BWI Werkstuk 2010

Datamining in E-mailmarketing: welk communicatiekanaal te gebruiken voor hoge respons

Naam: S. Balkaran

Studentnummer: 1415913

BWI Werkstuk begeleider: D. Roubos

BWI Werkstuk coördinator: H.M.J. van Goor

Opdracht: BWI Werkstuk 2010

Instelling: Vrije Universiteit Amsterdam

Datum: 01 – 07 – 2010

Datamining in E-mailmarketing

Welk communicatiekanaal te gebruiken voor hoge respons

Awien Balkaran
1415913

Begeleider:
Dennis Roubos

BWI Werkstuk 2010

Faculteit Exacte Wetenschappen
Vrije Universiteit Amsterdam
De Boelelaan 1081
1081 HV Amsterdam
Nederland

1 juli 2010

Voorwoord

Dit rapport met als titel *Datamining technieken in E-mailmarketing: welk communicatiekanaal te gebruiken voor hoge respons* heeft betrekking op het opstellen van een BWI Werkstuk. Dit werkstuk is een verplicht onderdeel van de master Business Mathematics and Informatics aan de Vrije Universiteit Amsterdam. Het werkstuk omvat grotendeels een literatuurstudie, aangevuld met eventueel een praktijkstudie waarin de beschreven theorie in de praktijk wordt toegepast.

Het verslag is verdeeld in een aantal delen. Allereerst wordt het begrip *e-mailmarketing besproken*. Vervolgens wordt er ingegaan op welke rol *datamining* al gespeeld heeft binnen dat concept. De datamining techniek die in een concrete situatie wordt toegepast is het *Genetisch Algoritme (GA)*. Deze techniek wordt in detail besproken. Daarnaast worden andere bekende technieken kort besproken. Dit dient ter vergelijking met GA. Tot slot wordt een praktijksituatie beschreven en hoe de beschreven theorie hierop toegepast kan worden.

Graag zou ik mijn begeleider bij het opstellen van mijn BWI Werkstuk, Dennis Roubos, willen bedanken. Het enthousiasme en de wil om iets moois op papier te willen zetten hebben mede bijgedragen aan het eindproduct. Daarnaast waren de aangereikte artikelen stuk voor stuk van toegevoegde waarde voor zowel theoretische kennis als een goede beeldvorming van de stof. Tot slot bedank ik Annemieke van Goor voor de coördinatie rondom het opstellen van een BWI Werkstuk en Sandjai Bhulai voor de pogingen om data van het CBS te mogen ontvangen.

Awien Balkaran

Inhoudsopgave

Samenvatting

Marketing wordt door de *American Marketing Association* gedefinieerd als *the activity, set of institutions, and processes for creating communicating, delivering, and exchanging offerings that have value for customers, clients, partners, and society at large*. Het concept van de marketingmix bestaat uit de zogenaamde *4 P's*: product, plaats, prijs en promotie. Deze worden later uitgebreid met: personeel, proces en *phycical evidence*. Door de toename aan elektronische technologieën is er een nieuwe term ontstaan binnen de marketing: internetmarketing. Eén van de meest effectieve instrumenten op dat gebied is e-mailmarketing, dat wordt gedefinieerd als het gebruik van email voor de marketing communicatie.

Op het gebied van direct marketing worden al datamining technieken toegepast. Hier stuit men op een drietal problemen die te maken hebben met lage respons in marketing databases, lage *accuracy* dat ongeschikt is voor *customer targetting*, en onderbelichte winstmaximalisatie. Om deze problemen te omzeilen zijn technieken bedacht die onder te verdelen zijn in segmentatie modellen en *response* modellen.

Het Genetisch Algoritme is voor het eerst uitgevonden door Holland met als doel om sommige processen van natuurlijke evolutie en selectie na te bootsen. Voor het toepassen van GA zijn de volgende componenten van belang: (chromosomale) representatie, initiële populatie, fitness evaluatie, selectie, cross-over en mutatie. Vijf methoden die gebruik maken van GA zijn: *binary decomposition*, *static range selection*, *dynamic range selection*, *class enumeration* en *evidence accumulation*.

BookCompany.nl, een bedrijf dat een digitaal platform heeft opgezet waar vraag en aanbod van tweedehands studieboeken bij elkaar worden gebracht, vraagt zich af welk communicatiekanaal ingezet dient te worden voor communicatie met haar klanten om zodoende een hoge respons te krijgen. Aan de hand van datamining kan verzamelde data nader onderzocht worden, iets dat ontbreekt in dit rapport. Ter ondersteuning wordt het gebruik van een beslisboom aanbevolen, voornamelijk doordat het een heldere techniek is in de interpretatie en snel toepasbaar is.

1. Inleiding

Bedrijfswiskunde en bedrijfsinformatica zijn bijna niet meer weg te denken bij het uitvoeren van de hedendaagse bedrijfsprocessen. Door het gebruik van technologische hulpmiddelen is het mogelijk om veel data te verzamelen. Deze data kan vervolgens gebruikt worden om van daaruit waardevolle informatie proberen te achterhalen. Dit valt onder het kopje *business intelligence*. Deze aanpak wordt onder andere gebruikt in de marketing. In dit rapport wordt ingezoomd op de e-mailmarketing: hoe werkt dit instrument, waar dien je rekening mee te houden en wat zijn de voor- en nadelen.

Om e-mailmarketing efficiënt te kunnen inzetten heeft men belang bij goede informatie. Deze informatie kan worden verkregen uit data. Na het verzamelen dient deze data te worden geanalyseerd. Een manier om dit aan te kunnen pakken is met behulp van datamining technieken. Deze technieken hebben als grondslag verschillende soorten algoritmen. Het algoritme waarop in dit rapport voornamelijk wordt ingezoomd is het *Genetisch Algoritme*, omdat dit algoritme vaak voorkomt in de literatuur in combinatie met het oplossen van problemen op het gebied van marketing. Dit algoritme wordt beoordeeld en vergeleken met bestaande (traditionele) datamining technieken in de context van e-mailmarketing.

Dit rapport begint met een korte introductie over de basisprincipes van marketing. Vervolgens wordt er specifiek aandacht gevestigd op e-mailmarketing. Hierop volgend wordt een hoofdstuk gewijd aan datamining technieken en modellen die toegepast worden in de (e-mail)marketing. Daarna volgt een hoofdstuk waarin het *Genetisch Algoritme* uitvoerig wordt besproken. Ook wordt in dit hoofdstuk een korte beschrijving gegeven van traditionele datamining technieken om het functioneren en bijbehorende condities te vergelijken met het *Genetisch Algoritme*. Tot slot wordt er een beschrijving gegeven van een praktijksituatie waarin de eerder beschreven theorie samenkomt. Dit deel wordt afgesloten met een aanbeveling.

2. E-mailmarketing

Dit hoofdstuk geeft een beschrijving van de basisprincipes van de marketing. Hiervoor wordt in het kort de marketingmix uitgelegd en de bijbehorende begrippen. Vervolgens komen verschillende marketingtechnieken aan het licht waarbij de technieken op een aantal aspecten met elkaar worden vergeleken. Tot slot wordt er extra aandacht besteed aan e-mailmarketing.

2.1 Korte introductie marketing

Bedrijven bieden producten en/of diensten aan en brengen deze op de markt. Om het laatst genoemde te laten slagen wordt er gedaan aan marketing. Voor dit begrip worden vele verschillende definities gebruikt. De *AMA (American Marketing Association)*, de grootste marketing associatie van Noord-Amerika [1], hanteert de volgende definitie:

> *Marketing is the activity, set of institutions, and processes for creating communicating, delivering, and exchanging offerings that have value for customers, clients, partners, and society at large*. [2,3]

Voor het invullen van de marketingstrategie is het concept van de marketing mix bedacht. Dit concept bestaat uit de zogenaamde 4 P's, namelijk: [3]

- Product (*product*) – bijvoorbeeld de verpakking van het product
- Plaats (*place*) – bijvoorbeeld de supermarkt waar de consument brood kan aanschaffen
- Prijs (*price*) – bijvoorbeeld de prijs die de concurrent vraagt voor hetzelfde product
- Promotie (*promotion*) – bijvoorbeeld de benadrukking van de sterke kanten van het product

Overigens wordt hierboven met 'product' ook diensten die door een bedrijf worden geleverd bedoeld.

Aan het eind van de jaren 70 bleek dat de hierboven beschreven P's met name van toepassing waren op fysieke producten, zoals CD's, alcoholische dranken en keukenzout. Deze traditionele 4 P's bleken niet geschikt voor diensten die worden aangeboden met een minimale fysieke component, zoals laptops met een uitgebreide garantie, kapper en een abonnement bij een sportvereniging. [3] Als gevolg hiervan hebben Booms en Bitner nog 3 P's toegevoegd aan de marketingmix: [4]

- Personeel (*people*) – bijvoorbeeld de professionals die in contact staan met de klanten bij het verlenen van de dienst
- Proces (*process*) – bijvoorbeeld het managen van de klant zijn verwachtingen, interactie en de tevredenheid van de dienst
- Fysiek bewijs (*physical evidence*) – bijvoorbeeld brochures van een universiteit voor studenten om te bepalen aan welke universiteit gestudeerd gaat worden

Deze 7 P's – product, plaats, prijs, promotie, personeel, proces en 'physical evidence', gaan tegenwoordig door het leven als de moderne marketing mix. [5]

2.2 Marketing communicatie mix

Om de aandacht van het publiek te trekken, gebruiken bedrijven een verscheidenheid aan instrumenten en media. Deze elkaar aanvullende en versterkende boodschappen, die een bedrijf uitzend met zijn reclame, voorlichting, Public Relations en andere marketingcommunicatie-instrumenten wordt gezamenlijk bestempeld als de marketing communicatie mix. Deze mix bestaat uit hoofdzakelijk vijf instrumenten, namelijk adverteren, promoties, Public Relations, Direct Marketing en persoonlijke verkoop. Hieronder volgt een nadere toelichting op elk van deze vijf instrumenten: [3]

- Adverteren; een niet-persoonlijke vorm van communicatie, vaak via een sponsor, dat wordt uitgezonden via media waarvoor is betaald
- Promoties; een communicatie-instrument dat van toegevoegde waarde is voor een product of dienst met de intentie om mensen ervan te overtuigen om nu te kopen dan op een later tijdstip

- Public Relations; een niet-persoonlijke vorm van communicatie dat gebruikt wordt door bedrijven om vertrouwen, goodwill, interesse, en uiteindelijk relaties op te bouwen met een groot aantal betrokkenen.
- Direct Marketing; een marketing communicatie instrument dat niet-persoonlijke media gebruikt en daarmee een persoonlijke communicatie zonder aanwezigheid van een tussenpersoon creëert met (potentiële) klanten, en andere belangrijke betrokkenen via bijvoorbeeld geadresseerde post.
- Persoonlijke verkoop; het gebruik van inter-persoonlijke communicatie met als doel om mensen aan te moedigen om een bepaald product of dienst te kopen om er persoonlijk beter van te worden

Door de toename aan elektronische technologieën, zoals het gebruik van de personal computer, is er een nieuwe term ontstaan: internet marketing. Dit wordt ook wel interactieve marketing of online marketing genoemd. [3] Eén van de meest effectieve instrumenten op het gebied van internet marketing is e-mailmarketing. [6] Hier wordt in de volgende paragraaf dieper op ingegaan. Naast e-mailmarketing bestaat internet marketing ook onder andere uit websites, online adverteren en zoekmachine marketing. [3]

2.3 Definitie e-mailmarketing

Zoals omschreven in de vorige paragraaf is e-mailmarketing één van de meest effectieve marketing instrumenten op het gebied van internet marketing. E-mailmarketing wordt gedefinieerd als het gebruik van email voor de marketing communicatie. [7] In de breedste zin het van het woord betekent dit elke mail die wordt verzonden naar een klant, potentiële klant of forum. In het algemeen dient het de volgende drie doelen:

- Het verzenden van direct promotie e-mails om daarmee nieuwe klanten te werven, of bestaande klanten over te halen om weer een aankoop te doen
- Het verzenden van e-mails die ontworpen zijn om daarmee de loyaliteit van klanten aan te moedigen en daarmee de relatie met de klant te benadrukken
- Het plaatsen van marketing berichten of advertenties in e-mails verzonden door andere mensen

Deze drie doelen kunnen gezien worden als de elektronische variant van direct mail, het verzenden van een geprinte nieuwsbrief en het plaatsen van advertenties in kranten en tijdschriften.

Voor bedrijven is het zeer aantrekkelijk om gebruik te maken van e-mailmarketing. Het gebruik maken van dit instrument is erg in trek vanwege de volgende redenen: [7]

- Het versturen van een email is veel goedkoper dan de meeste andere vormen van communicatie
- Het versturen van een email levert jouw bericht af bij de mensen (in tegenstelling tot een website, waar mensen zelf naar jouw bericht moeten komen)
- Email marketing is bewezen erg succesvol voor mensen die het goed toepassen [8]

Daarnaast blijkt uit onderzoek dat e-mailmarketing voor een hoog respons percentage zorgt. [6] De onderliggende factoren hiervan worden later in dit hoofdstuk besproken.

E-mailmarketing valt onder te verdelen in drie typen: direct email, retentie email en adverteren in andermans email. Hieronder volgt een korte toelichting op elk van deze typen: [8]

- *Direct mail*
  Dit type omvat het verzenden van een promotiebericht in de vorm van een email en bevat bijvoorbeeld de aankondiging van een speciale prijs voor een bepaald product. Wanneer je een lijst van klanten hebt of een lijst met beschikbare contactadressen waar je je promotie naar toe kan sturen, is het mogelijk om een lijst van e-mailadressen te verzamelen. Ook komt het voor dat een lijst van e-mailadressen gehuurd kan worden van een dienstverlenende organisatie. Zij zorgen ervoor dat je je bericht kan verzenden naar hun eigen e-mailadressen die zij bezitten. Deze dienst wordt meestal afgestemd naar de aard van het bericht dat je wilt verzenden.

- *Retentie email*
  In plaats van het versturen van promotie email met als doel om de ontvanger aan te moedigen om actie te ondernemen (een bepaald product kopen, je aanmelden voor een abonnement, etc.), kan men er ook voor kiezen om e-mails te versturen met als doel om klanten te behouden. Dit gebeurt vaak in de vorm van nieuwsbrieven. Een dergelijke brief bevat reclame en/of advertenties, maar met als doel om een lange termijn relatie op te bouwen met de lezers. Het zou meer waarde moeten leveren in de zin van meer dan alleen maar verkoopberichten. Daarom dient er meer de nadruk gelegd te worden op meer informatie dat de lezer informeert, vermaakt of voordelen aanbiedt.
- *Adverteren in andermans email*
  In plaats van het produceren van een eigen nieuwsbrief, is het mogelijk om nieuwsbrieven te vinden die zijn gepubliceerd door anderen en hen te betalen voor het plaatsen van jouw advertenties in de e-mails die zij versturen naar hun abonnees. In de praktijk blijkt het vaak zo te zijn dat veel nieuwsbrieven gemaakt zijn voor dit doel; het verkopen van advertentieruimte aan derden.

2.4 Verschillen tussen marketingtechnieken

In tabel 2.1 wordt een overzicht gegeven waarbij verschillende vormen van direct en internetmarketing met elkaar worden vergeleken. Hier wordt onder andere gekeken naar het potentiële bereik, de respons percentage, de respons tijd en of er sprake is van interactiviteit.

Uit tabel 2.1 blijkt onder andere dat de hoofdzakelijke karakteristieken van e-mailmarketing de lage kosten, het korte proces, hoge respons en klantgerichte campagnevoering zijn. Met het proces wordt de tijd bedoeld dat nodig is voor de voorbereiding, het versturen van de e-mails en het ontvangen van de respons. Daarnaast biedt het format – HTML, audio, video e-mail – waarin klanten worden bereikt ruimte voor creativiteit in e-mailmarketing.

	Email	Direct mail	Telemarketing	SMS	Internet advertenties
Bereik	Internet gebruikers	Alle huishoudens	Meeste huishoudens	Mobiele telefonie gebruikers	Internet gebruikers
Respons %	3,5 – 10%	± 2%	10 – 20%	10 – 20%	0,3%
Kosten per bericht	Erg laag 3p	Medium 60p	Hoog 6 pond	Laag 6p	Erg laag 1p
Tijd om te organiseren	Snel	Langzaam (drukwerk)	Langzaam (script)	Snel	Medium
Beschikbaarheid lijst klanten	Gelimiteerd	Erg goed	Good	Erg gelimiteerd	NVT
Gebruikte materialen	Tekst, video, visueel, audio	Verschillende visuele objecten	Alleen de stem	Korte tekst	Tekst, visuele objecten
Respons tijd klanten	Snel	Langzaam	Snel	Snelst	Snel
Interactiviteit	Ja	Nee	Ja	Ja	Ja

Tabel 2.1 Vergelijking verschillen marketingtechnieken[6]

2.5 Succesfactoren e-mailmarketing

Het succes van een marketing instrument wordt afgelezen aan de hand van de respons rate. [6] Wat zorgt er voor dat e-mailmarketing voor een hoog respons percentage zorgt? Oftewel, wat zijn de succesfactoren van e-mailmarketing? Uit een onderzoek van Ruth Rettie konden de volgende factoren worden onderscheiden die leiden tot een significante toename van de respons:

- Onderwerp van de e-mail
- Lengte van de e-mail
- Aanleiding van de e-mail
- Aantal plaatjes

Ook bleek uit datzelfde onderzoek dat respondenten van het onderzoek die al eerder online iets gekocht hadden, tussen de 30 en 34 jaar oud waren, en die een inkomen hadden van boven de 35.000 Engelse pond, een hogere respons laten zien.

2.6 Permission marketing en spam

Tot nu toe is er een zeer positief beeld geschetst van e-mailmarketing waar bijna geen nadelen aan vastkleven. Helaas is dat niet de realiteit. Er zijn een aantal zaken die de gemoederen bezig houden. Naast het feit dat het erg complex is om e-mailberichten te ontwerpen en te bezorgen aan de juiste groep mensen, ervoor te zorgen dat de berichten daadwerkelijk worden gelezen en beantwoorden, het meten en analyseren van resultaten, is er het probleem van toestemming, oftewel in het Engels aangeduid met *permission*. [7] Verantwoordelijke e-mailmarketing is gebaseerd op het idee van *permission*. Dit is een complexe kwestie en onderwerp van veel intensieve debatten in de wereld van de marketing. *Permission* marketing is een manier van adverteren via de e-mail waarbij de ontvanger van de e-mail zelf heeft toegestemd het te willen ontvangen. Het meest belangrijke aspect hiervan is dat je alleen de mensen een e-mail stuurt die erom gevraagd hebben of die toestemming hebben verleend. In het geval je geen toestemming hebt van een ontvanger, wordt de mail herkend als spam: ongewenste e-mail. [3] Wanneer je als bedrijf wordt beschuldigd van het versturen van spam, kan het voorkomen dat je e-mailaccounts worden geblokkeerd, je website uit de lucht wordt gehaald en je reputatie is aangetast. In sommige delen van de wereld ben je in overtreding van de wet. Alles komt er op neer dat toestemming van de ontvanger van essentieel belang is. Een manier om toestemming te verkrijgen is een klant die online een aankoop heeft gedaan de optie aan te bieden om op de hoogte gesteld te worden van aanbiedingen of een nieuwsbrief via de e-mail. Dit kan heel simpel door ergens in het proces van aanschaf van het product te vragen of de klant een vakje wil aankruisen. [7]

3. Datamining in Marketing

In dit hoofdstuk wordt gekeken hoe datamining wordt toegepast op het gebied van direct marketing. Hier komen een aantal knelpunten aan bod die in het kader hiervan zijn gesignaleerd. Om vervolgens deze problemen op te lossen worden target selection algoritmen voorgesteld. Deze zijn weer onder te verdelen in twee categorieën die elk hun eigen modellen hebben. Tot slot wordt er ingegaan op de rol die datamining speelt bij e-mailmarketing.

3.1 Direct marketing en datamining

Op het gebied van marketing worden dataminingtechnieken met name toegepast op het gebied van de direct marketing. Deze vorm van marketing houdt in dat er reclame gemaakt wordt van producten door potentiële klanten direct en persoonlijk te benaderen. [9] In dit kader wordt datamining gebruikt om bepaalde beslissingen efficiënter te laten verlopen. Een voorbeeld hiervan is het bepalen welke doelgroep een hogere respons oplevert en waarbij de kans groot is dat een bepaald product wordt aangeschaft. [9]

Het datamining proces toegepast op direct marketing wijkt enigszins af van het normale *multi phase* proces. Volgens Chuangxin Ou et al. dienen in ieder geval de volgende stappen vertegenwoordigd te zijn: [9]

- Data preparation and processing
- Finding patterns
- Promoting clients

Na toepassing van een aantal dataminingtechnieken is men tot de conclusie gekomen dat er geen oplossingen naar tevredenheid zijn gevonden. Dit komt door een aantal gesignaleerde problemen zoals: [9]

- De traditionele algoritmes kunnen niet worden gebruikt voor een situatie met een lage respons in marketing databases, doordat er dan simpelweg te weinig data beschikbaar is om het algoritme te trainen.

- De voorspelde *accuracy* kan niet gebruikt worden als een geschikte evaluatie criteria voor direct marketing. Een reden hiervoor is dat classificatiefouten anders behandeld moeten worden. Een andere reden is dat de voorspellende *accuracy* te zwak is voor *customer targeting*. Het biedt geen flexibiliteit in het kiezen van een aannemelijk percentage van waarschijnlijke kopers voor promotie.
- De meeste methoden houden alleen rekening met de respons. In de praktijk blijkt een direct marketeer meer geïnteresseerd in het maximaliseren van de winst dan het maximaliseren van de respons. Soms is het maximaliseren van de respons niet gelijk aan het maximaliseren van de winst. Het kan namelijk voorkomen dat je meer winst krijgt van mensen met een kleine kans op respons dan mensen met een hoge kans op respons.

Voor meer gesignaleerde problemen verwijzen wij naar het artikel *On data mining for direct marketing* van Chuangxin Ou et al.

3.2 Target selection algoritmen

Om de hierboven omschreven problemen op te lossen, worden *target selection* algoritmen voorgesteld als de basistechniek voor direct marketing. Deze algoritmen kunnen hoofdzakelijk onderverdeeld worden in twee categorieën: *segmentatie* modellen en *response* modellen. [9]

3.2.1 Segmentatie modellen

Segmentatie modellen verdelen individuen in groepen of clusters op basis van gelijkenissen in karakteristieken of attributen die worden beschreven. Deze groep modellen werkt zo homogeen mogelijk binnen de segmenten en zo heterogeen mogelijk tussen segmenten. De groep met de hoogste kans om te responderen wordt geselecteerd voor promotie. [9]

De meest gebruikte segmentatie modellen zijn *cluster analysis* technieken, zoals AID, CHAID en CART. Het resultaat van deze modellen is een beslissingsboom. Elke knoop in de boom stelt een groep voor waarin elke individu homogeen is. In de praktijk wordt de segmentatie uitgevoerd door het berekenen van een RFM-score

(score om de potentiële waarde van klantsegmenten te bepalen) en de totale lijst te verdelen in verschillende segmenten. [9]

Segmentatie kan niet goed werken op datasets met extreem ongebalanceerde klassenverdeling, omdat de *rate of positive instances* altijd laag is in marketing databases. Het is daardoor moeilijk om een geschikte boom te construeren en waarschijnlijk worden alleen simpele ongebruikelijke regels of patronen ontdekt. In zulke segmentatie modellen is het moeilijk om alle individuen te rangschikken in hetzelfde segment dat gelijke behandeling behoeft. [9]

3.2.2 Response modellen

Response modellen maken gebruik van andere modellen om de kans op respons van elke individu te berekenen en de respondenten met de hoogste kans te kiezen om te promoten. Een aantal response modellen die kunnen worden gebruikt zijn: *Rough Set Model*, *Logit/Probit model*, *Genetic Model*, *Neural Network Model*, en *Market Value Functions Model*. Hieronder volgt per model een korte toelichting.

1. **Rough Set Model**
   Dit model kan gebruikt worden voor het analyseren van potentiële klanten en het doen van voorspellingen wat betreft aankopen. Het algoritme van *Rough Classifier Generation*, waartoe het *Rough Set Model* toe behoort, bestaat uit twee fasen. De eerste fase is de globale segmentatie van de attributenruimte. Het probeert de gemiddelde globale kosten van beslissingen te minimaliseren. De tweede fase is het minimaliseren van het aantal beslissingsregels. In deze fase worden het aantal regels van de classifier geminimaliseerd.
   Het algoritme dat wordt gebruikt bij dit model gedraagt zich goed op training data waarin overbodige attributen aanwezig zijn. Het kan toegepast worden op twee kernproblemen: *prior probabilities* en *unequal misclassification costs*. De *resultant rough classifiers* zijn niet gevoelig voor uitbijters in de data en accepteert databases met veel ruis en inconsistente informatie. Gebruikers van dit model zijn niet gebonden aan het vasthouden van aannames omtrent de data of het model dat is geproduceerd.

2. Logit/Probit Model

Dit model kan toegepast worden op discrete respons data. Het kan worden gebruikt om de target score van elke individu in marketing databases te berekenen. Dit model neemt aan dat elke individu een zekere tendens heeft om te reageren r_t^* op een mailing dat is ontvangen op tijdstip *t*. Deze tendens wordt beïnvloed door X'_t, een stochastische waarde afhankelijk van tijdstip *t*. Hiervoor is de volgende formule van toepassing:

$$r_t^* = X'_t\beta + \varepsilon_t$$

Bij een waarde $r_t^* > 0$, nemen we aan dat de individu zal reageren. In alle andere gevallen nemen we aan dat het individu niet reageert. Een aantal kanttekeningen zijn dat het aanneemt dat de klant dezelfde hoeveelheid geld spendeert aan hetzelfde promotiemateriaal volgens de respons kans. Daarnaast zijn beide modellen alleen respons modellen. Het kan de maximale respons voorspellen, maar het kan niet de maximale winst voorspellen.

3. Genetic Model

Dit model wordt in de direct marketing gebruikt om modellen te bouwen die de verwachte respons maximaliseert. Elke model in de *genetic modelling* heeft een geassocieerde fitness waarde. Een model met een hogere fitness waarde lost het probleem beter op dan een model met een lagere fitness waarde.

Het voordeel van een *genetic model* is dat er geen aannames worden gemaakt, robuust is, non-parametrisch is, en goed functioneert onder zowel grote als kleine steekproeven. Het kan worden gebruikt om complexe relaties onder de loep te nemen. *Genetic models* hebben echter wel een probleem dat aan het begin dient te worden opgelost. Dat is namelijk het vinden van een geschikte fitness functie, zodat het model goed functioneert voor de te gebruiken dataset. Een ander probleem is het kiezen van de parameters: populatiegrootte en voortplanting, cross-over en de kans op mutaties.

4. Neurale Network Model

Een groot voordeel van dit model is dat het te gebruiken is voor eventuele non-lineariteit in de data voor het ontdekken van complexe relaties. Het neurale netwerk stelt een drempel vast, en alleen degene die daarboven scoren ontvangen promotiemateriaal. De complexiteit van het neurale netwerk is afhankelijk van het aantal non-lineaire problemen die opgelost dienen te worden. Het *feed-forward neural netwerk* werkt goed genoeg voor *target selection* problemen. Wanneer dit netwerk gebruikt wordt, kan er tegen het probleem van lokale minima aangelopen worden: in plaats van een globaal minimum levert het model een lokaal minimum. Een manier om dit te omzeilen is het gebruik maken van een genetisch algoritme voor het bepalen van de initiële gewichten in het neurale netwerk.
Neurale netwerken zijn geschikt voor het ontdekken van patronen, dat op zijn beurt weer non-lineaire/complexe relaties kan modelleren, en zijn non-parametrisch en robuust voor ruis. Het netwerk geeft een target score voor elke klant. Deze klanten worden vervolgens gerangschikt volgens de score. De hoogst gerangschikte klanten krijgen promotiemateriaal. Een aantal zwakke punten van het neurale netwerk is het genereren van een complexe formule dat moeilijk te interpreteren is, en het neurale netwerk is moeilijk te formeren.

5. Market Value Functions Model

Dit model stelt een lineair model voor om *target selection* problemen op te lossen in direct marketing. Er wordt aangenomen dat elk object wordt gerepresenteerd door waarden van een eindige set van attributen. De marktwaardefunctie is een lineaire combinatie van *utility* functie op waarden van attributen. Het is afhankelijk van twee delen: *utility* functie en de gewichten van de attributen. Voor het bepalen van de gewichten wordt de entropie bepaald, waarbij een lage entropie staat voor een meer informatieve attribuut. De schatting van de utility functies hangt samen met *probability* modellen van informatie ontvangst. Een *market value* functie is vervolgens een lineaire combinatie van de gewichten van de attributen en de utility functie. Dit kan als volgt worden weergegeven:

$$r(x) = \sum_{\alpha \in At} \omega_\alpha \upsilon_\alpha (I_\alpha (x))$$

…met ω_a het gewicht van attribuut a, en $u_a(I(x))$ de utility functie. *At* is de eindige niet-lege set van attributen. Deze functie kan worden gebruikt voor het berekenen van de target score van elke individu.
De *market value* functie heeft een aantal voordelen. Allereerst, het rangschikt individuen op basis van hun marktwaarde in plaats van classificeren. Ten tweede, is de functie goed te interpreteren. Tot slot is dit model uit te voeren zonder aanwezigheid van expertise.

3.3 E-mailmarketing in datamining

Zoals aangegeven in de vorige paragraaf zijn er meerdere datamining technieken in de literatuur bekend die toegepast kunnen worden op het gebied van direct marketing. Op het gebied van e-mailmarketing is er nog relatief weinig bekend. Wel wordt er veelal links gemaakt met het toepassen van genetische algoritmes. In het volgende hoofdstuk gaan we hier dieper op in.

4. Genetisch algoritme

In dit hoofdstuk wordt het genetisch algoritme besproken, waarbij een stappenplan wordt geschetst hoe dit algoritme dient te worden toegepast. Daarnaast worden er een vijftal methoden beschreven die gebruik maken van het genetisch algoritme. Om een vergelijking te maken met overige datamining technieken worden de voor- en nadelen van een aantal veel voorkomende technieken gegeven. De reden hiervoor is om te zien welke methode het meest geschikt is voor een bepaalde situatie.

4.1 Introductie

Het Genetisch Algoritme (GA) is voor het eerst uitgevonden door Holland met als doel om sommige processen van natuurlijke evolutie en selectie na te bootsen. [10] In de natuur dient elke soort zich aan te passen aan de gecompliceerder en steeds veranderende omgeving om de kans van overleven te maximaliseren. Deze kennis is voor elke soort vastgelegd en gecodeerd in chromosomen, die transformaties ondergaan wanneer er wordt voortgeplant. Chromosomen zijn de dragers van het erfelijkheidsmateriaal en bevinden zich in de celkern. [11] Over een bepaalde periode zorgen deze veranderde chromosomen voor soorten waarvan het aannemelijker is dat ze overleven en een grotere kans hebben om de verbeterde (erfelijkheids)kenmerken door te geven aan toekomstige generaties. Niet alle veranderingen zijn voordelig; in ieder geval hebben de nadelige kenmerken de neiging om uit te sterven.

Grofweg kunnen er drie factoren worden gesteld die van invloed zijn op het genetisch materiaal van de volgende generatie: [12]

1. Individuen met de beste aanpassing planten zich hoogstwaarschijnlijk voort; als gevolg hiervan is de kans groot dat hun eigenschappen worden doorgegeven aan de volgende generatie.

2. Bij het voortplantingsproces van planten en dieren zijn meestal twee individuen betrokken; beiden dragen een deel van hun genetisch materiaal af aan de volgende generatie.
3. In de natuur treden soms mutaties op; er ontstaat een afwijking in het genetisch materiaal, waardoor de kans op overleven afneemt. In uitzonderlijke situaties komt het voor dat juist de mutatie ervoor zorgt dat het individu meer in aanmerking komt voor overleven in hun omgeving.

De sturende factor achter het hele proces van veranderende erfelijkheid is de omgeving. Deze factor bepaalt uiteindelijk welke eigenschappen ervoor zorgen dat de kans van overleven van een individu toeneemt. Hiermee wordt ook een selectie gemaakt van het genetisch materiaal dat doorgegeven wordt naar volgende generaties om de kans van overleven eveneens te doen vergroten.

4.2 Stappenplan GA

Het genetisch algoritme van Holland probeert het natuurlijke genetische proces als volgt te simuleren. De eerste stap is het kiezen van een legale oplossing voor het oplossen van het probleem door het kiezen van een reeks van genen die een bepaalde waarde kunnen aannemen uit een eindige voorgedefinieerde set of alfabet. Deze reeks van genen die een mogelijke oplossing voorstelt staat bekend als een chromosoom. Daarna wordt een initiële populatie van willekeurige legale chromosomen geconstrueerd. Voor elke generatie wordt de *fitness* van elke chromosoom berekend: een hoge fitness waarde is een indicatie voor een betere oplossing dan een lage fitness waarde. De meest fitte chromosomen worden geselecteerd om de volgende generatie te produceren, die daarmee de beste karakteristieken van de ouders overerven. Na enkele generaties van selectie van fitte chromosomen is het resultaat hopelijk een populatie dat substantieel fitter is dan in de beginsituatie. [10]

Om GA toe te kunnen passen zijn de volgende componenten van belang: [10]

- (Chromosomale) representatie
- Initiële populatie
- Fitness evaluatie
- Selectie

- Cross-over en mutatie

We lichten elke van deze componenten apart toe.

4.2.1 (Chromosomale) representatie

Elke chromosoom stelt een mogelijke oplossing voor van het probleem en bestaat uit een reeks van genen. Er bestaan verschillende manieren om een chromosoom weer te geven. De representatie is afhankelijk van de toepassing. De volgende representaties zijn toepasbaar: [10]

- **Binaire representatie**
  Dit is één van de eerste representaties en het vaakst gebruikt. In dit geval bestaat een chromosoom uit nullen en enen {0, 1}.
- **Integer representatie**
  Wanneer een gen op een chromosoom meerdere waarden kan aannemen is een integer een betere keuze: {1, 2, 3, 4} of {Noorden, Oosten, Westen, Zuiden}
- **Geheeltallige of Floating-point representatie**
  Als de genen bestaan uit een continue verdeling dan is een geheeltallige of floating-point representatie een betere keus: $\{x_1, x_2, \ldots, x_k\}$

4.2.2 Initiële populatie

Wanneer een geschikte representatie van de chromosomen is gekozen, is het zaak om een geschikte populatie samen te stellen dat dient als startpunt. Deze populatie kan willekeurig gekozen worden of door gebruik te maken van speciale, probleem specifieke informatie. Voorbeelden hiervan zijn: [13]

- **Fitness Proportional Selection (FPS)**: de kans dat een individu met fitness f_i wordt gekozen is evenredig met zijn fitness waarden ten opzichte van de totale fitness waarde van alle individuen:

$$\frac{f_i}{\sum_{j=1}^{\mu} f_j}$$

- **Ranking selectie**: het sorteren van de populatie of basis van de fitness, waarna selectiekansen worden toegewezen aan de individuen op basis van hun ranking, in plaats van hun eigenlijke fitness waarde.

4.2.3 Fitness evaluatie

Voor het uitvoeren van een fitness evaluatie is eerst het vinden van een geschikte fitness functie van belang. Vervolgens wordt met deze functie berekend wat de fitness van een bepaalde oplossing is. Dit dient ter ondersteuning in hoeverre er met een bepaalde oplossing verder gewerkt kan worden.

4.2.4 Selectie

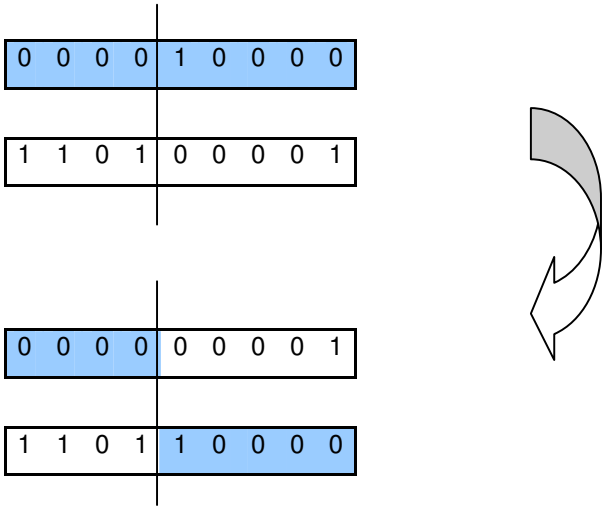
Voor de fase van voortplanting dienen er chromosomen geselecteerd te worden vanuit de huidige populatie. Wanneer de populatie bestaat uit *2n*, waarbij *n* een positieve geheeltallige waarde is, pikt het selectiemechanisme twee chromosomen op basis van hun fitness waarde. Vervolgens vinden er operaties plaats, zoals cross-over en mutaties (hieronder beschreven) om zodoende tot twee nieuwe chromosomen te komen voor de nieuwe populatie. Deze cyclus wordt herhaald totdat de nieuwe populatie *2n* chromosomen bevat, dus na *n* cycli. Hoe hoger de fitness waarde, hoe hoger de kans dat een chromosoom wordt gekozen voor reproductie.

4.2.5 Cross-over (recombinatie) en mutaties

Wanneer een paar van chromosomen is geselecteerd, kan cross-over plaatsvinden om de volgende generatie te vormen. De cross-over rate p_c, een waarde meestal tussen 0,5 en 1.0, bepaalt of er cross-over plaatsvindt. Wanneer twee chromosomen worden geselecteerd, wordt random een waarde tussen 0 en 1 getrokken. Als deze waarde kleiner is dan p_c, dan vindt er cross-over plaats tussen de twee chromosomen. Er zijn verschillende vormen van cross-over te onderscheiden. Voorbeelden hiervan zijn:

- **One-point crossover**
  Hier wordt willekeurig een waarde in het bereik [0, *L*–1] gekozen waarbij de *L* staat voor de totale lengte van het gecodeerde deel van de chromosoom. Op

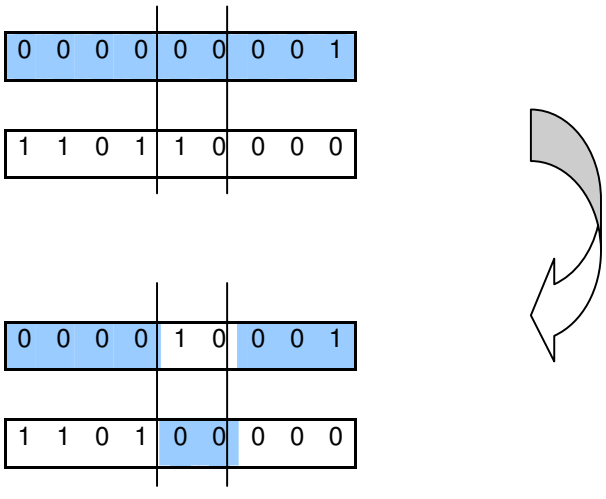
dat punt worden de chromosomen gesplitst en worden de staarten verwisseld om zo de nieuwe chromosomen voor de volgende generatie te krijgen.



Figuur 4.1: One-point crossover

- **N-point crossover**

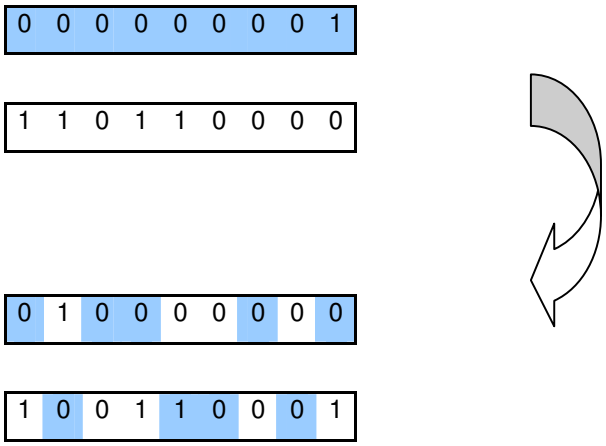
  Dit is een vorm van one-point crossover, waarbij er meerdere punten zijn om de chromosoom in segmenten te verdelen. In de praktijk betekent dit dat er *n* random crossover punten in het interval [0, *L*-1] gekozen dienen te worden. Zie tekening hieronder voor een voorbeeld met *n* = 3.



Figuur 4.2: N-point crossover

- **Uniform crossover**

  Deze methode evalueert elke gen apart om te bepalen of het wordt overgegeven naar de volgende generatie of niet. Hiervoor wordt een string met lengte L bestaande uit random variabelen gegenereerd. Deze waarden variëren tussen de 0 en 1. Voor elke positie geldt het volgende: als de random gekozen waarde onder een bepaalde waarde ligt (bv. 0,5), dan wordt het gen van het eerste chromosoom gekozen; anders van de tweede. De tweede nieuwe generatie chromosoom wordt geconstrueerd door het spiegelbeeld te nemen van de eerste nieuwe generatie chromosoom. Een voorbeeld hiervan wordt weergegeven in figuur 4.3.



Figuur 4.3: Uniform crossover. In dit voorbeeld is de reeks [0,35 0,62 0,18 0,42 0,83 0,76 0,39 0,51 0,36] aan random variabelen getrokken uit [0,1) om de overerving te bepalen

De hierboven beschreven methoden voor crossover hebben betrekking op een binaire representatie. Deze zijn eveneens goed te gebruiken voor integer representaties. Voor floatingpoint representaties bestaan de discrete en *arithmetic recombination* (*simple*, *single* en *whole*), en voor permutatie representaties bestaan de *Partially Mapped Crossover* (PMX), *Edge crossover*, *Order Crossover* en de *Cycle Crossover*.

De pseudocode voor een vorm van een genetisch algoritme ziet er als volgt uit:

```
P(0) = new Population(N, RANDOM)
P(0).fitness();
for (t=1; t<=Ngen.; t++) {
        P(t) = new Population(N);
        for (i=1; i<=Npselectie; i++) {
                P(t).add(P(t-1).fittest());
        }
        for (i=Npselectie + 1; i<=N; i++) {
                individu1 = new Individu(P(t-1).select());
                individu2 = new Individu(P(t-1).select());
                if (random() < pcrossover)
                        crossover(individu1, individu2);
                if (random() < pmutatie)
                        mutate(individu1);
                if (random() < pmutatie)
                        mutate(individu2);
                P(t).add(individu1);
                P(t).add(individu2);
        }
        P(t).fitness();
}
return P(t).fittest();
```

4.3 Classificatie met gebruik van GA

Nu we een beschrijving hebben hoe het Genetisch Algoritme werkt, schakelen we over naar de toepassing ervan in de praktijk. In het artikel van Thomas Loveard en Victor Ciesielski, *Representing Classification Problems in Genetic Programming*, worden vijf methoden toegepast die allen gebruik maken van GA. Het gaat hier om de volgende methoden: *binary decomposition*, *static range selection*, *dynamic range selection*, *class enumeration* en *evidence accumulation*. Deze methoden worden toegepast voor een multi-classificatie taak om vervolgens te onderzoeken welke

methode het meest geschikt is. Dit wordt uitgedrukt in termen van *accuracy of classification*. Ook wordt er gekeken naar de tijd die nodig is om tot die oplossing te komen. [14]

4.3.1 Waarom GA voor classificatietaken?

Op het gebied van classificatie biedt GA flexibele mogelijkheden om een oplossing te geven van zeer complexe problemen. Dit geeft GA een voorsprong ten opzichte van algemeen gebruikte classificatietechnieken zoals beslisbomen, neurale netwerken en statistische classifiers. Daarnaast geldt dat wanneer er meer tijd is om de dataset te trainen, des te accurater de classifier kan worden getraind, in tegenstelling tot bijvoorbeeld C4.5 die altijd dezelfde classifier zullen opleveren ondanks de toegenomen trainingstijd. Een ander voordeel is dat elke run probabilistisch is, waardoor verschillende runs die hetzelfde probleem beogen op te lossen zo goed als nooit dezelfde oplossing zullen geven. Hieronder bespreken we kort de essentie van de vijf gepresenteerde methoden die gebruik maken van GA. [14]

4.3.2 Binary decomposition

Wanneer twee of meerdere classes betrokken zijn bij een classificatieprobleem kan *binary decomposition* uitkomst bieden. Bij deze methode wordt een multi-class probleem behandeld als een set van binaire classificatie problemen. Gegeven een probleem *P* met een set van *n* classes $P = \{c1, c2, \ldots, cn\}$. Dit probleem kan vervolgens verdeeld worden in $n - 1$ binaire classificatie problemen. Het eerste binaire subprobleem wordt het classificeren tussen $\{c1, x\}$, waarbij *x* een samenvoeging is van alle overige classes: $P - \{c1\}$. Het tweede binaire subprobleem wordt het classificeren tussen $\{c2, y\}$, waarbij *y* staat voor $P - \{c1, c2\}$. Dit proces gaat door totdat het laatste probleem bestaande uit de classificatie set $\{cn - 1, cn\}$. Voor deze aanpak wordt gebruik gemaakt van standaard functies en terminal sets weergegeven in tabel 4.1 en 4.2. [14]

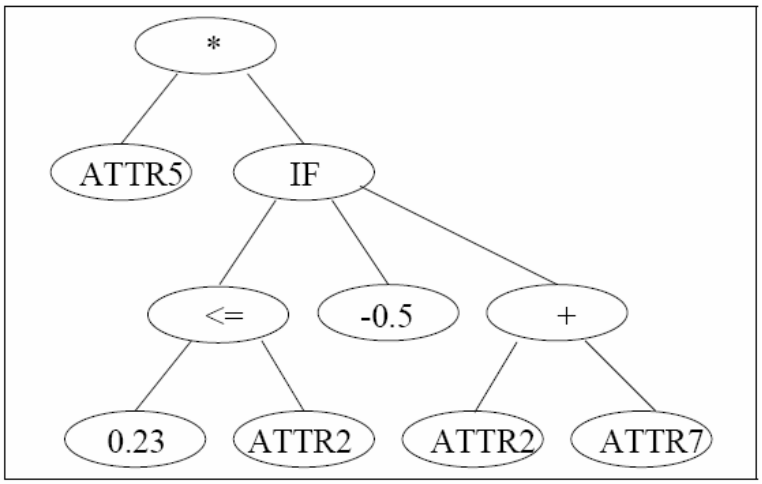
Name	Return type	Arguments	Argument types	Description
Plus	Double	2	Double, double	Arithmetic addition
Minus	Double	2	Double, double	Arithmetic subtraction
Mult	Double	2	Double, double	Arithmetic multiplication
Div	Double	2	Double, double	Protected arithmetic division (divide by zero returns the value zero)
IF	Double	3	Boolean, double, double	Conditional. If arg1 is true, then return arg2, otherwise return arg3
<=	Boolean	2	Double, double	True if arg1 is less than of equal to arg2
>=	Boolean	2	Double, double	True if arg1 is greater than or equal to arg2
=	Boolean	2	Double, double	True if arg1 is equal to arg2
Between	Boolean	3	Double, double, double	True if value of arg1 is between the values of arg2 and arg3

Tabel 4.1: standaard functie set

Name	Return type	Description
Random(-1,1)	Double	Randomly assigned constant with value between -1 and 1
Attribute[x]	Double	Value of attribute x

Tabel 4.2: standaard terminal set

Na gebruikmaking van deze regels kan een programma er uit zien zoals grafisch weergegeven in figuur 4.4. In dit voorbeeld is de output een geheeltallige waarde. Door de *IF*, wordt de output van attribuut 5 vermenigvuldigd met; de constante waarde -0,5 wanneer 0,23 gelijk of kleiner dat de waarde van attribuut 2 is en anders met de som van de waarde van attributen 2 en 7. [14]



Figuur 4.4: een voorbeeld 'program tree' voor Binary Decomposition, Static Range Selection of Dynamic Range Selection

4.3.3 Static range selection

Genetische programma's die classificatie uitvoeren retourneren vaak een geheeltallige waarde waardoor het moeilijk is om betekenisvolle punten te selecteren voor het afgrenzen van classes, met name bij multiclass classificatie problemen. Bij deze methode worden er ranges willekeurig gekozen waarbinnen een waarde als grens wordt gekozen. De kans is echter wel groot dat deze ranges niet geheel optimaal zijn gekozen. Ook deze methode maakt gebruik van de standaard functies en terminal sets beschreven in tabel 4.1 en 4.2. [14]

4.3.4 Dynamic range selection

Deze methode werkt volgens de *static range selection*, maar hier is het voor elk programma toegestaan om andere ranges toe te passen die dynamisch zijn bepaald. Dit wordt gedaan met behulp van *training examples*: de waarden die worden geretourneerd van een subset van een *training example* kunnen worden gebruikt om grenzen van een class op te stellen. Ook deze methode maakt gebruik van de standaard functies en terminal sets beschreven in tabel 4.1 en 4.2. [14]

4.3.5 Class enumeration

Bij deze methode wordt een nieuwe datatype geïntroduceerd: *ClassType*, een enumerated type. De set van waarden dat dit type kan opslaan is gelimiteerd door het aantal verschillende class types van het gegeven probleem. Een nieuwe terminal is eveneens geïntroduceerd, namelijk *ClassNum*, dat een waarde retourneerd van het type *ClassType*. Verder is de *IF* aangepast (zie Tabel X.X en X.X). [14]

Name	Return Type	Arguments	Argument Types	Description
IF	ClassType	3	Boolean, ClassType, ClassType	Conditional. If arg1 is true, then return arg2, otherwise return arg3.

Tabel 4.3: Class Enumeration; aanpassing van de standaard functie set

Name	Return Type	Description
ClassNum	ClassType	One, out of the possible set of classes for the given problem

Tabel 4.4: Class Enumeration; aanvulling op standaard terminal set

In zijn aanpak lijkt deze methode veel op de C4.5. een groot verschil is dat elke conditionele knoop in de boom bijna elke aritmetische expressie dat mogelijk is kan aannemen, gegeven de functies en de terminal set.

4.3.6 Evidence accumulation

Deze methode staat vele verschillende branches van de program boom toe om bij te dragen aan de beslissing voor het kiezen van een bepaalde class. Naast de program boom bevat het ook een *vector data storage area*, genaamd de *certainty vector*. Deze vector bevat één element voor elke mogelijke class van het probleem. Voordat het program wordt gestart, wordt elk element van de vector geïnitialiseerd op nul. Als het program wordt gerund worden er waarden toegevoegd (of afgetrokken) van

de desbetreffende elementen uit de vector via de operatie van een nieuwe terminal *AddToClass[x](-1,1)*, weergegeven in tabel X.X. Deze terminal voegt waarden toe in de range van -1 tot 1 aan één van de elementen uit de *certainty vector*. Wanneer het programma stopt wordt de *certainty vector* geëvalueerd en de element met de hoogste waarde in de vector wordt gezien als de meest zekere uitkomst voor classificatie. In het geval dat twee elementen dezelfde waarde hebben, wordt het resultaat gerapporteerd als een error. [14]

Aangezien de uitkomst van de classificatie gelegen ligt in de *certainty vector*, is het niet noodzakelijk voor het programma om een waarde te retourneren. Als gevolg hiervan wordt de *IF* functie aangepast (zie tabel 4.5). daarnaast wordt er een functie genaamd *BLOCK* toegevoegd. Deze functie accepteert van twee tot vier argumenten en evalueert elk argument in een bepaalde volgorde. Deze functie is nodig, zodat het mogelijk is voor meerdere toevoegingen (of aftrekkingen) van waarden in de *certainty vector*.

Name	Return Type	Arguments	Argument Types	Description
IF	No Value	3	Boolean, No Value, No Value	Conditional. If arg1 is true, then evaluate arg2, otherwise evaluate arg 3.
BLOCK(2, 4)	No Value	2 – 4	All No Value	Evaluate all arguments sequentially

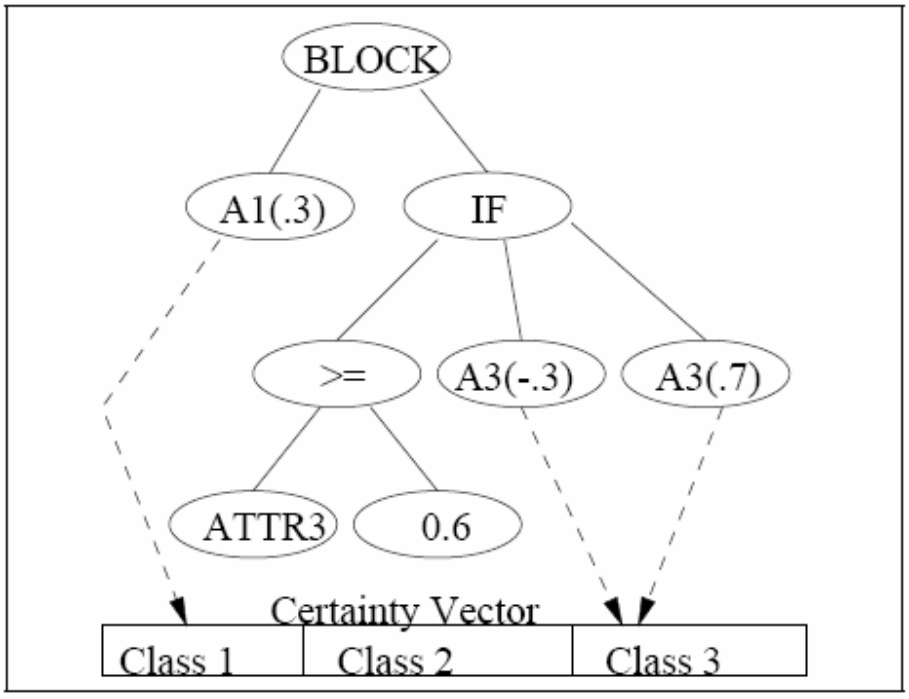
Tabel 4.5: Evidence Accumulation; aanpassing standaard functie set

Name	Return Type	Description
AddToClass[x](-1,1)	No Value	Add a value between -1 and 1 to a certainty value of class x

Tabel 4.6: Evidence Accumulation; toevoeging standaard terminal set

Een voorbeeld van deze methode wordt weergegeven in figuur 4.5. Een terminal van de vorm *An(x)* houdt in dat een *AddToClass* functie wordt toegepast waarbij de

certainty value x wordt toegevoegd aan class *n*. in dit voorbeeld wordt automatisch de waarde van 0,3 toegevoegd aan class 1 element van de *certainty vector*. Vervolgens wordt er 0,3 van afgetrokken van class 3 van de *certainty vector* wanneer de waarde van attribuut 3 groter of gelijk is dan 0,6, en anders wordt er 0,7 opgeteld bij class 3 van de *certainty vector*. [14]



Figuur 4.5: een voorbeeld 'program tree' van Evidence Accumulation

4.4 Overige datamining technieken

Naast het toepassen van verschillende technieken die gebruik maken van het genetisch algoritme zijn er meerdere datamining technieken. Deze technieken worden met name toegepast in de Artificial Intelligence. Het gaat hierbij om de volgende methoden: [15]

- Beslisbomen (*Decision trees*)
- Neurale netwerken
- *k*-Nearest neighbor
- Naive Bayes
- Support Vector Machines

Hieronder volgt een korte toelichting van elke techniek met hun voor- en nadelen. Voor een uitgebreidere toelichting verwijs ik naar het boek *Data Mining: Practical Machine Learning Tools and Techniques* van Ian H. Witten, Eibe Frank en Morgan Kaufman.

Beslisbomen

Beslisbomen volgen een regelgebaseerde methode. Door deze regels af te lopen wordt uiteindelijk gekomen tot een classificatie. Bij het opstellen van de beslisboom wordt gekeken naar de belangrijkste eigenschappen om uiteindelijk tot de classificatie te komen. Deze methode maakt gebruikt van regels in de vorm van "als … dan …".

Voordelen:

- Snelle toepasbaarheid
- Eenvoudige opzet
- Makkelijk te herleiden keuze
- Redelijke nauwkeurigheid

Nadelen:

- Matig bruikbaar wanneer met grote hoeveelheden diverse informatie wordt gewerkt
- Gebruik van storende gegevens leidt tot onjuiste antwoorden

Neurale netwerken

Een neuraal netwerk is een statistische analyse gebaseerd op leermodellen van biologische zenuwstelsels. Patronen worden herkend aan de hand van voorbeelden. Het klassieke model bestaat uit drie lagen: de inputlaag, een outputlaag en daartussen de verborgen laag (*hidden layer*). Data uit de inputlaag wordt via een knooppunt in de verborgen laag gekoppeld aan een bepaalde output. De knooppunten in de verborgen laag krijgen bepaalde waarden toegekend, waardoor het mogelijk is om een netwerk te bouwen waarin alle output met elkaar verbonden zijn.

Voordelen:

- Goed en betrouwbaar na voldoende training
- Kan op een groot aantal verschillende problemen toegepast worden

Nadelen:

- Het trainen van een neuraal netwerk kost veel tijd
- Het is niet mogelijk het model eenvoudig aan te passen
- De keuzes die een neuraal netwerk maakt zijn niet te verklaren

_k_-Nearest neighbor

Deze methode gaat uit van een verdeling van de data in groepen. Bij elke classificatie wordt een object vergeleken met *k* andere al bekende resultaten die daar het meest op lijken. Vervolgens wordt de classificatie die het vaakst voorkomt bij die *k* andere resultaten gekozen voor het nieuw te classificeren object. [16]

Voordelen:

- Leren gaat snel
- Mogelijk om complexe outputfuncties te leren
- Geen verlies van informatie uit trainingsvoorbeelden

Nadelen:

- Traag tijdens het beantwoorden van queries
- Minder geschikt bij aanwezigheid van irrelevante inputvariabelen

Naive Bayes

Bij deze methode staan waarschijnlijkheid en statistiek centraal. Hierbij wordt gebruik gemaakt van de theorie van Thomas Bayes. Deze theorie maakt inzichtelijk hoe inschattingen van de kans op een onzekere gebeurtenis *A* worden beïnvloed door het optreden van een andere onzekere gebeurtenis *B*. [17]

Voordelen:

- Het heeft weinig training data nodig
- Werkt snel bij zeer uitgebreide datasets met veel features, zowel bij het bouwen van de classifier als bij het classificeren van nieuwe objecten
- Onafhankelijkheid tussen variabelen wordt aangenomen, waardoor covarianties niet nodig zijn

Nadelen:

- Aanname van onafhankelijkheid tussen parameters is niet altijd gerechtvaardigd

- Kleine groepen worden afgestraft, omdat de priorkans een doorslaggevende rol speelt bij de classificatie

Support Vector Machines

Deze machines bestaan uit een set van gerelateerde *supervised learning* methoden die gebruikt worden voor classificatie en regressie. Gegeven een set van trainingsvoorbeelden, waarbij elk object gemarkeerd is als horende bij één van de twee categorieën, bouwt een SVM training algoritme een model dat voorspeld in welke categorie het nieuwe object valt. Hierbij wordt gebruik gemaakt van representaties in een meerdimensionale ruimte, zodanig dat de trainingsvoorbeelden vallen in twee duidelijke gebieden die zijn te verdelen door een duidelijke *gap*. Deze gap is zo groot mogelijk. Nieuwe voorbeelden worden dan voorspeld aan de hand van de ruimte waarin ze vallen ten opzichte van de *gap*. Daarnaast maakt dit algoritme gebruik van een *kernelfunctie*. [18,19]

Voordelen:

- Hoge mate van flexibiliteit, schaalbaarheid en snelheid
- Kan werken met complexe, echte problemen
- Presteert goed op datasets die vele attributen hebben; er is geen bovengrens voor het aantal attributen

Nadelen:

- Lage uitvoersnelheid; het opbouwen van de classifier is kwadratisch
- De keuze van kernel heeft grote invloed op het construeren van de classifier
- Hoge algoritmische complexiteit en de vereiste geheugen daarvoor

5. Business Case

In dit hoofdstuk wordt een praktijksituatie geschetst. Deze situatie beschrijft een probleem dat een tweedehands studieboeken verkoopsite ervaart. Door middel van een beschrijving wordt middels een kort onderzoek getracht een antwoord te vinden op de onderzoeksvraag. Ter afsluiting wordt er een aanbeveling gedaan.

5.1 Bookcompany.nl

Het bedrijf BookCompany.nl bestaat sinds september 2009 en is opgezet door twee studenten aan de Vrije Universiteit van Amsterdam. Het bedrijf werkt via een digitaal platform waar vragers en aanbieders van tweedehands studieboeken bij elkaar worden gebracht. Het concept is zodanig opgezet dat er een lokale markt wordt gecreëerd: de vragers en aanbieder bevinden zich op dezelfde fysieke locatie, namelijk de instelling waaraan ze studeren. Hier kan de inzage van het studieboek en de overdracht van het geld plaatsvinden. Daarnaast besteedt het bedrijf veel tijd en energie aan promotie om zodoende voor meer naamsbekendheid te zorgen ten koste van overige organisaties die een dergelijke dienstverlening aanbieden. BookCompany.nl biedt daarnaast een dienst op maat. Dat houdt in dat aan de hand van profielgegevens notificaties kunnen worden gemaakt van studieboeken die de desbetreffende persoon mogelijk nodig zou kunnen hebben.

5.2 Probleemstelling

In hoofdstuk 1 is een beschrijving gegeven van marketing en hoe dit wordt geïntegreerd in bedrijfsactiviteiten. Er zijn verschillende manieren om de beoogde doelgroep te bereiken. Voor verschillende situaties zijn verschillende manier het best toe te passen. Zo kan voor het bedrijf dat product A op de markt brengt telefonische enquête een hoge respons opleveren en voor een ander bedrijf dat product B op de markt brengt face 2 face verkoop de hoogste respons opleveren. Het doel van deze Business Case is om te achterhalen welk communicatiekanaal BookCompany.nl het

best kan toepassen met het oog op het behalen van een hoge respons. Hierbij dient gedacht te worden aan promotie van het bedrijf en de boodschap om je aan te melden op de website. De vraagstelling luidt als volgt:

> *Welk communicatiemiddel dient BookCompany.nl in te zetten om een hoge respons te bereiken bij haar klanten?*

In de volgende paragrafen geven we een toelichting hoe de beantwoording van deze onderzoeksvraag tot stand komt. Er dient gefocust te worden op studenten en recentelijk afgestudeerden, oftewel de leeftijdsgroep van ongeveer 18 tot 28 jaar.

5.3 Materialen en methoden

5.3.1 Data

Voor de beantwoording van de onderzoeksvraag is data nodig. De data die is verkregen dient te worden hoort te bestaan uit de volgende onderdelen:

- Persoonlijke gegevens
  - Leeftijd
  - Geslacht
  - Woonplaats
  - Burgerlijke status
  - Opleidingsniveau
  - Sociaal Economische Status (SES)

- Eventuele extra karakteristieken
  - Student?
  - Hoeveel jaar geleden student?
  - Lid van studievereniging?
  - Actief op of ooit gebruik gemaakt van Marktplaats.nl, Bol.com of dergelijke site waar tweedehands producten worden aangeboden?

- Communicatiemiddelen:
  - Online enquête

- Telefonische enquête
- Schriftelijke enquête
- Mondelinge enquête (interview)
- Focusgroepen
- SMS enquête
- Fax enquête
- Gereageerd/respons op enquête (uitkomstvariabele)

5.3.2 Inclusie/exclusie criteria

Voor de analyse van de data worden eerst een aantal criteria toegepast. De data bevat gegevens van studenten. Daarnaast is de eis dat wanneer iemand geen student meer is, maximaal 4 jaar geleden moet zijn afgestudeerd. De looptijd van een tweedehands studieboek is gemiddeld 4 jaar, dus heeft het na maximaal 4 jaar geen zin om een boek te koop aan te bieden. Vervolgens wordt er onderscheid gemaakt tussen de volgende twee groepen:

- Lid van studievereniging vs. lid van geen studievereniging
- Actief gebruikmakend van tweedehands producten sites vs. niet actief gebruikmakend van tweedehands producten sites

5.3.3 Analyse data

Voor het analyseren van de geprepareerde data wordt gebruik gemaakt van datamining technieken. De technieken die worden toegepast zijn dezelfde als beschreven in dit rapport:

- Genetisch algoritme (5 methoden)
- Beslisbomen (*Decision trees*)
- Neurale netwerken
- k-Nearest neighbor
- Naive Bayes
- Support Vector Machines

5.4 Resultaten

Doordat er geen geschikte training data gevonden kon worden is er in dit rapport geen mogelijkheid gevonden om de gepresenteerde theorie toe te passen in de praktijk. Daarnaast is een poging om data vanuit het CBS te verkrijgen niet succesvol gebleken. Het zelfstandig verzamelen van data bleek daarnaast een te tijdrovend karwei te zijn voor dit rapport. Als gevolg hiervan wordt een advies uitgebracht voor de toe te passen datamining techniek, dat zorgt voor de hoogste kwaliteit van de classifier. Deze techniek dient toegepast te worden op de geprepareerde data zoals omschreven in de paragrafen eerder in dit hoofdstuk.

5.5 Conclusie

Als we kijken naar onze dataset, dan zijn er weinig attributen nodig. Wat dat betreft hebben we niet te maken met een complex vraagstuk, waardoor een Support Vector Machine niet aan te raden is. Het zoeken naar een goede fitness functie zorgt ervoor dat een Genetisch Algoritme niet verstandig is te gebruiken. Een neuraal netwerk neemt veel tijd in beslag om te trainen en gezien de praktische toepasbaarheid van de business case is het belangrijk om de keuzes van het netwerk ook te kunnen verklaren. Ook de aanname van onafhankelijkheid tussen verschillende variabelen kan een onterechte aanname zijn: het is namelijk aannemelijk dat er een verband bestaat tussen SES en opleidingsniveau. Aangezien er een aantal classes zijn, is het niet raadzaam om *k*-NN te gebruiken. Als conclusie kunnen we daarom trekken dat een beslisboom aan te raden valt als te gebruiken datamining techniek. Deze techniek is helder in de interpretatie en snel toepasbaar.

Literatuurlijst

[1] **About American Marketing Association**. *Marketingpower*. Laatst geraadpleegd op 26 januari 2010. http://www.marketingpower.com/AboutAMA/Pages/default.aspx

[2] **Definition of Marketing**. *Marketingpower*. Laatst geraadpleegd op 26 januari 2010 http://www.marketingpower.com/AboutAMA/Pages/DefinitionofMarketing.aspx

[3] Paul Baines et al. **Marketing**. *Oxford University Press*; 2008.

[4] J. Bitner en B. Booms. **Marketing strategies and organizational structures for service firms**. *American Marketing Association*; Chicago; 1981

[5] **The Extended Marketing Mix (7P's)**. *The Times 100*. Laatst geraadpleegd op 4 februari 2010. http://www.thetimes100.co.uk/theory/theory--the-extended-marketing-mix-(7ps)--319.php

[6] R. Rettie. **Email marketing: Success Factors**. *Kingston University*. 2002

[7] M. Brownlow. **What is email marketing?** *Email Marketing Reports.* Laatst geraadpleegd op 2 maart 2010. http://www.email-marketing-reports.com/intro.htm

[8] M. Brownlow. **Why do email marketing?** *Email Marketing Reports.* Hersteld mei 2010, eerste publicatie november 2010. Laatst geraadpleegd op 2 maart 2010.

http://www.email-marketing-reports.com/basics/why.htm

[9] C. Ou et al. **On Data Mining for Direct Marketing**. *Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing*. Springer Berlin / Heidelberg; 2003.

[10] S. Hurley, L. Moutinho, N.M. Stephens. **Solving Marketing Optimization Problems Using Genetic Algorithms**. *European Journal of Marketing*. Volume 29; Issue 4; Page 39 – 56; 1995.

[11] **Chromosomen**. *Erfelijkheid.nl*. Laatst geraadpleegd op 6 april 2010 http://www.erfelijkheid.nl/erfelijkheid/chromosomen.php

[12] G.J. Bex. **Genetische Algoritmen**. 1 december 2000 http://alpha.uhasselt.be/~gjb/info/pubs/geneticAlgorithms.pdf

[13] A.E. Eiben en J.E. Smith. **Introduction to Evolutionary Computing**. *Springer, Natural Computing Series*. 1st edition, 2003.

[14] T. Loveard en V. Ciesielski. **Representing Classification Problems in Genetic Programming**. *Department of Computer Science, Royal Melbourne Institute of Technology*

[15] M. Grootveld en W. Huijsen. **Automatische Classificatie – De Technieken**. 21 december 2005.

[16] Steven L. Salzberg en David W. Aha. **Learning to Catch: Applying Nearest Neighbor Algorithms to Dynamic Control Tasks**. http://www.aaai.org/Papers/Symposia/Fall/1992/FS-92-02/FS92-02-025.pdf

[17] B. van Herum. **Automatic opinion extraction and classification from text**. 22 augustus 2007 https://doclib.uhasselt.be/dspace/bitstream/1942/3731/1/van-hertum.pdf

[18] Oracle Data Mining Concepts 11g. Hoofdstuk 18: Support Vector Machines September 2007 http://www.comp.dit.ie/btierney/oracle11gdoc/datamine.111/b28129/algo_svm.htm

[19] **Voordelen Support Vector Machines**. Laatst geraadpleegd op 18 mei 2010. http://www.svms.org/disadvantages.html