak zowel het opschonen als het koppelen met zorg te doen. Immers, zoals bij elk proces is het ook bij data mining: 'garbage in, garbage out.' Als je de bestanden gekoppeld hebt, is het zaak om na te gaan hoe betrouwbaar de uiteindelijke velden zijn. Net als traditionelere statistische technieken is data mining redelijk robuust tegen fouten in de data, maar als bekend is dat een bepaald veld veel fouten heeft is het niet echt zinvol deze in de analyse mee te nemen. Immers, resultaten die gebruik maken van dit veld zijn niet erg betrouwbaar en in de praktijk niet erg bruikbaar.

Al met al is het opbouwen van een mining-bestand dus een behoorlijke hoeveelheid werk. Als je dan ook verwacht dat je bepaalde mining-taken vaker gaat uitvoeren ligt het voor de hand om de organisatie en het beheer van zulke bestanden tot een aparte taak te verheffen. In dat geval spreekt men vaak van een data warehouse. Een data warehouse is een database waarin gegevens vanuit verschillende productiedatabases geïntegreerd opgeslagen worden ten behoeve van, bijvoorbeeld, data mining.

Data warehouses hebben een aantal voordelen voor data mining. Ten eerste kan er sneller met de analyse begonnen worden. Ten tweede zorgt het ervoor dat men de resultaten van analyses op verschillende momenten kan vergelijken. Immers, het warehouse zorgt ervoor dat het mining bestand elke keer op dezelfde manier is samengesteld.

Data Mining

Als het mining bestand klaar is, kan het minen eindelijk beginnen. Er is geen standaard interface voor data mining-tools. Ik zal in deze sectie data minen toelichten op de manier waarop dat met Data Surveyor(2) gaat; al was het alleen maar omdat ik daar het meeste ervaring mee heb. Data mining is in de inleiding al gedefinieerd als het ontdekken van strategische informatie. In het geval van Data Surveyor is dit geoperationaliseerd als het ontdekken van interessante deelgroepen.(3)
In ons voorbeeld is een groepen klanten met een relatief hoog percentage credit cards een voorbeeld van een interessante deelgroep. Voordat Data Surveyor interessante deelgroepen kan gaan zoeken zullen we eerst moeten specificeren wat een groep klanten interessant maakt. Een dergelijke specificatie heeft twee aspecten: Syntax: We moeten de vorm van mogelijk interessante deelgroepen specificeren. Dit gebeurt met behulp van descriptions.

Semantiek: We moeten specificeren wat de ene deelgroep interessanter maakt in vergelijking met de andere. Dit gebeurt met een kwaliteitsfunctie, die voor elke deelgroep de mate van belangrijkheid bepaald. We zullen nu beide aspecten verder toelichten aan de hand van ons voorbeeld.

Descriptions

Als een data mining systeem een interessante deelgroep vindt, zijn we natuurlijk niet geïnteresseerd in een opsomming van alle klanten in deze deelgroep. Een korte kernachtige beschrijving is in eerste instantie veel nuttiger; de technische term voor zo'n beschrijving is een description, bijvoorbeeld:

Auto = Volvo & Sport = Golf

is veel inzichtelijker als een opsomming van alle Volvorijders die golf spelen. Bovendien kan zo'n description direct worden opgevat als een 'query' op de database. Met andere woorden, als we willen weten wie de golfspelende Volvorijders zijn, dan kunnen we daar met een druk op de knop achterkomen. Het tweede grote voordeel van descriptions is dat de analist zijn domeinkennis kan gebruiken in de specificatie van welke descriptions potentieel een antwoord kunnen zijn. In ons voorbeeld weet bbi dat een description als

leeftijd in [18, 24] & geslacht = man

een aanvaardbare groep aanduidt, terwijl

leeftijd in {18, 32, 47, 56} & geslacht = man

dat niet doet. Het is immers aannemelijk dat jonge mannen een grotere voorkeur voor credit cards hebben dan gemiddeld, dat zo iets ook zou gelden voor een groep mannen van wat willekeurige leeftijden lijkt veel meer toeval. Bovendien levert de eerste description een duidelijke doelgroep voor de marketingstrategie, terwijl de tweede dat zeker niet doet.

Regio's bieden een ander voorbeeld. Stel je voor dat het gebruik van een credit card regioafhankelijk is; dit zou te maken kunnen hebben met de dichtheid van credit card-acceptanten. Als het systeem dan regio's vindt waarin het gebruik hoger ligt dan gemiddeld, moeten die regio's op een kaart er ook als regio's uitzien, als een soort cirkeltje. Zo is de regio Utrecht, Vleuten/de Meern, Nieuwegein, Houten wel aanvaardbaar, terwijl Nieuwegein, Boskoop, Asten, Lemmer dat niet is.

Tenslotte kun je Data Surveyor nog hiërarchieën meegeven. Met een hiërarchie geeft de analist bijvoorbeeld aan hoe de favoriete dranken van de klanten gegroepeerd mogen worden. Met andere woorden, het systeem mag descriptions maken die gebruik maken van:

favoriete drank = bier = {Heineken, Grolsch}

maar niet van:

favoriete drank = {Grolsch, druivensap}.

Kwaliteitsfunctie

>Nadat je de toelaatbare descriptions gespecificeerd hebt, moet je Data Surveyor vertellen wat zo'n descriptie interessant maakt. Dat doe je met behulp van een kwaliteitsfunctie. Dat is een functie die, gegeven een database, aan elke description een getal toevoegt. Hoe hoger het getal, hoe beter de kwaliteit van deze description.
In ons voorbeeld zijn we geïnteresseerd in profielen van credit card-gebruikers; we zijn op zoek naar groepen klanten met een (relatief) hoog percentage credit card-bezitters. Hoe hoger het relatieve aantal credit card-bezitters onder de klanten die aan een description voldoen, hoe beter die description is. Onze kwaliteitsfunctie moet dus van elke description twee dingen weten: hoeveel klanten voldoen aan deze beschrijving en hoeveel daarvan bezitten reeds een credit card. Door deze te delen hebben we dan een aanzet tot een kwaliteitsmaat.

Het is alleen nog maar een aanzet omdat we nog geen rekening hebben gehouden met de statistiek. Stel dat descriptions x1 en x2 beide een fractie 0.8 opleveren, maar dat x1 een uitspraak over 5 klanten doet, terwijl x2 over 10.000 klanten gaat. Dan is x2 natuurlijk een veel betere description als x1. Immers x2 biedt een veel betrouwbaardere uitspraak dan x1. Om rekening te houden met de betrouwbaarheid van uitspraken maakt Data Surveyor gebruik van betrouwbaarheids-intervallen. Het 95 procent betrouwbaarheidsinterval voor de fractie van x1 is [0.4, 1], terwijl die voor x2 [0.79, 0.81] is.

Behalve betrouwbaarheids-intervallen spelen bij data mining nog veel meer statistische aspecten een rol. Zo weten we dat hoe korter een description is, hoe betrouwbaarder de uitspraak is; dit volgt uit het Minimum Description Length Principle. Het voert te ver om in het kader van dit verhaal diep in te gaan op al deze statistische aspecten die samenhangen met kwaliteitsfuncties. Bovendien houdt Data Surveyor er al rekening mee voor de gebruiker. Zij die er meer van willen weten, worden verwezen naar [4].

Omdat statistiek in data mining zo'n belangrijke rol speelt is kan het slechts een hulpmiddel zijn voor de professionele gebruiker is. Immers, de marketeer die gebruik maakt van dit gereedschap zal genoeg van statistiek moeten weten om tot betrouwbare conclusies te komen

De Data Mining Taak
Nu de descriptions en de kwaliteitsfunctie gespecificeerd zijn, hebben we aangegeven welke deelgroepen van ons bestand interessant zijn. Het enige wat nu nog rest is wat we met deze deelgroepen willen. Er is een aantal mogelijkheden:
Ten eerste zou het kunnen zijn dat we alleen de description van de hoogste kwaliteit willen hebben. In ons voorbeeld omdat bbi alleen een mailing wil doen aan klanten met de hoogste kans dat zij op het credit cardaanbod in zullen gaan.

Ten tweede zou het kunnen zijn dat we de database - en daarmee de klantenkring - willen partitioneren (in stukken knippen). Dat wil zeggen, dat we alle klanten in homogene groepen willen onderverdelen. Alle klanten in een groep hebben daarbij dan dezelfde kans om het credit card aanbod te accepteren. Zo'n partitionering is bijvoorbeeld nuttig als we een mailing willen doen naar 20 procent van het klantenbestand en de return on investment willen maximaliseren.

In andersoortige toepassingen zouden we ook geïnteresseerd kunnen zijn in alle descriptions met een kwaliteit boven een gegeven minimum. Dit zou je bijvoorbeeld gebruiken om te bepalen welke producten in een winkel vaak tegelijk aangeschaft worden; de zogenaamde "Basket-Analysis". In ons voorbeeld is dit minder nuttig, omdat de klanten dan aan meerdere descriptions gaan voldoen. In de data mining taak moeten we dit laatste aspect, "wat willen we", dan ook expliciet specificeren.

Het vinden van resultaten
Als de data mining taak gespecificeerd is, kan het systeem aan de slag. Stel, we hebben gevraagd om de groep klanten met de grootste kans op het accepteren van ons aanbod; we zoeken de description met de maximale kwaliteit.
Het eenvoudigste algoritme wat deze taak oplost berekent gewoon de kwaliteit van alle descriptions, bepaalt welke de hoogste kwaliteit heeft en retourneert die als antwoord. Het voordeel van dit algoritme is dat het correct is; het levert inderdaad de gezochte description. Het nadeel is dat het zo verschrikkelijk lang duurt voordat we het antwoord krijgen, omdat het aantal descriptions dat we moeten beschouwen exponentieel is ten opzichte van de grootte van de database. Stel bijvoorbeeld dat we een database met 10 klanten hebben. Dan zijn er 2^10 - 1 = 1023 groepen klanten te verzinnen. Als we de kwaliteit van 1000 groepen per seconde kunnen berekenen, zijn we in iets meer dan een seconde klaar. Maar stel nu dat we 100 klanten hebben, dan loopt het aantal te verzinnen groepen klanten op tot:

2^100-1=1267650600228229401496703205375

Als we nog steeds de kwaliteit van 1000 groepen per seconde kunnen berekenen, dan kost het nu 10^19 jaar. Dat is langer als de leeftijd van het heelal.(4)

Kortom, alle descriptions afzoeken is niet te doen. We zullen slimmere zoekmethoden moeten gebruiken. In de kunstmatige intelligentie heten dat heuristieken. Het nadeel van een heuristiek is dat we niet kunnen garanderen dat we de groep met de hoogste kwaliteit hebben gevonden. Het voordeel is dat we een redelijk goed antwoord in relatief korte tijd krijgen.

Er zijn vele soorten heuristische zoekmethoden, zoals: Hill Climbers, Simulated Annealing, Taboo Search, Genetische Algoritmen en Neurale Netwerken. Welke techniek het beste is, hangt helemaal af van de specifieke data mining-taak en van allerlei karakteristieken van de mining database, zie [2]. Met andere woorden, het moet niet de taak van de gebruiker zijn om te kiezen welke techniek wordt toegepast, dat moet het data mining-systeem maar bepalen. Bovendien is het niet interessant, de resultaten die gevonden worden wel.

Een auto is misschien de beste analogie. Wat voor soort motor er in de auto moet zitten, hangt af van wat je met die auto wil. Een vrachtwagenmotor presteert niet in een raceauto, maar andersom komt er ook niets van terecht. Een data mining-tool is een auto met een heleboel verschillende motoren die de juiste motor op het juiste moment inzet.

Interpretatie en gebruik
Als het data mining-systeem de gezochte descriptions gevonden heeft, zijn we er nog niet. Immers, de resultaten moeten nu gebruikt gaan worden. Dat is minder eenvoudig als het lijkt. Stel dat bbi de profielen wil gebruiken om haar eigen klanten aan te schrijven. Stel bovendien dat ze tot nu toe alle klanten die om zo'n kaart gevraagd hebben er een gegeven heeft. Dan kan zij de profielen rechtstreeks gebruiken; immers de mining database gaf een getrouwe weergave van haar klanten.
Als zij echter in het verleden klanten een kaart geweigerd heeft omdat zij ze als niet-kredietwaardig beschouwde, ligt het anders. Immers, de kaarthouders die we gebruikt hebben om te bepalen wie zo'n kaart zouden willen hebben, zijn niet alleen kaarthouder maar ook kredietwaardig. Als we ons dat van te voren bedacht hadden, dan hadden we alleen in

een bestand van kredietwaardige klanten gezocht en was er geen vuiltje aan de lucht. Als we het ons alleen maar achteraf realiseren, kunnen we de resultaten (met wat voorzichtigheid) nog wel gebruiken maar moeten we op een andere wijze, (handwerk) voor zorgen dat we alleen kredietwaardige klanten aanschrijven (data mining is overigens wel geschikt hiervoor).

Als bbi z'n profielen zou willen gebruiken om niet-klanten tot klant te maken via een lucratief credit card aanbod, ligt de zaak nog gecompliceerder. De klanten van de bank zijn tenslotte geen onafhankelijke steekproef uit de Nederlandse bevolking. Al was het alleen maar omdat bbi een beleid toepast bij het al dan niet accepteren van nieuwe klanten. We kunnen er niet zo maar vanuit gaan dat de resultaten die we voor ons klantenbestand gevonden hebben ook gelden voor de Nederlandse bevolking. Dit is een veel moeilijker probleem, ook als we het ons van te voren al bedacht hadden. Immers, waar halen we de gegevens vandaan over niet-klanten? Een mogelijkheid is om vooraf een steekproef uit de niet-klanten te nemen. Zo zou men enige tijd iedereen die er om vraagt tot klant kunnen maken. Potentieel is dit een kostbare oplossing, maar bij zorgvuldig gebruik hoeft het niet excessief te zijn.

Oplossingen achteraf bestaan ook. Zo kun je natuurlijk je profielen testen op de markt. Je neemt dan een steekproef van de Nederlandse bevolking die aan je profiel voldoet en je test of de response binnen deze groep overeenkomt met de response die je vanuit je klantenbestand verwacht. Onder statistici, econometristen en data miners heet dit probleem reject-inference. Het laatste woord over hoe dit probleem het beste kan worden opgelost is zeker nog lang niet gezegd.

Conclusie

Om succesvol te kunnen werken is een goed inzicht in de markt nodig. Data mining is een krachtig hulpmiddel om dat inzicht te verwerven. De resultaten die als descripties gegeven worden zijn direct te begrijpen en te gebruiken.

In de kracht van data mining schuilt ook het gevaar. Het is heel gemakkelijk om met een data mining-tool elk willekeurig bestand te analyseren. Echter, om tot betrouwbare en dus bruikbare resultaten te komen, moet het hele kdd proces met zorg doorlopen worden. Als dat gebeurt, is data mining een hulpmiddel wat zich zelf terugverdient.

De vraag is of data mining daarmee beter is dan de andere in zwang zijnde technieken. Er ontbreekt de ruimte om op deze vraag in te gaan. De ervaring leert echter dat als een database voldoende rijk is, je met data mining relevante en nieuwe strategische informatie boven water haalt.

Bijlage X Datavoorbereiding bepaald kwaliteit

Datavoorbereiding bepaalt kwaliteit dataminingprocessen

'DE VOORBEREIDINGSFASE IS HET MOEILIJKSTE GEDEELTE VOOR MENSEN DIE NOG GEEN OF WEINIG ERVARING MET DATAMINING HEBBEN'

'DE KEUZE VAN DE DATA DIE BIJ DE ANALYSE WORDT BETROKKEN, HEEFT EEN DUIDELIJKE INVLOED OP WELK MODEL GEVONDEN KÁN WORDEN'

'HET DATAMININGPROCES IS NIET PER SE AFGESLOTEN, WANNEER HET MININGRESULTAAT IS BEREIKT'

Datamining is een iteratief proces, zeker in de voorbereidende fase. Die fase is vaak kostbaar, omdat de data-analist 'onproductief' bezig is met uitzoeken van de benodigde data en het transformeren ervan tot een geschikt bestand. Ervaring, uniforme beschrijving, evaluatie van de data-analyse en de mining zelf maken het totale proces efficiënter en effectiever. Dat laat onverlet dat menselijke keuzes in het begin van het traject de mininguitkomst bepalen.

Om een juist dataminingmodel te vinden voor het leveren van een nuttig pattern (betekenisvolle samenhang tussen data) moet een voorbereidingsfase plaatsvinden. Hierbij hoort het kiezen van een initieel databestand en het voorbereiden van de data, zodat deze geschikt zijn voor analyse. Voor de volledigheid: het dataminingsproces volgt in zijn voobereidende fase de volgende stappen:

- data verzamelen,
- data structuren en classificeren,
- data rationaliseren,
- data objectiveren.
- data analyse.

Tijdens practica met dataminingtools en afstudeerwerk is echter gebleken, dat deze voorbereidingsfase het moeilijkste gedeelte is voor mensen die nog geen of weinig ervaring met datamining hebben. De dataminingtools zelf bieden hier weinig directe steun. Dezelfde ervaring is ook in de praktijk opgedaan met de opzet van dataminingprojecten.
Met het kiezen van het initiële databestand wordt vastgelegd welke data überhaupt getoetst worden op hun relevantie voor het vinden van een geschikt dataminingmodel. Dat is moeilijker dan het op het eerste gezicht lijkt, maar wel belangrijk, want hiermee wordt ook de uitkomst van het dataminen in een bepaalde richting 'gedrukt'.

Beperkte input
Bij open datamining wordt geen bepaalde vraagstelling gevolgd; men wil gewoon betekenisvolle samenhangen vinden tussen data, want die zijn op voorhand niet evident. Als er al een vermoeden bestaat van de samenhang, is het miningproces niet meer open, maar gericht. Omdat men in de open variant toch niet weet wat men zoekt, wordt vaak genoegen genomen met beschikbare bestanden die relevant lijken voor de te nemen beslissingen.

Als men bijvoorbeeld beslissingen over verkoopstrategieën wil onderbouwen, moeten natuurlijk modellen afleidbaar zijn die over verkoop informatie verstrekken. Data over koopgedrag van individuele klanten kan relevante informatie leveren, maar is als dataminginput toch beperkt. Het aantal personen dat bij de huishouding van een klant hoort, kan im-

mers ook van invloed zijn op diens koopgedrag of het aantal kinderen van een bepaald leeftijd, of de inkomsten van de huishouding in het geheel.
Het is heel moeilijk om vast te stellen of de juiste datasets gekozen zijn. Er komt immers altijd een resultaat uit de dataminingexercitie. Geen pattern vinden (van betekenisvolle samenhangen tussen data(waarden), is óók een van de mogelijke resultaten. Uit elkaar lopende patterns kunnen juist zijn en in dezelfde richting wijzende ook. Het is best mogelijk dat precies één feature (een eigenschap met een voorspellende waarde) niet meegenomen is in de analyse, maar dat deze het juiste pattern - de in realiteit bestaande samenhang tussen feiten voor het aankoop van een bepaald product in een bepaald situatie - zou opleveren. Een op het eerste gezicht waardeloze (weinig voorspellende waarde) feature heeft de potentie om in combinatie met een andere feature van grote voorspellende waarde te zijn. Een feature als 'Geslacht' is daar een goed voorbeeld van. Uit het merendeel van de uit te voeren data-analyses wordt deze feature een random voorspellende waarde toegedicht, met andere woorden: geen. Echter, bij het toevoegen van deze feature aan het voorspellergezelschap blijkt deze van grote waarde te zijn in combinatie met andere features. In dat geval dient de feature als zogenaamde waterdrager om de rest van het featureteam een toegevoegde voorspellende waarde mee te geven. Het probleem met de waterdragers is het herkennen van het nut die deze eventueel voor het gehele dataminingproces kunnen hebben. Het bepalen van deze (verborgen) potentiële voorspellende waarde van features wordt nog niet geautomatiseerd ondersteund en betekent extra inspanning voor de dataminingconsultant.
Bij gerichte datamining is de stap van het kiezen van het initiële databastand iets anders ingericht. Er moet worden vastgelegd welke vermoedens men wil bevestigen of juist uitsluiten. Dat vermoeden wijst de weg naar een bepaald inhoudelijk domein van de data. Als het de verkoop van een prijzig sportartikel betreft en men wil uitvinden of het lidmaatschap van een sportclub invloed op het koopgedrag zal hebben of niet, moet er data beschikbaar zijn over leden van sportclubs (over hun aankopen van vergelijkbare prijzige sportartikelen) en data over aankopen van niet-leden.

Beperkingen
Wellicht overbodig te melden dat de gerichtheid van zo'n dataminingquery een vooroordeel bevat, namelijk dat het lidmaatschap van de sportclub invloed heeft. Dat beperkt veel onderzoekers in de samenstelling van hun beslissingsboom en trainings- en testbestanden. Wellicht is een ander feature dan het lidmaatschap van een sportclub van veel grotere invloed, maar wordt die niet gevonden, om-dat men bij de datasetsamenstelling te gefocused is op de vraagstelling.

Of een geschikt model gevonden is, kan door vergelijking met de realiteit worden vastgesteld (test-bestand). Er zijn methoden om de waarschijnlijkheid van een model te berekenen, maar daar gaat het even niet om. Wij willen benadrukken, dat de keuze van de data die bij de analyse wordt betrokken, een duidelijke invloed heeft op welk model gevonden kán worden. Deze keuze maken is voor een aanzienlijk deel subjectief en op ervaring gebaseerd: de keuze van de data (de identificatie van de te mi-nen data) eist ervaring met datamining. In het bijzonder is ervaring vereist met het gedrag van het algoritme en het zogenoemde datamessaging, het vastleggen van de value ranges en hun verhoudingen.
Om bij de keuze van de features zo veel mogelijke objectiviteit te bereiken, kunnen features 'onherkenbaar' worden gemaakt voor en door een consultant. Middels het het anonimiseren van eigenschappen bijvoorbeeld wordt een consultant niet verleid om een databestand zó aan te passen, dat iets wat hij vanzelfsprekend vindt opeens wél binnen het voorspelde potentieel valt. Als de consultant geen gevoelens heeft bij de gegevens, staat hij/zij anders tegenover het resultaat van het miningsproces.

Naast de features hebben ook de waarden van de data die bij een feature horen - beter: hun betrokken waardebereik (de range van waarden die als geldig wordt beschouwd voor het betreffende dataminingmodel) - invloed op de kwaliteit van het model. Als bijvoorbeeld een verkoopgebied wordt uitgebreid kunnen er andere demografische structuren van potentiële klanten bijkomen (features), en dat kan aanpassing van de ranges van de datawaarden noodzakelijk maken. Zo zou het kunnen zijn, dat er nu ook welgesitueerde families in het nieuwe marktgebied wonen en dat die hun kinderen langer financieel ondersteunen, hetgeen invloed heeft op het koopgedrag. Dat betekent dat de leeftijd van kinderen een groter waardebereik moet hebben in het onderhavige model.
Het kan ook zijn dat in de nieuwe situatie het aantal kinderen in de huishouding helemaal geen invloed heeft. Dan valt een feature weg, maar slechts voor het nieuwe verkoopgebied. Het tot dan toe geschikte dataminingmodel zal geen bruikbare uitkomsten leveren als het niet wordt aangepast.
Het mag duidelijk zijn: elke situatie is anders, maar ervaring scheelt. Het helpt om in iedere situatie een iteratieve aanpak te kiezen (data prepareren, model maken, toetsen, vergelijken met de realiteit, model aanpassen) en langzamerhand zekerder te worden in het gebruik van datamining ter ondersteuning van beslissingen.

Bronnen beoordelen
Laten we aannemen dat de data voor het dataminen inhoudelijk geïdentificeerd zijn. Men weet welke features met data 'bediend' moeten worden. De volgende stap is te besluiten uit welke bronnen de betreffende data gehaald worden. Die bronnen moeten dus beoordeeld op geschiktheid. Als men geluk heeft, zijn de databases en andere databestanden goed gedocumenteerd en komen de fysieke bronbestanden overeen met de ooit opgemaakte documentatie daarover. Onze ervaring leert echter, dat men toch meer bestanden met 'dezelfde' data tegenkomt. Meestal zijn de bestanden niet gesynchroniseerd, dus de data verschillen in actualiteit. Soms ontbreken data of er zijn NULL-waarden gebruikt op plekken waar deze eigenlijk niet horen of datawaarden zijn niet plausibel, et cetera. Men zal moeten beslissen welke van de bestanden worden overgenomen in de dataminingbestanden - na eventuele verbetering van de kwaliteit of aanvulling uit ander bestanden. Daartoe hoort ook datarationalisatie. Daarmee worden bijvoorbeeld dubbele records uitgesloten die in feite dezelfde informatie betreffen, maar er syntactisch verschillend uitzien. Bijvoorbeeld dezelfde persoon, maar anders geschreven.

Als er data uit verschillende bestanden overgenomen moeten worden, is het belangrijk om erop te letten dat deze dezelfde actualiteit hebben - ook als het om historische gegevens gaat. Als verschillend geüpdate gegevens met elkaar geassocieerd worden (bijvoorbeeld door middel van een join via dezelfde waarden van een bepaald attribuut dat in tabellen van twee databases voorkomt) kan dat tot foutieve informatie leiden. Zo lang zo'n foutief aandeel in het te minen databestand klein is, is de invloed op het resultaat te verwaarlozen. Daar mag alleen niet te snel genoegen mee worden genomen, want het is slecht te voorspellen wanneer de balans in het tegendeel overgaat.
Desalniettemin, soms is de historische ontwikkeling niet meer te achterhalen, omdat bijvoorbeeld geen tijdswaarden opgeslagen zijn. De noodzaak daartoe ontbrak klaarblijkelijk in het verleden. Ook als er wel sprake is van een systematisch opgebouwd digitaal archief, loopt men bijna zeker tegen twee problemen op:

1 - verschillen in de actualiseringgraad van verschillende bestanden en onzekerheden hierover;
2 - het ophalen van data uit oude files en het re-engineren van hun betekenis is gewoonlijk tijdrovend en soms ook erg kostbaar, omdat bijvoorbeeld specialisme nodig is, die schaars is geworden (oude systeemkennis).

Men moet redelijk zeker weten dat de bestanden de moeite waard zijn. Soms laten historische data zich makkelijker en betrouwbaarder uit analoge documenten reconstrueren. Waarschijnlijk zijn zo geen grote historische databestanden op te bouwen, tenminste niet voor één dataminingproces. Als men echter slechts kleine bestanden ter beschikking heeft, is het analytische potentieel beperkt. Het (additioneel) kopen van betrouwbare databestanden is in zulke situaties een alternatief. Er moet rekening mee worden gehouden, dat formatteren van de data eveneens kostbaar is.

Evaluatie
Wij hebben het hier niet over de noodzaak van terugkoppeling om bijvoorbeeld het partitioneren van datawaarden te evalueren en voor de volgende keer te verbeteren. Waar het om gaat, is dat een evaluatie er zeer waarschijnlijk toe leidt dat de inhoudelijke doelstellingen van het dataminen aangepast of uitgebreid worden. Een gevolg hiervan zou kunnen zijn, dat nieuwe (bijkomende) data geïntegreerd en dus ook voorbereid moet worden. Het model moet aangepast worden; kortom, het hele proces wordt opnieuw doorlopen.

Rekening houdend met dit feit van inhoudelijke aanpassingen en herhalen van kostbare voorbereidende en analytische werkzaamheden, helpt het om de bevindingen bij iedere analyse van gegevensbestanden te documenteren. Wij verwijzen hiervoor naar het artikel over het voorbereidende werk voor datawarehouses (Beyond 4, december 2000). De inspanningen die nodig zijn voor een data-analyse om bruikbare data te identificeren en het voorbereiden van de datasets moeten niet voor elke vraagstelling opnieuw worden gedaan. Datamining en andere analyses zijn zo 'gewoon' aan het worden in de bedrijfsvoering, dat ze dus efficiënt en duurzaam ondersteund moeten worden vanuit een visie op de totale analysebehoeftes van de onderneming.
Stel, er is een datawarehouse opgebouwd en het wordt ook onderhouden. Er zijn al meermalen succesvolle voorspellingen gedaan met de data uit het warehouse via een bepaald datamining model. Maar als een bruikbaar model voor een bepaalde vraagstelling gevonden is, betekent dat niet, dat dit model altijd geldig blijft. In beginsel is hier geen verschil met dataschema's; het miningmodel is in principe een dataschema voor miningdoeleinden. Het is een abstractie, een model van de werke-lijkheid. Als daarin - in de werkelijke in- en externe factoren van een onderneming (of andere grootheid) - veranderingen optreden, moet het dataschema aangepast worden, ongeacht of het een model van een operationele database is of van een dataminingbestand. Een bestaand model moet dus op zijn geldigheid worden geëvalueerd als men het over grotere tijdsperioden gebruikt. Deze evaluatie van de betrokken data moet over de featurekeuze gaan, maar ook over het waardebereik en de onderverdeling van de waarde-ranges.
Prof.dr.ir.Waltraud Gerhardt

Bijlage XI Datamining Cursus

TECHNIEKEN VOOR DATAMINING

Tutorial en hands-on

Een 4-daagse cursus op 18-19-20 maart en 11 april 2003 in Delft

DOELGROEP

Personen die in kort tijdsbestek kennis willen opdoen over enige theoretische grondslagen en toepassingen van technieken voor datamining. De cursus is zowel bedoeld voor personen met een zakelijk motief als voor personen die vanuit een technisch perspectief zijn geïnteresseerd in data analyse.

OPLEIDINGSNIVEAU

Academisch/HBO niveau óf een door ervaring verkregen gelijkwaardig kennisniveau. Er wordt uitgegaan van basiskennis van de wiskunde en statistiek.

DOELSTELLING

Inzicht geven in het proces van datamining, kennis verstrekken over het bewerken van verzamelingen data, het prepareren van data, het moduleren van de data en het toepassen van diverse technieken voor datamining, alsmede het evalueren van de resultaten. In het bijzonder wordt beoogt inzicht te geven in en het leren toepassen van de verschillende modelleringsmethoden die daarbij een rol spelen. Hiertoe behoren methoden en algoritmen voor classificatieregels, voor associatieregels en voor regressieregels. Na afloop van de cursus bent u in staat het geleerde in uw eigen werksituatie toe te passen.

TOELICHTING

Datamining is het proces van het analyseren van data vanuit verschillende perspectieven met als doel relaties en patronen in de data samen te vatten in bruikbare informatie. Anders gezegd, datamining is het niet triviale proces dat tot doel heeft geldige, nieuwe, potentieel bruikbare en begrijpbare relaties en patronen uit data te vergaren en te representeren.

Het ontdekken van relaties en patronen (klassen, clusters, associaties en reeksen) is, in traditioneel zakelijke terminologie, van groot belang voor bijvoorbeeld het identificeren van klanten- of marktsegmenten en het introduceren van nieuwe diensten of producten, alsmede het kunnen voorspellen van opbrengsten en type transacties gerelateerd aan bijvoorbeeld de investeringen in die nieuwe diensten of producten.

CURSUSOPZET

De cursus beslaat vier dagen, waarvan drie aaneengesloten dagen en een terugkomdag na enige weken. Een cursusdag bestaat uit een ochtendprogramma en een middagprogramma. In de ochtend wordt nieuwe theorie behandeld die aan de hand van een uitgebreid voorbeeld wordt toegelicht. In de middag worden opgaven aangereikt die u in groepjes kunt uitwerken. De dag wordt afgesloten met reflectie op de theorie en de opgaven. Op 19 maart is er een avondprogramma met een diner, een gastspreker en tijd voor reflectie.

Verder kiest u tijdens de drie cursusdagen een voorbeeldprobleem uit uw eigen praktijk welke u op eigen locatie verder uitwerkt en waarvan de resultaten tijdens de terugkomdag aan elkaar worden gepresenteerd.

De volgende onderwerpen komen aan bod:

Data voorbereiding

Datatypen en missende attribuutwaarden
Oplossingsmethoden voor missende attribuutwaarden
Inconsistente data

Modelleringsmethoden

Classificatiemethoden: beslisbomen, k-naaste-nabuur classificatie
Cluster methoden: k-means clustering, fuzzy c-means clustering
Regelgeneratie: associatieregels, beslisregels, regressieregels

Evaluatiemethoden

Cross-validatie
Visualisatie van resultaten

Schaalbaarheid

Gevolgen van zeer grote datasets voor de behandelde technieken
Mogelijke oplossingen zoals: k-d tree, k-nn condensing
Voorbeelden van algoritmen die speciaal zijn ontworpen voor grote datasets

VERZORGENDE INSTANTIE

De Sectie Elektrotechniek van Stichting PATO in samenwerking met de Technische Universiteit Delft.

CURSUSLEIDING

Prof.dr.ir. E. Backer (TU Delft)

DOCENTEN

Prof.dr.ir. E. Backer
Dr.ir. M.J.T. Reinders
Dr. C.J. Veenman (allen TU Delft)

DATA/PLAATS

De cursus wordt gehouden op 18-19-20 maart en 11 april 2003 in Delft. De cursus begint om ca. 9.00 uur en eindigt om ca. 17.00 uur. Op 19 maart is er een avondprogramma.

KOSTEN

De deelnamekosten bedragen **€ 1.985,-** per persoon (geen BTW verschuldigd). Inbegrepen zijn de kosten van het cursusmateriaal, de lunches en koffie/thee.

AANMELDING

U kunt zich direct aanmelden door het in de website opgenomen aanmeldingsformulier ingevuld aan ons te retourneren. Aanmeldingen bij voorkeur inzenden vóór 26 februari 2003.

ANNULERING

Bij annulering tot twee weken voor aanvang van de cursus wordt € 500,- in rekening gebracht, daarna 50% van het cursusgeld. Bij plotselinge verhindering kan men zich zonder extra kosten door een collega laten vervangen. De Stichting PATO behoudt zich het recht voor de cursus te annuleren bij onvoldoende aanmeldingen.

INLICHTINGEN

Mw. drs. M.A.C.H. Bergmans, Stichting PATO, tel.: (070) 3644957, info@pato.nl

Bijlage XII Het gebruik van data voor beslissingsondersteuning

A.E. Eiben
Bedrijfswiskunde en Informatica
Vrije Universiteit Amsterdam
www.cs.vu.nl/~gusz

Inleiding
Computers worden voor een groot deel gebruikt voor “simpele” taken: kantoorautomatisering, (b.v. tekstverwerking), administratie (b.v. van salarissen en betalingen) en het afwerken van transacties, (b.v. aankopen in een supermarkt of telefoongesprekken). Een meer geavanceerd soort gebruik betreft geautomatiseerde beslissingsondersteuning. In deze applicaties is er altijd iets waarover te beslissen valt – en dus beslist moet worden.

Denk bijvoorbeeld aan een transportbedrijf dat dagelijks een aantal vrachtwagens laat rijden. Het opstellen van routes voor de vrachtwagens behoeft beslissingen (wie, wanneer, waar naar toe). Deze beslissingen kunnen heel moeilijk zijn doordat het betreffende probleem te complex is om het goed te overzien. Tevens kunnen er zeer grote verschillen zijn tussen verschillende beslissingen, de ene route kan veel langer zijn en dus meer tijd en geld kosten. Het is dus zaak om de routes zo goed mogelijk uit te kienen. Computers kunnen dan hulp bieden door de mogelijkheden door te rekenen en optimale, of in ieder geval zeer goede, routes aan de gebruiker voor te stellen. Deze tak van sport is bekend als beslissingsondersteuning door middel van optimalisatie. Tijdens deze, derde, les van de introductiecursus wordt dit geïllustreerd door middel van een computerprogramma dat de kortste route door een aantal steden berekent.

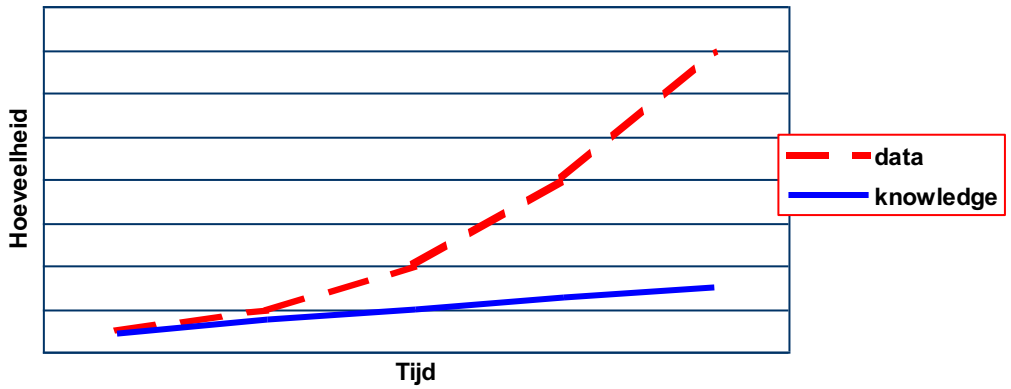
Een andere aanpak betreft het ondersteunen van beslissingen waar het overwegen en vergelijken van verschillende opties van meerdere criteria afhangt. Is bij het transportbedrijf maar één kwaliteitsmaat (de lengte van de routes), bij het kopen van een nieuwe auto komt er meer bij kijken: prijs, benzineverbruik, veiligheid, betrouwbaarheid, imago, etc. Voor deze zogenaamde multi-criteria situaties zijn er weer andere soort beslissingsondersteunende systemen. Een voorbeeld wordt tijdens de cursus toegelicht en je mag het programma mee naar huis nemen.

Het centrale thema van de derde les van de introductiecursus is weer een ander verwant gebied: beslissingsondersteuning door middel van data analyse. Om een voorbeeld te geven, denk aan een marketing afdeling van een grote firma die producten verkoopt aan consumenten. Voor het opzetten van een direct mail campagne moet er een lijst van geadresseerden samengesteld worden. Een goede lijst bevat namen die met een hoge kans positief reageren op de campagne. Zo’n lijst valt op te stellen op basis van gegevens van eerdere campagnes. Daarin staan gegevens van de gemailde personen en hun reactie (of het uitblijven daarvan). Een ervaren deskundige, of een goed computerprogramma, kan uit de historische data kennis destilleren, b.v. in de vorm van “mannen tussen 15-25 jaar reageren meestal positief”. Deze kennis kan dan worden ingezet bij het selecteren van namen voor een nieuwe campagne. Tijdens de tweede helft van de derde les krijg je ruim de mogelijkheid om met een data-analysesysteem te werken.

De kracht van data
In dit deel wordt er dieper ingegaan op het benutten van gegevens ten behoeve van genereren van kennis, en uiteindelijk het ondersteunen van (betere) beslissingen.

1. De kenniskloof

In de vijftiger jaren zijn commerciële en overheidsinstellingen begonnen met het verzamelen van steeds meer gegevens. Er wordt echter slechts een fractie van deze gegevens gebruikt, d.w.z. daadwerkelijk omgezet in zinvolle en bruikbare kennis. Er zijn bijvoorbeeld gegevens, die klinkklare feiten over consumenten en hun transacties representeren. Deze kunnen worden omgezet in bepaalde kennis, zoals de stelling dat de consumentengroep X een grote interesse heeft in het product of de dienst Y. Zulke "kennis brokstukken" zijn echter niet makkelijk met de hand te halen uit de enorme beschikbare gegevensbergen. Er is gigantische kloof tussen beschikbare en gebruikte gegevens en deze kloof is alleen maar aan het verbreden.



Figuur 1: Illustratie van de kenniskloof

2. Machinaal leren

De academische wetenschap heeft methoden ontwikkeld om uit pure gegevens kennis te kunnen halen. Deze methoden, die zijn ontstaan vanuit de statistiek, kunstmatige intelligentie, enz., kunnen interessante relaties blootleggen tussen de verschillende attributen van de gegevens. Zulke relaties kunnen te ingewikkeld en te diep verborgen in de data zijn, zodat mensen onmogelijk in staat kunnen zijn om ze te vinden. In het geval van interessante relaties is dit zeker het geval. De algemene naam van dit wetenschappelijk gebied en zijn methoden is machine learning. Bekende machine learning methodes zijn o.a. statistische regressie, beslissingsbomen, neurale netwerken, genetisch programmeren, associatie regels. Het toepassen van een machine learning procedure op een bepaalde verzameling gegevens levert een zogenaamd model, die bij de data past. Grofweg zijn er twee soorten modellen: beschrijvend en voorspellend. Een beschrijvend model verklaart relaties tussen gegeven attributen, een voorspellend model zegt iets over een attribuut op basis van andere attributen, b.v. response op een direct mail op basis van leeftijd en geslacht.

3. Data mining en knowledge discovery

Data mining is voortgekomen uit machine learning. Ondanks dit is data mining eerder een kunst of praktijk dan een wetenschap. We kunnen data mining definiëren aan de hand van de belangrijkste doelstelling en het proces. De hoofddoelstelling lijkt zeer veel op die van machine learning: het uithalen van niet triviale en bruikbare kennis uit gegevens. Het proces bestaat echter uit meerdere stappen en het uitvoeren van machine learning methodes is daar slechts één van. We kunnen in principe de volgende fases onderscheiden binnen een data mining proces: het verzamelen, opschonen en integreren van de gegevens, het bou-

wen, toetsen, en implementeren van het model en later het onderhouden van gegevens en model. De echte data mining gebeurt bij het bouwen van het model door middel van een machine learning methode. “Knowledge discovery” wordt meestal gebruikt in een brede zin: het staat voor het gehele proces. De bovengenoemde fases van het knowledge discovery proces worden meestal niet in een lineaire manier doorlopen, terugkoppeling naar een eerder fase is heel gewoon. Dit maakt het hele proces nogal complex in de praktijk.

4. Business intelligence

Business intelligence (BI) is de technologie, die diverse ontwikkelingen van de laatste jaren verenigt. BI kan gezien worden als de moderne benadering van beslissings ondersteunende systemen. Terwijl vroeger het optimalisatie aspect centraal lag, ligt tegenwoordig de nadruk op het gebruik van gegevens. De belangrijkste twee doelen van BI zijn:
De gegevensstroom te kanaliseren en te integreren (d.m.v. zogenaamde data warehouses). Tools voor intelligent gebruik van data te bieden. Data mining is een van de mogelijke tools. Volgens veel bedrijfsanalisten zijn de winnaars van de toekomst de bedrijven met de beste klantengegevens. Volgens mijn zienswijze moeten we aan deze (ware) bewering iets toevoegen. De winnaars zullen diegenen zijn met de best gebruikte klantgegevens. Natuurlijk, het gebruiken heeft vele gezichten: gegevens worden gebruikt als ze simpelweg toegankelijk worden gesteld, geraadpleegd en gekopieerd voor het opstellen van een rapport, van meerdere kanten bekeken door een interactieve vragensessie met de database (zogenaamde OLAP technieken) en uiteindelijk als ze gebruikt worden voor modelbouw door datamining. Competitieve voordelen komen voort uit het segmenteren van klanten, het optimaliseren van post selecties, het voorspellen van de verkoop van game consoles, enz. Het bedrijfspotentieel van gegevens is bewezen door de recente commercialisering van dit veld. In het begin van de 90er jaren was data mining een hoofdzakelijk academische activiteit, die werd uitgevoerd door op universiteit en onderzoeksinstituten ontwikkelde hulpmiddelen. Vandaag de dag bieden grote software en technologie verkopers datamining voorzieningen aan als deel van, of als een toegevoegde module aan, hun hoofdproduct. Ook complete business intelligence systemen (die niet meer met academici worden geassocieerd) worden aangeboden door steeds meer professionele leveranciers en gebruikt door consultants.

Samenvatting

Het veld van het managen en gebruiken van data is ontegenzeggelijk steeds belangrijker aan het worden gedurende de laatste 10 jaar. Zijn relevantie komt voort uit de observatie van de groeiende kennis kloof en uit de technologische beloftes om deze kloof (deels) te dichten, In andere woorden is inmiddels alom erkend dat gegevens, wanneer ze juist worden gebruikt, de kwaliteit van besluitvorming in een hoge mate verbeteren en zelfs ook de kwaliteit verhogen van operationele bedrijfsvoering. Dus is het veld van beslissingsondersteuning op basis van gegevens nu al volledig volwassen?
Het antwoord is nee, de doorbraak ligt nog steeds voor ons. Enerzijds bevindt de verspreiding van technologie zich nog voor de grote boom, experts noemen dit een overgangsfase van “early adoption to early majority”. Verder zijn er nog onderwerpen, die aandacht behoeven voor een grote doorbraak van de technologie, zoals bijvoorbeeld schaalvergroting (het kunnen behandelen van zeer grote gegevens verzamelingen), verdere automatisering in het datamining proces (het verlagen van de benodigde vaardigheden van gebruikers), de rol van internet (Web mining) en privacy (internationale en nationale wetgeving betreffende het gebruik van persoonsgegevens). Het is toch met een grote zekerheid te zeggen dat beslissingsondersteuning op basis van gegevens (business intelligence) in populariteit zal groeien dat het steeds meer waarde aan ondernemingen zal toevoegen.