LIMBURGS UNIVERSITAIR CENTRUM
PARTNER IN DE UNIVERSITEIT LIMBURG

Optimaliseren van e-business campagnes met behulp van data mining technieken

Sabrina NOBLESSE

Promotor : Prof. Dr. K. VANHOOF

Thesis voorgedragen tot het behalen van de graad
van Handelsingenieur in de Beleidsinformatica
Academiejaar 2004-2005

Woord vooraf

In de eerste plaats zou ik graag mijn promotor Prof. Dr. Koen Vanhoof bedanken, zonder zijn hulp was deze thesis niet mogelijk geweest. Verder wil ik me ook richten tot Tom Moers en Benoît Depaire die me steeds bijstonden met hun kennis en advies. Ook Lene Clijsters en andere medestudenten die me geholpen hebben tijdens het schrijven van mijn thesis wil ik graag bedanken. Als laatste zou ik mijn ouders, mijn zus en mijn vriend Tom Moers willen bedanken omdat zij mij steunen in alles wat ik doe.

Samenvatting

Meer en meer bedrijven willen het Internet gebruiken om klanten aan te trekken en te selecteren. Door de overvloed aan data die men via dit medium krijgt, is er nood aan technieken en methodes om deze data te analyseren en om te zetten naar informatie. Ook het verkennen en modelleren van betekenisvolle relaties en patronen die gevonden werden in de beschikbare data is belangrijk.

Data mining is het verkennen en analyseren van grote hoeveelheden observationele data met als doelen het samenvatten van de data op een nuttige en begrijpbare manier voor de eigenaar van de data en het ontdekken, verkennen en modelleren van eventuele betekenisvolle relaties, regels en patronen in de beschikbare data. Observationele data zijn data die niet door een bedrijf worden aangemaakt, verzameld of bijgehouden met het doel gebruikt te worden voor data mining. Om aan data mining te kunnen doen werden er diverse data mining technieken, zoals beslissingsbomen, ontwikkeld.

Iedere data mining techniek kan echter een verschillend resultaat opleveren en is niet geschikt om op alle soorten data toegepast te worden. Ook de manier waarop de resultaten van de gebruikte data mining techniek(en) beoordeeld zullen worden kan belangrijk zijn en voor verschillende uitkomsten zorgen. De situatie en het bedrijf waarin het data mining project moet uitgevoerd worden, hebben eveneens een invloed op de keuze van een bepaalde data mining techniek. Een bedrijf zal dus zijn e-business campagne proberen te optimaliseren aan de hand van zorgvuldig geselecteerde data mining technieken en evaluatiemethodes.

Het kiezen van de juiste evaluatiemethode voor een specifieke data mining techniek is niet altijd evident. Er bestaan namelijk verschillende methodes om de resultaten van data mining technieken te beoordelen. Deze evaluatiemethodes kunnen bovendien ook verschillend zijn voor elke data mining techniek. Aangezien de keuze van een evaluatiemethode een invloed kan heb-

ben op het succes van een data mining onderzoek, is het belangrijk om de meest optimale methode te vinden.

In deze thesis gaan we aan de hand van een gevalstudie onderzoeken of het sensitiviteitsalgoritme gebruikt kan worden voor het beoordelen en rangschikken van resultaten gegeven door binaire beslissingsbomen. Het sensitiviteitsalgoritme werd oorspronkelijk voorgesteld door Isabelle Alvarez in het artikel *Sensitivity Analysis of the Result in Binary Decision Trees: Giving More Information to the End-user* met als doel het uitvoeren van een sensitiviteitsanalyse van het resultaat van binaire beslissingsbomen zodat er meer informatie kan gegeven worden aan de eindgebruiker. In het sensitiviteitsalgoritme wordt per geclassificeerd geval de afstand tussen het geval en het bijhorende beslissingsoppervlak berekend. We veronderstellen dat deze afstand omgekeerd evenredig is met de kans dat de besslissingsboom een verkeerde classificatie heeft gedaan.

We hebben onderzocht op welke manier het sensitiviteitsalgoritme kan gebruikt worden om de resultaten of classificaties gegeven door een beslissingsboom te beoordelen en de geclassificeerde gevallen te rangschikken. Met waarschijnlijkheidsschatting bedoelen we de kans dat de voorspelling die de beslissingsboom gedaan heeft juist is, of met andere woorden dat de voorspelde klasse ook de werkelijke klasse zal zijn van het specifieke geval.

Afhankelijk van welke doelen men vooropstelt betreffende de gewenste resultaten zal het sensitiviteitsalgoritme, al dan niet in combinatie met de waarschijnlijkheidsschattingen, gebruikt kunnen worden om de door een binaire beslissingsboom geclassificeerde gevallen te rangschikken. Een voorbeeld van een doel is het vooropstellen dat vooral de gevallen met een weinig frequente klasse juist geclassificeerd zullen worden.

Inhoudsopgave

Lijst van figuren

Lijst van tabellen

Hoofdstuk 1

Inleiding en probleemstelling

Dit hoofdstuk bevat de probleemstelling waarop deze thesis gebaseerd is. De werkwijze van het onderzoek wordt eveneens weergegeven en toegelicht. Als laatste geeft dit hoofdstuk een overzicht van de inhoud van deze thesis.

1.1 Praktijkprobleem: omschrijving en situering

Het Internet heeft de laatste jaren een spectaculaire groei gekend. Dit in combinatie met de lage kosten om data te verzamelen via dit medium, heeft als gevolg dat meer en meer bedrijven het Internet willen gebruiken om klanten aan te trekken en te selecteren. Door de overvloed aan data die men via het Internet ter beschikking heeft, is er nood aan technieken en methodes om deze data te analyseren en om te zetten in informatie. Ook het verkennen en modelleren van betekenisvolle relaties en patronen die gevonden werden in de beschikbare data is hierbij belangrijk. Data mining technieken, zoals beslissingsbomen, kunnen hiervoor een oplossing bieden.

Iedere data mining techniek kan echter een verschillend resultaat opleveren en is niet geschikt om op alle soorten data toegepast te worden. Ook de manier waarop de resultaten van de gebruikte data mining techniek(en) beoordeeld en gerangschikt zullen worden kan van belang zijn en voor verschillende uitkomsten zorgen. De situatie en het bedrijf waarin het data mining project moet uitgevoerd worden, hebben eveneens een invloed op de keuze van een bepaalde data mining techniek of een combinatie van data mining technieken. Een bedrijf zal dus zijn e-business campagne proberen te optimaliseren aan de hand van zorgvuldig geselecteerde data mining technieken en evaluatiemethodes.

1.2 Centrale onderzoeksvraag

Het kiezen van de juiste evaluatiemethode voor een specifieke data mining techniek is niet altijd evident. Er bestaan namelijk verschillende methodes om de resultaten van data mining technieken te beoordelen. Deze evaluatiemethodes kunnen daarbij ook verschillend zijn voor elke data mining techniek. Aangezien de keuze van een dergelijke methode een invloed kan hebben op het succes van een data mining onderzoek, is het belangrijk om de meest optimale methode te vinden.

In september 2004 tijdens de vijftiende conferentie over Machine Leren in Pisa (Italië), werd er door Isabelle Alvarez een nieuwe methode voorgesteld om het resultaat gegeven door een binaire beslissingsboom te kwalificeren. Het artikel *Sensitivity Analysis of the Result in Binary Decision Trees: Giving More Information to the End-user*, geschreven door Isabelle Alvarez,

beschreef deze methode en formuleerde hiervoor een algoritme. In dit artikel wordt ook aangehaald dat het sensitiviteitsalgoritme gebruikt kan worden om de resultaten gegeven door een binaire beslissingsboom te beoordelen en te rangschikken. Aan de hand van dit artikel hebben we de centrale onderzoeksvraag opgesteld.

Op welke manier kan het sensitiviteitsalgoritme gebruikt worden als evaluatiemethode voor de resultaten van binaire beslissingsbomen?

1.3 Werkwijze

Dit onderzoek is zowel theoretisch als praktijkgericht. In het theoretische gedeelte wordt een inleiding gegeven tot data mining en wordt er dieper ingegaan op de data mining techniek beslissingsbomen. Omdat een groot aantal begrippen betreffende data mining Engels zijn, kiezen we ervoor om deze niet allemaal te vertalen naar het Nederlands omdat zij dan misschien hun betekenis verliezen.

Het praktijkgedeelte van deze thesis bestaat uit het implementeren van een bestaand algoritme, meer bepaald het sensitiviteitsalgoritme ontwikkeld door Isabelle Alvarez. Dit algoritme kan gebruikt worden voor de sensitiviteitsanalyse van het resultaat van binaire beslissingsbomen zodat er meer informatie kan gegeven worden aan de eindgebruiker. In het kader van deze thesis zal, aan de hand van een gevalstudie, bekeken worden of het algoritme ook zijn vruchten afwerpt voor het beoordelen en rangschikken van de resultaten gegeven door binaire beslissingsbomen. Er is namelijk nood aan een methode voor het ordenen van de gevallen die reeds geclassificeerd werden door een beslissingsboom. Deze gevallen rangschikken volgens betrouwbaarheid, ook wel waarschijnlijkheid of waarschijnlijkheidsschatting genoemd, is namelijk niet altijd de beste werkwijze. Dit omdat de betrouwbaarheid die door bestaande beslissingsboomalgoritmes berekend wordt, hetzelfde is voor de verschillende gevallen uit eenzelfde blad. Hierdoor is de betrouwbaarheid dus eigenlijk geen goede, nauwkeurige meting voor de waarschijnlijkheid dat een geval de juiste klasse toegekend heeft gekregen (Zadrozny en Elkan 2001, Provost en Domingos 2000, Margineantu en Dietterich 2001).

1.4 Overzicht inhoud

Deze thesis is georganiseerd in vijf hoofdstukken:

- Hoofdstuk 2 is een inleidend hoofdstuk dat de basisconcepten van data mining bespreekt. Er wordt getracht een allesomvattende definitie voor data mining te geven maar ook het doel en de activiteiten van data mining worden kort aangehaald.
- In hoofdstuk 3 wordt er dieper ingegaan op beslissingsbomen. Dit is een techniek die gebruikt kan worden voor data mining. De werking evenals het scoren met beslissingsbomen en eventuele problemen hieromtrent zullen uitgelegd worden in dit hoofdstuk.
- De data mining software Weka en de implementatie van het algoritme zullen in hoofdstuk 4 besproken worden.
- Hoofdstuk 5 handelt over de toepassing van het algoritme in een werkelijke situatie. De werking van het algoritme zal beoordeeld worden door het toe te passen op de data die verkregen werden in het kader van een internationale data mining competitie.
- In het laatste hoofdstuk, hoofdstuk 6, formuleren we enkele conclusies betreffende het onderzoek van deze thesis en de resultaten ervan. Er zullen ook enkele suggesties en aanbevelingen voor verder onderzoek gedaan worden.

Hoofdstuk 2

Data mining

In dit hoofdstuk bespreken we enkele basisconcepten van data mining. Een aantal bestaande definities zullen gebruikt worden om een zo volledig mogelijke definitie van data mining te geven. Ook het verschil tussen experimentele en observationele data wordt uitgelegd. Als laatste zullen we dan de verschillende activiteiten van data mining kort beschrijven.

2.1 Definitie van data mining

Data mining kan op verschillende manieren gedefinieerd worden. Hieronder volgen enkele definities die we in de vakliteratuur zijn tegengekomen.

> *Data mining is het analyseren van observationele datasets om ongekende relaties te ontdekken en om de data samen te vatten op manieren die zowel begrijpbaar als nuttig zijn voor de eigenaar van de data (Hand* et al. *2001).*
>
> *Data mining is de verkenning en analyse van grote hoeveelheden data op automatische of semi-automatische manieren om zo betekenisvolle patronen en regels te ontdekken (Berry en Linoff 1997).*
>
> *Data mining is het verkennen en modelleren van relaties uit grote hoeveelheden data met behulp van geavanceerde methodes (Walsh 2003).*

Zoals Hand, Mannila en Smyth weergeven in hun definitie, kan data mining gebruik maken van observationele data. Dit zijn data die niet door een bedrijf worden aangemaakt, verzameld en bijgehouden om gebruikt te worden voor data mining, maar om bijvoorbeeld een gedetailleerd klantenbestand te maken (Hand *et al.* 2001). Data die wel aangemaakt en verzameld worden met het oog op onderzoek noemt men experimentele data. Tabel 2.1 (Walsh 2003) geeft een overzicht van de verschillen tussen deze twee soorten data.

Tabel 2.1: Vergelijking experimentele en observationele data

Data	Experimenteel	Observationeel
Doel	Onderzoek	Operationeel
Waarde	Wetenschappelijk	Commercieel
Ontstaan	Actief beheerst	Passief geobserveerd
Hoeveelheid	Klein	Enorm groot
Aard	Zuiver en regelmatig	Onzuiver en onregelmatig
Toestand	Statisch	Dynamisch

De bovenstaande definities van data mining beschrijven data mining elk op hun manier. Als we de verschillende aspecten van deze drie definities combineren, komen we tot de volgende definitie voor data mining:

Data mining is het verkennen en analyseren van grote hoeveelheden observationele data met als doelen het samenvatten van de data op een nuttige en begrijpbare manier voor de eigenaar van de data en het ontdekken, verkennen en modelleren van eventuele betekenisvolle relaties, regels en patronen in de beschikbare data.

2.2 Doel van data mining

Het doel van data mining is de onderneming inzicht laten krijgen in de inhoud en de samenhang van de verzamelde observationele data. Hierdoor kan het bedrijf ongekende patronen, relaties en trends ontdekken en modelleren. Data mining kan als een hulpmiddel bij onderzoek of verbeteringen in het bedrijf gebruikt worden. De meest succesvolle toepassingen van data mining zijn echter terug te vinden in databasemarketing. Met database wordt hier de verzameling van gegevens over potentiële of bestaande klanten bedoeld. Data mining kan bijvoorbeeld gebruikt worden om te bepalen welke mensen zullen reageren op een bepaalde aanbieding uit een campagne. Een ander toepassingsgebied uit het bedrijfsleven is Customer Relationship Management. Het bedrijf kan met behulp van data mining zijn klanten beter leren kennen wat betreft hun behoeften, verwachtingen en gedrag (Berry en Linoff 2000).

2.3 Activiteiten van data mining

De activiteiten van data mining, ook wel taken genoemd, kunnen op verschillende manieren ingedeeld worden. De indeling van Michael J.A. Berry en Gordon Linoff in de boeken *Mastering Data Mining* en *Data Mining Techniques For Marketing, Sales and Customer Support* lijkt ons echter het meest geschikt omdat deze indeling correspondeert met de redenen die een persoon, een instelling of een bedrijf kan hebben om data te analyseren.

Alvorens deze verschillende activiteiten van data mining kort te bespreken, maken we een onderscheid tussen supervised en unsupervised data mining. Bij supervised data mining kan het uiteindelijke model vergeleken worden met een 'zwarte doos' omdat we enkel aandacht hebben voor de voorspellingen die het model maakt en niet voor de manier waarop het model werkt. Unsupervised data mining is data mining waarbij we wel willen weten hoe het model werkt en hoe het tot zijn resultaten komt. Men zal hierbij dus patronen en gelijkenissen tussen groepen van data zoeken maar zonder eerst een doelvariabele of voorgedefinieerde klassen te specificeren. Beide soorten

data mining kunnen naast elkaar gebruikt worden. Eerst kan unsupervised data mining gebruikt worden om meer inzicht te krijgen in het domein en het algoritme. Daarna kan supervised data mining, op basis van de eerder verworven inzichten, toegepast worden om meer specifieke vragen over de data te beantwoorden. Vereenvoudigd voorgesteld zal unsupervised data mining de verschillende klassen ontdekken terwijl supervised data mining de records zal toewijzen aan de reeds ontdekte en gekende klassen. De data mining activiteiten kunnen toegewezen worden aan één van deze twee soorten data mining, wat wordt weergegeven in tabel 2.2 (Berry en Linoff 1997).

Tabel 2.2: Activiteiten van data mining

Supervised data mining	Unsupervised data mining
Classificatie	Affinity Grouping
Schatten	Clustering
Voorspellen	Beschrijven en visualiseren

2.3.1 Classificatie

Bij classificatie wordt een model ontwikkeld dat toegepast kan worden op data die nog niet geclassificeerd werden. De kenmerken van nieuwe records in de database worden onderzocht en gebruikt om de records toe te wijzen aan een eerder gedefinieerde klasse (Berry en Linoff 2000). Kenmerkend voor deze classificatietaak is het gebruik van een training dataset met reeds geclassificeerde records en duidelijke, nauwkeurige definities voor de klassen. De waarden die klassen kunnen aannemen zijn categorisch en discreet. Een voorbeeld hiervan is het classificeren van een leningaanvraag als laag, middelmatig of hoog risico (Berry en Linoff 1997).

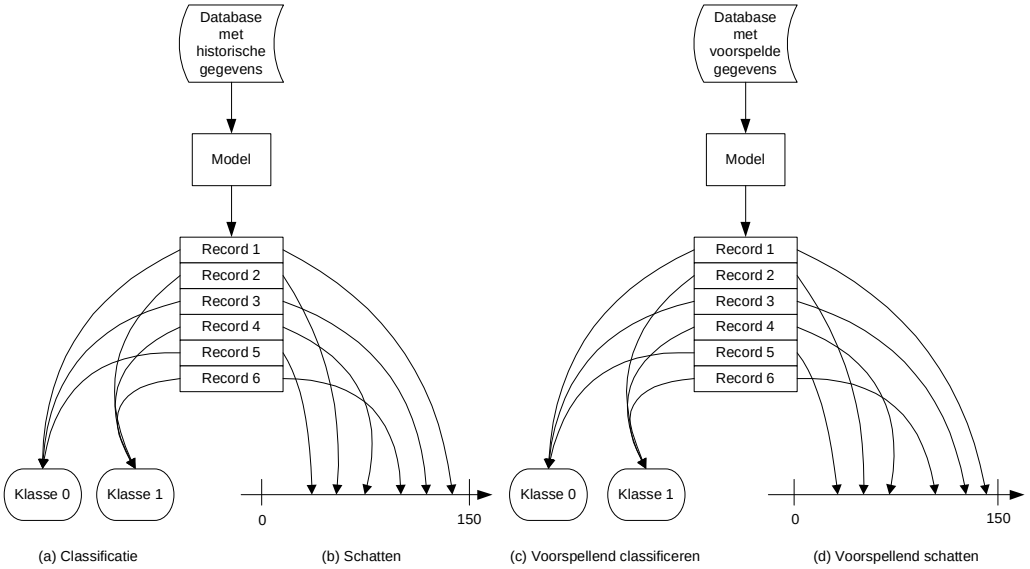
2.3.2 Schatten

De taak van de schattingsactiviteit is het ontwikkelen van een model waarmee men de waarde van een kwantitatieve en continue variabele kan bepalen. Het schatten van de leeftijd of het gewicht van een huisdier is een voorbeeld van deze taak. Schatten kan echter ook gebruikt worden om een classificatietaak uit te voeren met het verschil dat de klassen dan kwantitatieve, continue waarden kunnen aannemen (Berry en Linoff 2000).

2.3.3 Voorspellen

Deze activiteit van data mining heeft min of meer dezelfde functie als de classificatie- en schattingstaken. Het enige verschil is dat het ontwikkelde model een variabele in de toekomst zal classificeren of schatten aan de hand van geschatte of voorspelde waarden of op basis van voorspeld gedrag van de betreffende variabelen in de toekomst. Bij de classificatie- en schattingsactiviteiten wordt daarentegen gebruikt gemaakt van historische data (Berry en Linoff 2000). Dit onderscheid is ook duidelijk zichtbaar in figuur 2.1.

Het doel van deze activiteit is dus een model ontwikkelen dat de waarde van een variabele voorspelt aan de hand van reeds voorspelde waarden van andere variabelen uit de dataset (Hand *et al.* 2001).



Figuur 2.1: Verschil tussen classificatie, schatten en voorspellen

2.3.4 Affinity Grouping

De taak van affinity grouping bestaat eruit te bepalen welke transacties en/of gebeurtenissen samen plaatsvinden of welke dingen samenhoren. Een bekend voorbeeld hiervan is *Market Basket Analysis* waarbij men gaat onderzoeken welke producten meestal samen aangekocht worden in een warenhuis (Berry en Linoff 2000). Door zich op deze activiteit van data mining toe te spitsen

kan een bedrijf bijvoorbeeld bepalen welke producten en/of diensten het best samen aangeboden worden.

2.3.5 Clustering

Clustering wordt gebruikt om een heterogene populatie te segmenteren in homogene subgroepen. Het verschil tussen clustering en classificatie is dat er bij clustering geen voorbeelden of voorgedefinieerde klassen aanwezig zijn zoals dit bij classificatie wel het geval zou zijn. Records worden gegroepeerd op basis van hun overeenkomstige kenmerken (Berry en Linoff 1997).

2.3.6 Beschrijven en visualiseren

De taak van het beschrijven en visualiseren van data komt neer op het verkennen van de data zonder dat men hierbij een bepaald doel voor ogen heeft. Een veelgebruikt synoniem voor deze activiteit is *Explorative Data Analysis* (Hand *et al.* 2001). Het doel van deze activiteit is dus het beschrijven en visueel voorstellen van de data in de database zodat de gebruiker een beter inzicht krijgt in de data en de objecten die door deze data voortgebracht werden (Berry en Linoff 2000).

Hoofdstuk 3

Beslissingsbomen

Er bestaan verschillende data mining technieken maar in deze thesis wordt enkel de techniek van beslissingsbomen behandeld. De reden hiervoor is dat het sensitiviteitsalgoritme dat we gaan onderzoeken alleen op deze data mining techniek toegepast kan worden. De eerste paragraaf van dit hoofdstuk geeft definities voor enkele veelgebruikte termen en verder zal dit hoofdstuk de verschillende types van beslissingsbomen bespreken. Ook de manier waarop de resultaten van deze data mining techniek momenteel geordend worden en de problemen die hiermee gepaard gaan zullen kort aangehaald worden. Als laatste stellen we dan een reeds bestaand algoritme voor dat eventueel gebruikt kan worden om de resultaten van beslissingsbomen te ordenen.

3.1 Definities

In deze paragraaf geven we definities voor enkele veelgebruikte termen. De lezer die al enige ervaring heeft met beslissingsbomen en de terminologie hieromtrent kan deze paragraaf overslaan.

Attribuut
Een kenmerk van een object. Het object auto heeft bijvoorbeeld als attributen bouwjaar, type, merk en kleur (Negnevitsky 2001).

Beslissingsboom
Grafische voorstelling van een dataset. De data worden beschreven aan de hand van een boomachtige structuur. Een beslissingsboom is samengesteld uit knopen, takken en bladeren zoals ook zichtbaar is in figuur 3.1 op p 13. De boom start altijd vanuit de wortelknoop en groeit doordat de data verder gesplitst worden in nieuwe knopen op ieder niveau (Negnevitsky 2001).

Binaire beslissingsboom
Beslissingsboom waarbij er in iedere knoop niet meer dan twee takken of kinderen zijn.

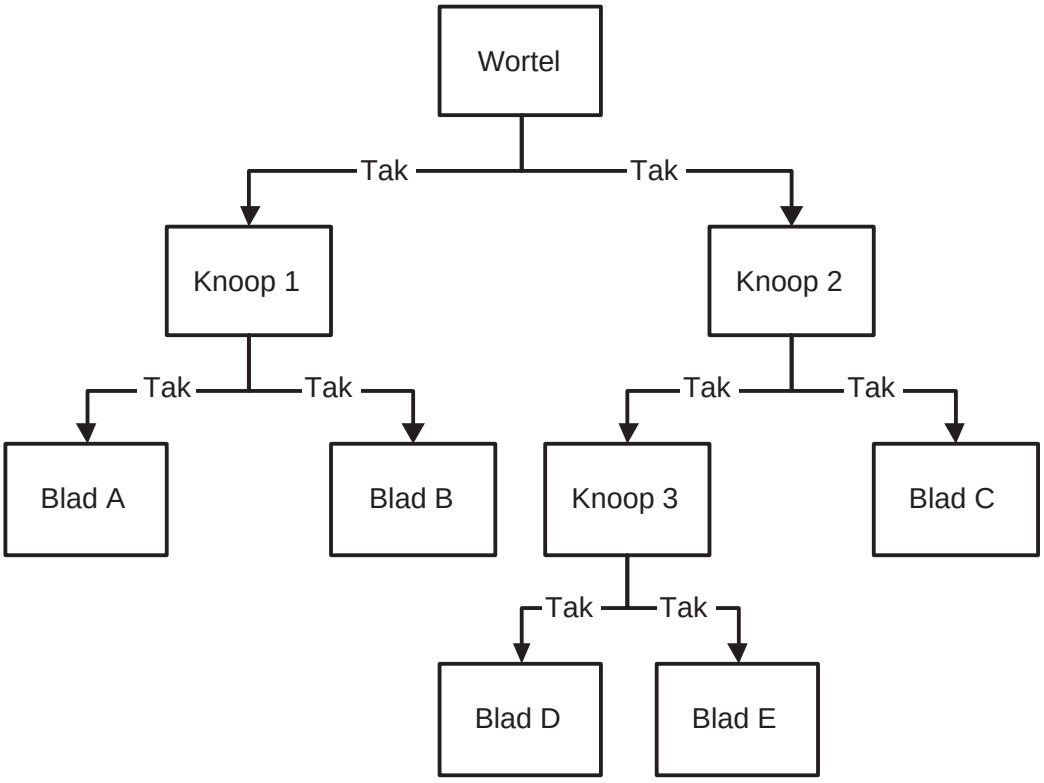
Blad
Een knoop van een beslissingsboom aan het uiteinde van een tak of met andere woorden een knoop zonder kinderen. Dit wordt ook wel een eindknoop genoemd (Negnevitsky 2001). Voorbeelden hiervan zijn Blad A tot en met E op figuur 3.1 op p 13.

Categorische data
Data die door een klein aantal discrete categorieën geclassificeerd kunnen worden (Negnevitsky 2001). Categorische data kunnen geordend (ordinaal) of ongeordend (nominaal) zijn.

Conjunctie
Een product van AND-operatoren (Weisstein 2005a). In het kader van beslissingsbomen komt een conjunctie van attribuuttesten overeen met een pad van de wortel van de boom tot in een blad (Mitchell 1997). Dit is eigenlijk de weg die een instantie aflegt of moet afleggen om in een specifiek blad terecht te komen en geclassificeerd te worden.

Continue data
Data die een oneindig aantal waarden kunnen aannemen binnen een

bepaald interval. Voorbeelden van continue data zijn temperatuur, gewicht en oppervlakte (Negnevitsky 2001).



Figuur 3.1: Voorbeeld van een (binaire) beslissingsboom

Data
Gegevens over fysieke verschijnselen of zakelijke transacties die verzameld werden met behulp van objectieve metingen van de attributen van mensen, plaatsen, dingen en/of gebeurtenissen (O'Brien 1998).

Dataset
Een verzameling data of een groep gevallen, instanties of records.

Discrete data
Data die slechts een eindig aantal verschillende waarden kunnen aannemen. Voorbeelden hiervan zijn het aantal kamers in een huis en het aantal zitplaatsen in een auto (Negnevitsky 2001).

Disjunctie
De bewerkingsoperator OR (Weisstein 2005b). Een beslissingsboom

kan gezien worden als een disjunctie van conjuncties van beperkingen op de attribuutwaarden van instanties (Mitchell 1997).

Disjunctie van conjuncties
Een verzameling van verschillende mogelijkheden. In de context van beslissingsbomen komt dit overeen met de paden die de instanties uit een dataset in een beslissingsboom kunnen afleggen alvorens geclassificeerd te worden.

Inductie
Een techniek die veralgemeningen maakt op basis van informatie die uit de dataset werd gehaald (Two-Crows-Corporation 2005).

Instantie
Een specifiek object van een klasse ook wel een geval of een record genoemd (Negnevitsky 2001).

Kans (Probabiliteit, waarschijnlijkheidsschatting)
Een kwantitatieve beschrijving van de waarschijnlijkheid van een bepaalde gebeurtenis. Probabiliteit wordt wiskundig uitgedrukt als een getal tussen 0 en 1 maar het kan ook als een percentage uitgedrukt worden (Negnevitsky 2001).

Kind
Een knoop geproduceerd door het splitsen van de data van de knoop in het bovenliggende hiërarchische niveau van de beslissingsboom. Een kind bevat dus een subset van de dataset die vervat ligt in de ouder (Negnevitsky 2001). In figuur 3.1 op p 13 is blad A een kind van Knoop 1 en Knoop 3 een kind van Knoop 2.

Klasse
Een groep van objecten, gevallen of instanties met gezamenlijke kenmerken of overeenkomstige attribuutwaarden (Negnevitsky 2001).

Knoop
Een beslissingspunt van een beslissingsboom (Negnevitsky 2001). Voor een grafische voorstelling van een knoop verwijzen we naar figuur 3.1 op p 13.

Nominale data
Ongeordende categorische data zoals bijvoorbeeld geslacht dat de waarden mannelijk of vrouwelijk kan aannemen (Two-Crows-Corporation 2005).

Ordinale data
Geordende categorische data zoals bijvoorbeeld temperatuur die laag, middelmatig of hoog kan zijn (Two-Crows-Corporation 2005).

Ouder
Een knoop in een beslissingsboom die data verdeelt tussen knopen op het volgende hiërarchisch niveau van de beslissingsboom. De ouderknoop bevat een complete dataset, terwijl de kinderen van deze knoop subsets van deze dataset zullen bevatten (Negnevitsky 2001). In figuur 3.1 op p 13 zijn zowel knopen 1 tot en met 3 als de wortel voorbeelden van ouders.

Pad
De weg die een instantie, geval of record aflegt van in de wortel tot in een blad van een beslissingsboom. Dus een conjunctie van attribuuttesten.

Record
Een set van specifieke attribuutwaarden die bij een specifiek object horen. Een record is eigenlijk een rij in een database (Negnevitsky 2001). Soms wordt dit ook wel een geval of een instantie genoemd.

Tak
De verbinding tussen twee knopen in een beslissingsboom (Negnevitsky 2001). Zie ook figuur 3.1 op p 13.

Test set
Een dataset die gebruikt wordt om de prestaties van een beslissingsboom of een model te testen of te evalueren. Deze dataset is volledig onafhankelijk van de training set en bevat gevallen die de boom nog niet eerder gezien heeft. Eens de training van de beslissingsboom voltooid is zal de test set als validatie gebruikt worden (Negnevitsky 2001).

Training set
Een dataset gebruikt voor het trainen of ontwikkelen van een beslissingsboom of een beslissingsmodel (Negnevitsky 2001).

Wortel
Dit is de bovenste knoop van een beslissingsboom en wordt soms ook de wortelknoop van een beslissingsboom genoemd. De boom begint altijd vanuit deze knoop en groeit naar beneden door de data op ieder niveau verder in nieuwe knopen te splitsen. De wortelknoop bevat de volledige dataset (alle records, gevallen of instanties) en de kinderknopen bevatten subsets van deze dataset (Negnevitsky 2001). In figuur 3.1 op p 13 is ook een wortelknoop aanwezig.

3.2 Inleiding

Leren met behulp van beslissingsbomen is een manier van leren waarbij men discrete doelfuncties tracht te benaderen. De bestudeerde functie zal dan voorgesteld worden door een beslissingsboom (Mitchell 1997). Deze grafische voorstelling van een dataset geeft een beschrijving van de data aan de hand van een boomachtige structuur. Een beslissingsboom is, net zoals een echte boom, samengesteld uit knopen, takken en bladeren. De beslissingsboom start altijd vanuit de wortelknoop en groeit dan verder doordat de dataset op ieder niveau verder opgesplitst wordt in nieuwe knopen (Negnevitsky 2001). Beslissingsbomen vertegenwoordigen dus een disjunctie van conjuncties van beperkingen op de attribuutwaarden van instanties. Hierbij correspondeert ieder pad van de wortel van de boom tot in een blad met een conjunctie of verzameling van attribuuttesten (Mitchell 1997).

Beslissingsbomen zijn vooral goed in het oplossen van classificatieproblemen. Een ander belangrijk voordeel is dat ze de data ook visualiseren (Negnevitsky 2001). Niet alle classificatieproblemen zijn echter geschikt om opgelost te worden met beslissingsbomen zoals uit volgende paragraaf zal blijken.

3.3 Problemen geschikt voor het toepassen van beslissingsbomen

Beslissingsbomen kunnen gebruikt worden voor het oplossen van classificatieproblemen van allerlei aard maar ze zijn volgens Mitchell het meest geschikt voor problemen met onderstaande karakteristieken (Mitchell 1997).

- De instanties of gevallen uit de dataset worden voorgesteld en beschreven door een vaste verzameling van attributen.
- De doelfunctie van het probleem neemt discrete waarden aan.
- Disjuncte beschrijvingen kunnen een vereiste zijn.
- Er mogen fouten in de training set aanwezig zijn want leermethodes gebaseerd op beslissingsbomen zijn niet zo gevoelig voor fouten in de training set. Met fouten bedoelen we zowel classificatiefouten als fouten in de attribuutwaarden die de instanties beschrijven.

- De trainingdata mogen attributen bevatten waarvoor er bij sommige instanties of gevallen geen waarde aanwezig is.

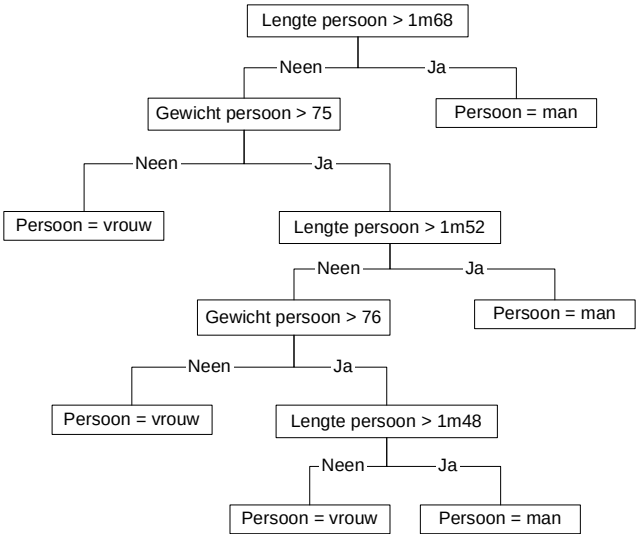
3.4 Opbouw en werking van beslissingsbomen

Een instantie wordt door een beslissingsboom geclassificeerd door te starten in de wortel van de boom en hier het attribuut gespecificeerd door deze beginknoop te testen. Daarna gaat men via de tak van de boom die overeenkomt met de waarde van het geteste attribuut naar beneden. Dit proces wordt dan telkens herhaald door de subboom in de knoop op het volgende niveau (Mitchell 1997). De instantie zal dus neerwaarts door de beslissingsboom bewegen volgens de waarden van de attributen die getest worden in de opeenvolgende knopen. Wanneer een blad bereikt wordt, zal de instantie de klasse toegewezen krijgen die aan het blad is toegekend (Witten en Frank 2000).

Aan de hand van het voorgaande kunnen we afleiden dat de diverse knopen elk hun specifieke functie in de beslissingsboom hebben. Iedere knoop in de boom, met uitzondering van een blad, specificeert een beperking of test inzake een attribuut van de instantie. Dit wordt ook duidelijk in figuur 3.2 op p 18. Elke tak die vanuit een knoop vertrekt, correspondeert met één van de mogelijke uitkomsten van de test in deze knoop (Mitchell 1997). Sommige van deze takken zullen eindigen in een blad. Het blad geeft dan een classificatie die van toepassing zal zijn op alle instanties aanwezig in het betreffende blad. Het blad kan echter ook een kansverdeling over de mogelijke classificaties geven (Witten en Frank 2000).

Zoals reeds werd aangehaald komt een knoop in een beslissingsboom overeen met een beperking of test op een bepaald attribuut. Meestal vergelijkt deze test de waarde die de instantie voor het attribuut heeft met een constante. In sommige bomen worden in een knoop twee attributen met elkaar vergeleken of wordt er een functie van één of meerdere attributen als test gebruikt (Witten en Frank 2000). In het algemeen kunnen we twee soorten attributen onderscheiden, numerieke en niet numerieke.

Als het attribuut dat getest wordt in een knoop niet numeriek is, dan is het aantal kinderen dat deze knoop zal hebben meestal gelijk aan het aantal mogelijke waarden van het attribuut. Aangezien er dan één tak is voor iedere mogelijke waarde zal het attribuut niet opnieuw getest worden naarmate men dieper in de boom terechtkomt. Het kan echter gebeuren dat de attribuut-



Figuur 3.2: Voorbeeld van een beslissingsboom

waarden in twee subsets verdeeld worden en dan zal de betreffende knoop slechts twee vertakkingen hebben. In dit geval kan het attribuut meer dan één keer getest worden doorheen de beslissingsboom (Witten en Frank 2000).

Indien het attribuut wel numeriek is, dan bepaalt de test in een knoop meestal of de attribuutwaarde strikt groter (>) of groter (>=) is, strikt kleiner (<) of kleiner (<=) is, of gelijk (=) is aan een bepaalde constante. Bij reële waarden zal er dan weer getest worden of de attribuutwaarden in een specifiek interval liggen, of groter of kleiner zijn dan de waarden in een interval. Een numeriek attribuut kan verschillende keren getest worden in een boom, waarbij iedere test dan een andere constante of een ander interval zal bevatten (Witten en Frank 2000).

3.5 Types van beslissingsbomen

Er kunnen twee types van beslissingsbomen onderscheiden worden: classificatiebomen en regressiebomen.

- *Classificatie-beslissingsbomen*: beslissingsbomen die gebruikt worden om categorische variabelen in klassen onder te brengen (Two-Crows-Corporation 2005). Deze bomen zullen de records een label toekennen

en ze hierdoor toewijzen aan een klasse (Berry en Linoff 2000). In het praktijkgedeelte van deze thesis zullen we een classificatieboom ontwikkelen.

- *Regressie-beslissingsbomen*: beslissingsbomen die de waarden van continue variabelen kunnen voorspellen (Two-Crows-Corporation 2005).

Het type beslissingsboom bepaalt ook het beslissingsboomalgoritme dat gebruikt wordt om de boom te produceren. Afhankelijk van het type beslissingsboom dat ze voortbrengen maken deze algoritmes gebruik van andere parameters of van dezelfde parameters maar met andere standaardwaarden. Voorbeelden waarop de algoritmes van elkaar kunnen verschillen zijn bijvoorbeeld het aantal splits dat toegelaten is op elk niveau van de boom of hoe deze splits gekozen worden. Enkele voorbeelden van algoritmes zijn CHAID (Chi-Square Automatic Interaction Detector), CART (Classification and Regression Trees), ID3 en C4.5/C5 (Berry en Linoff 2000). In bijlage A gaan we dieper in op het C4.5-algoritme.

3.6 Scoren met beslissingsbomen

Zoals reeds werd aangehaald in hoofdstuk 1 is het dikwijls niet voldoende om enkel te voorspellen tot welke klasse een specifiek geval behoort. Er is ook een rangschikking nodig van de gevallen die door de beslissingsboom geclassificeerd werden. Deze gevallen zouden bijvoorbeeld geordend kunnen worden volgens de kans dat het geval tot de voorspelde klasse behoort. Een probleem hierbij is echter dat de waarschijnlijkheidsschattingen van beslissingsbomen onbetrouwbaar en onnauwkeurig kunnen zijn (Zadrozny en Elkan 2001).

Uit het voorgaande kunnen we al afleiden dat beslissingsbomen niet enkel weergeven tot welke klasse een specifiek geval zal behoren. Ze geven ook de waarschijnlijkheid aan waarmee dat geval tot de voorspelde klasse zal behoren. In hetgeen wat volgt beschrijven we hoe beslissingsbomen deze waarschijnlijkheidsschattingen genereren en welke problemen hierdoor veroorzaakt kunnen worden.

Beslissingsboomalgoritmes bepalen de waarschijnlijkheidsschatting voor een specifiek geval aan de hand van de ruwe trainingsfrequentie van het blad waartoe het geval behoort. In de ruwe trainingsfrequentie wordt het aantal positieve trainingsvoorbeelden in een blad gerelateerd aan het totale aantal

trainingsvoorbeelden van dat blad (Zadrozny en Elkan 2001). Met positieve trainingsvoorbeelden bedoelen we gevallen waarvan de klasse waartoe ze behoren overeen komt met de klasse die verbonden is aan het blad waar ze deel van uitmaken. Formule 3.1 kan gebruikt worden om de ruwe trainingsfrequentie van een blad te berekenen.

$$\begin{aligned} n &= totaal\ aantal\ trainingsvoorbeelden\ in\ een\ bepaald\ blad \\ k &= aantal\ positieve\ trainingsvoorbeelden\ in\ een\ bepaald\ blad \\ P &= waarschijnlijkheidsschatting \\ &= ruwe\ trainingsfrequentie \\ &= \frac{k}{n} \end{aligned} \tag{3.1}$$

Algoritmes om beslissingsbomen te ontwikkelen leggen de nadruk op het maximaliseren van de classificatienauwkeurigheid en het minimaliseren van de grootte van de beslissingsboom (Provost en Domingos 2000). Dit in combinatie met de manier waarop deze beslissingsboomalgoritmes de waarschijnlijkheidsschattingen genereren zorgt voor enkele problemen.

- *Extreme waarden voor de waarschijnlijkheidsschattingen:* beslissingsboomalgoritmes proberen de bladeren van de te ontwikkelen beslissingsboom zo homogeen mogelijk maken. De geobserveerde frequenties van de positieve trainingsvoorbeelden zullen hierdoor automatisch naar 0 en 1 verschuiven (Zadrozny en Elkan 2001, Provost en Domingos 2000). Met homogeen bedoelen we dat er geen of zo weinig mogelijk gevallen van verschillende klassen in een blad zitten.

- *Waarschijnlijkheidsschattingen zijn niet statistisch betrouwbaar:* dit geldt vooral wanneer het aantal trainingsvoorbeelden geassocieerd met een blad klein is (Zadrozny en Elkan 2001). Stel dat we een blad hebben dat samengesteld is uit gevallen die allen tot dezelfde klasse behoren. Het is moeilijk om te aanvaarden, zeker wanneer er slechts een klein aantal gevallen tot dit specifieke blad behoort, dat andere gevallen die zullen voldoen aan de beperkingen opgelegd door dit blad ook tot dezelfde klasse zullen behoren (Provost en Domingos 2000).

- *Er is slechts één waarschijnlijkheidsschatting per blad:* iedere instantie van hetzelfde blad krijgt dezelfde waarschijnlijkheidsschatting van het beslissingsboomalgoritme toegewezen (Provost en Domingos 2000).

> We weten dat ieder blad van een beslissingsboom overeenkomt met een zekere beslissingsruimte. Ondanks het feit dat de verschillende gevallen uit een blad zich op verschillende punten in deze ruimte bevinden, krijgen ze toch allemaal dezelfde waarschijnlijkheidsschatting toegekend door het beslissingsboomalgoritme (Margineantu en Dietterich 2001).

Met onderstaande voorbeelden trachten we deze problemen wat te verduidelijken. Het eerste voorbeeld werd overgenomen uit de paper *Improved Class Probability Estimates from Decision Tree Models* van Dragos D. Margineantu en Thomas G. Dietterich (Margineantu en Dietterich 2001). Het andere voorbeeld komt uit het artikel van Foster Provost en Pedro Domingos, *Well-Trained PETs: Improving Probability Estimation Trees* (Provost en Domingos 2000).

Voorbeeld 1: kans op hartziekte

Stel dat klasse Y samengesteld is uit personen die kans hebben op een hartziekte. De kans op een hartziekte gaan we schatten met behulp van het attribuut bloeddruk *bp*. Stel dat er een blad is in de ontwikkelde beslissingsboom dat overeenkomt met de beperking $bp > 160$. Als er nu van de 100 gevallen die in dit blad terecht zullen komen zo'n 90 personen zullen zijn met een hartziekte dan wordt de waarschijnlijkheid dat een geval tot klasse Y zal behoren geschat op $P(Y|x) = 0,90$. Er zullen wellicht van deze 100 gevallen personen zijn waarvan de bloeddruk groter is dan die van andere personen in dit blad. Het is haast vanzelfsprekend dat we een grotere kans op hartziekte willen toewijzen aan personen met bloeddruk $bp = 250$ dan aan personen met een lagere bloeddruk zoals bijvoorbeeld $bp = 161$ (Margineantu en Dietterich 2001).

Voorbeeld 2: blad dat is samengesteld uit een klein aantal gevallen

Stel dat we een blad hebben in een beslissingsboom dat vijf gevallen bevat die allemaal tot dezelfde klasse behoren. Het beslissingsboomalgoritme zal aan dit blad dan een waarschijnlijkheidsschatting van $P = \frac{5}{5} = 1$ toekennen. Hierdoor zal elk geval dat aan de beperkingen of voorwaarden van dit blad zal voldoen, toegewezen worden aan de klasse die toegekend is aan dit blad. Hierbij is het niet onterecht om op te merken dat een aantal van vijf gevallen eigenlijk niet voldoende is om een voorspelling te maken omtrent het al dan niet behoren tot een klasse (Provost en Domingos 2000).

3.7 Sensitiviteitsalgoritme

In deze paragraaf bespreken we het sensitiviteitsalgoritme van Isabelle Alvarez (Alvarez 2004). Eerst definiëren we enkele termen die belangrijk zijn in het kader van dit algoritme en als laatste volgt dan de bespreking van het sensitiviteitsalgoritme.

3.7.1 Enkele definities

Afstand

$$d = \|x - y\| = \sqrt{\sum_{i=1}^{a} (x_i - y_i)^2}$$

Beslissingsoppervlak

De grens of het grensoppervlak tussen gebieden met een verschillend klasselabel (Alvarez 2004).

Euclidische ruimte

Een Euclidische n-ruimte, soms ook wel een Cartesische ruimte of simpelweg n-ruimte genoemd, is de ruimte van alle n-tuppels van reële getallen, $(x_1, x_2, \ldots, x_n)$. Het wordt meestal genoteerd als $\Re^n$ en het is een vectorruimte waardoor de elementen ervan n-vectoren genoemd worden (Weisstein 2005e).

Element

Als x deel uitmaakt van een set A dan wordt x een element van A genoemd (Weisstein 2005d). Onderstaande notatie wordt gebruikt om aan te geven dat x een element is van A:

$$x \in A$$

Hypervlak

Laat $a_1, a_2, \ldots, a_n$ scalairen zijn die niet allemaal gelijk zijn aan nul. De set

$$X = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

die bestaat uit alle vectoren in $\Re^n$ zodat

$$a_1x_1 + a_2x_2 + \ldots + a_nx_n = 0$$

is een subruimte van $\Re^n$ en wordt een hypervlak genoemd.

Lineaire beslissingsboom
Een beslissingsboom waarvan de testen in de knopen gebaseerd zijn op een lineaire combinatie van attributen (Murthy 1998).

Metrische ruimte
Een set S met een globale afstandsfunctie g die voor iedere twee punten x en y in S de afstand ertussen geeft als een niet-negatief reëel getal $g(x, y)$. Deze afstandsfunctie moet eveneens voldoen aan onderstaande voorwaarden:

1. $g(x, y) + g(y, z) \geq g(x, z)$ $(driehoeksongelijkheid)$
2. $g(x, y) = g(y, x)$ $(symmetrie)$
3. $g(x, y) = 0 \Leftrightarrow x = y$

Niet-lineaire beslissingsboom
Een beslissingsboom waarvan de testen niet-lineaire combinaties van attributen zijn (Ittner en Schlosser 1998).

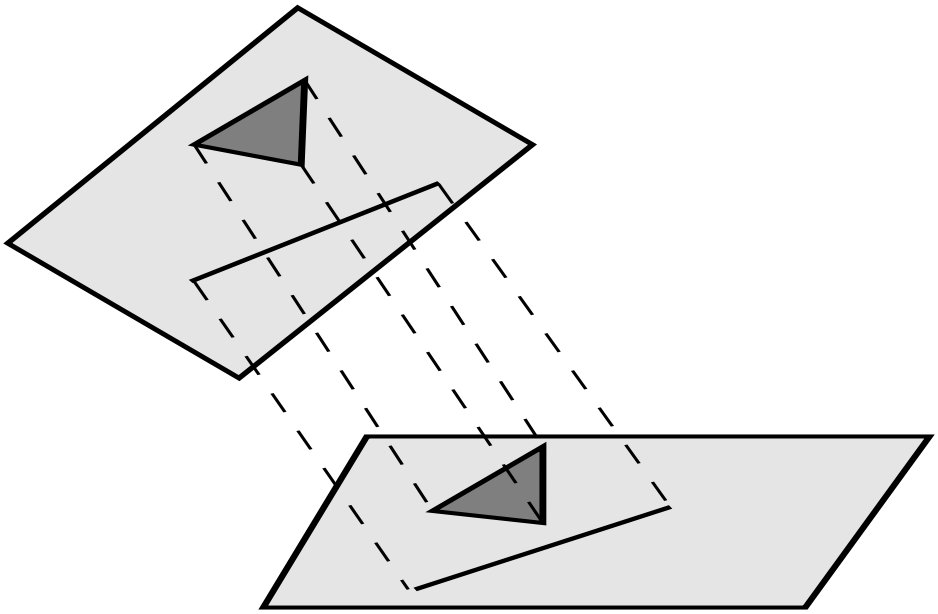
n-tuppel
Wanneer n niet gekend is wordt dit simpelweg een tuppel genoemd. Het is een synoniem voor een lijst van een geordende set van n elementen. Het kan ook geïnterpreteerd worden als een n-vector (Weisstein 2005f).

Open bol
Een n-dimensionale open bol $B_r(p)$ met straal r en middelpunt p is de verzameling van punten waarvoor de afstand tot p strikt kleiner is dan r (Weisstein 2005g). In formulevorm geeft dit:

$$B_r(p) = \{y : \|y - p\| < r\}$$

Projectie
De transformatie van punten en lijnen uit een vlak op een ander vlak door de corresponderende punten op de twee vlakken te verbinden met parallelle lijnen (Weisstein 2005h). Dit is ook zichtbaar in figuur 3.3 op p 24.

Scalair
Een getal in één dimensie (Weisstein 2005i).



Figuur 3.3: Projectie

Sensitiviteit

De sensitiviteit in punt x is $d(x, \Gamma)$ (Alvarez 2004). Hierbij wordt $d(x, \Gamma)$ gezien als de afstand tussen geval x en het bijhorende beslissingsoppervlak Γ.

Sensitiviteitsverplaatsing

De sensitiviteitsverplaatsing in punt x is de vector $\overrightarrow{xp(x)}$ met $d(x, p(x)) = d(x, \Gamma)$ (Alvarez 2004). Hierbij wordt $d(x, \Gamma)$ gezien als de afstand tussen geval x en het bijhorende beslissingsoppervlak Γ. De afstand tussen het inputgeval x en de projectie van x op het beslissingsoppervlak Γ wordt gegeven door $d(x, p(x))$.

Set

Een verzameling van objecten waarbij rangorde van geen belang is (Weisstein 2005j).

Vector

Een vector A wordt bepaald door n coördinaten en kan gespecificeerd worden als $(A_1, A_2, \ldots, A_n)$. Een vector van punt A tot punt B wordt genoteerd als $\overrightarrow{AB}$ (Weisstein 2005k).

3.7.2 Sensitiviteitsalgoritme

Deze paragraaf is gebaseerd op het artikel *Sensitivity Analysis of the Result in Binary Decision Trees: Giving More Information to the End-user*, geschreven door Isabelle Alvarez (Alvarez 2004).

Wanneer de attribuutwaarden van de gevallen uit een dataset numeriek zijn, kan met behulp van dit algoritme de afstand berekend worden tussen een geval uit de dataset en het daarbij horende beslissingsoppervlak. Dit beslissingsoppervlak is samengesteld uit delen van verschillende hypervlakken en niet uit één uniek hypervlak. De reden hiervoor is dat voor ieder geval het bijhorende beslissingsoppervlak bepaald wordt door de testen die het geval in de beslissingsboom moet ondergaan alvorens het in een specifiek blad terechtkomt. De berekende afstand kan dan gebruikt worden om de waarschijnlijkheidsschattingen geval-afhankelijk te maken. De waarschijnlijkheidsschattingen van beslissingsbomen zijn namelijk niet altijd even betrouwbaar en nauwkeurig zoals in de paragraaf *Scoren met beslissingsbomen* reeds besproken werd.

Het sensitiviteitsalgoritme werd ontwikkeld om te werken op beslissingsbomen waarvan de hypervlakken loodrecht staan op een attribuutas. Deze hypervlakken worden door de testen in de beslissingsboom gedefinieerd. Er zijn twee situaties waarin de hypervlakken van een beslissingsboom loodrecht staan op een attribuutas.

- *Situatie 1:* er wordt een lineaire beslissingsboom gebruikt en de testen in deze beslissingsboom bevatten slechts één attribuut.
- *Situatie 2:* men maakt gebruik van een niet-lineaire beslissingsboom. Bij dit soort beslissingsbomen staan de hypervlakken altijd loodrecht op een attribuutas.

Een beslissingsboom veroorzaakt een verdeling van de inputruimte E. Laat x een punt zijn van de inputruimte E en laat Γ het beslissingsoppervlak zijn dat bij x hoort. De sensitiviteit in punt x is dan de afstand tussen x en het beslissingsoppervlak Γ. Om deze afstand, $d(x, \Gamma)$, te kunnen berekenen moet x geprojecteerd worden op het beslissingsoppervlak Γ.

Ieder punt in de open bol met centrum x en straal $r = d(x, \Gamma)$ heeft dezelfde voorspelde klasse als x. Er bestaat echter minstens één punt in eender welke open bol met centrum x en straal $r + \epsilon$ waarvoor de klasse verschillend is

van deze van x. Als E daarbij een volledige metrische ruimte is, dan bestaat er tenminste één punt $p(x)$ waarvoor $d(x, p(x)) = d(x, \Gamma)$. Indien E ook een Euclidische ruimte is, dan is $p(x)$ de projectie van x op Γ.

Laat f het blad zijn waartoe x behoort en laat $(h(x, H_i))_{i \in I})$ de lijst van testen zijn die x moet ondergaan om in f terecht te komen. De set C_f van gevallen geclassificeerd door blad f is een hyperrechthoek en we associëren met dit blad f de lijst van testen $(h(x, H_i))_{i \in I})$. Het blad f zal enkel de gevallen classificeren die tot de doorsnede C_f van de halfruimten $E(Hi)$ behoren. Deze halfruimten werden gevormd door de testen die het geval x moest ondergaan om in het blad f terecht te komen. De doorsnede C_f van de halfruimten $E(Hi)$ kan met behulp van formule 3.2 voorgesteld worden.

$$C_i = \bigcap_{i \in I} E(H_i) \tag{3.2}$$

Het algoritme $sensitivityAt(x, DT)$ berekent de sensitiviteit en de sensitiviteitsverplaatsing in punt x van beslissingsboom DT. Hiervoor wordt de afstand berekend tussen x en de bladeren waarvan het klasselabel $c(f)$ verschillend is dan dat van de voorspelde klasse $c(x)$ van x. Om deze afstanden te kunnen berekenen moet het geval x echter eerst geprojecteerd worden op ieder blad f waarvan $c(f) \neq c(x)$. Dit gebeurt met behulp van een tweede algoritme, $projectionOntoLeaf(x, (E(H_i))_{i \in I})$ op p 27. Na deze projectie waarvan $p_f(x)$ het resultaat is, worden de afstanden tussen x en de bijhorende projecties berekend en gerangschikt. De kleinste afstand zal dan geselecteerd worden als de sensitiviteit in x.

Algoritme 1: $sensitivityAt(x, DT)$

1. *Verzamel de set F van bladeren f van de beslissingsboom waarvoor de klasse $c(f) \neq c(x)$*
2. *Voor iedere $f \in F$:*
3. *bereken $p_f(x) = projectionontoLeaf(x, f)$*
4. *bereken en rangschik $d(x, p_f(x))$*
5. $(d(x, p_n(x)), \overrightarrow{xp_n(x)})$ *met* $n = argmin_{f \in F}(d(x, p_f(x)))$

Algoritme 2: $projectionOntoLeaf(x, (E(H_i))_{i \in I})$

1. $y = x$
2. $Voor\ i = 1\ tot\ grootte(I)$:
3. $\quad als\ y \notin E(H_i)\ dan\ y_u = b$
4. $\quad\quad u\ de\ coördinaat\ van\ het\ attribuut\ dat\ H_i\ definieert$
5. $\quad\quad b\ de\ grenswaarde\ voor\ H_i$
6. $Geef\ de\ nieuwe\ y\ terug$

In het verdere gedeelte van deze thesis gaan we het sensitiviteitsalgoritme implementeren. We gaan ook kijken of het, al dan niet in combinatie met de waarschijnlijkheidsschattingen, gebruikt kan worden om de resultaten van binaire beslissingsbomen te rangschikken.

Hoofdstuk 4

Implementatie

Alvorens het algoritme toegepast kan worden op een gevalstudie of dataset moet het geïmplementeerd worden. Deze implementatie moet het mogelijk maken om met behulp van de door Weka ontwikkelde beslissingsboom de afstanden te berekenen die nodig zijn voor het sensitiviteitsalgoritme. In dit hoofdstuk gaan we eerst de data mining software van Weka kort bespreken en daarna geven we meer uitleg over de implementatie van het sensitiviteits-algoritme.

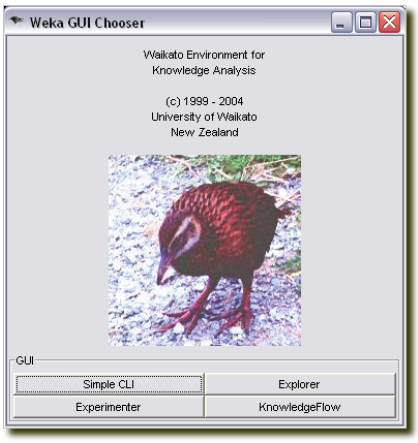
4.1 Weka

In deze paragraaf bespreken we Weka, dit is het programma dat gebruikt wordt om de beslissingsboom te ontwikkelen waarop het algoritme toegepast kan worden. Er wordt ook een korte beschrijving gegeven van het bestandsformaat dat Weka gebruikt. We baseren ons hierbij op informatie die rechtstreeks of via links ter beschikking wordt gesteld op de website van Weka, `http://www.cs.waikato.ac.nz/ml/weka/`.

4.1.1 Algemeen

Weka of *Waikato Environment for Knowledge Analysis* is software die gebruikt wordt voor het uitvoeren van verscheidene data mining taken. De onderliggende machine leeralgoritmes kunnen rechtstreeks toegepast worden op een dataset of aangeroepen worden vanuit Java-code. Tools voor zowel het analyseren als het modelleren van data aan de hand van allerlei technieken zijn opgenomen in Weka. Deze software wordt verspreid onder de open-bron-software licentie oftewel GNU General Public License (Frank 2005). Dit houdt in dat zowel de broncode als het programma zelf gratis ter beschikking wordt gesteld.

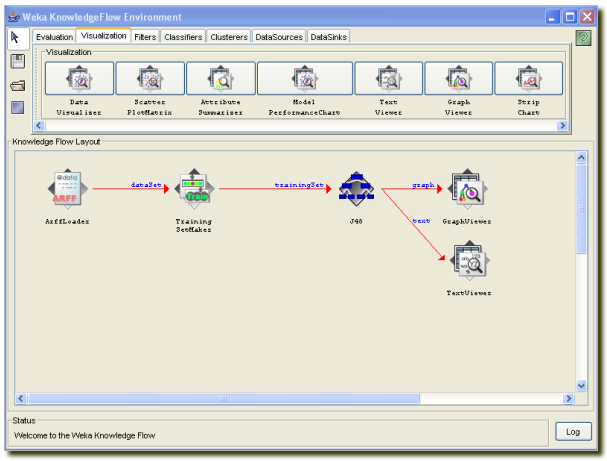
Als de gebruiker Weka opent, wordt het scherm van figuur 4.1 zichtbaar. Dit is de *GUI Chooser* waarmee de gebruikers dienen aan te geven met welk onderdeel van Weka ze willen werken.



Figuur 4.1: Startscherm van Weka

In figuur 4.1 op p 29 kan men duidelijk zien dat de gebruiker vier keuzemogelijkheden heeft zodra Weka geopend is.

- *Simple CLI:* een simpele prompt waarin men commando's kan typen. Dit onderdeel van Weka wordt gebruikt in het praktijkgedeelte van deze thesis.
- *Explorer:* dit kan gebruikt worden om interactief de gewenste data te laden, te analyseren en te verwerken. Met het analyseren en verwerken van data bedoelen we onder andere het bestuderen van de diverse attributen evenals het selecteren en visualiseren ervan. De *Explorer* zal ook aangewend kunnen worden om modellen te ontwikkelen.
- *Experimenter:* de gebruiker heeft hier de mogelijkheid om experimenten te ontwerpen en uit te voeren maar ook om ze te wijzigen en te analyseren. De *Experimenter* kan eveneens gebruikt worden om de prestaties van verschillende leeralgoritmes op diverse datasets te testen. De resultaten kunnen dan geëvalueerd en vergeleken worden.
- *Knowledge Flow:* deze interface heeft ongeveer dezelfde functionaliteiten als de *Explorer*. Het verschil is dat de gebruiker hier een visuele voorstelling krijgt van het Knowledge Discovery of data mining proces zoals zichtbaar is in figuur 4.2.



Figuur 4.2: Voorbeeld van de Weka Knowledge Flow Interface

4.1.2 ARFF bestandsformaat

Een ARFF of Attribuut-Relation File Format bestand is een ASCII tekstbestand dat een lijst van instanties of gevallen beschrijft. De gevallen van deze lijst hebben allemaal dezelfde attributen. Een ASCII tekstbestand is een bestand dat enkel karakters bevat die tot de ASCII karakterset behoren. Voorbeelden van ASCII karakters zijn letters, getallen en interpunctie symbolen. In een ASCII tekstbestand zal nooit enige vorm van formattering aanwezig zijn. Het Attribuut-Relation File Format werd ontwikkeld door het Machine Learning Project aan het Departement van Computerwetenschappen van de Universiteit van Waikato om samen met Weka te gebruiken. In figuur 4.3 is een voorbeeld opgenomen van een ARFF-bestand.

```
%Titel of naam van de data set
%
%Extra informatie over de data set
%wordt in commentaar bovenaan het
%arff-bestand vermeld
%
@RELATION NaamDataSet

@ATTRIBUTE NaamAttribuut1 DataTypeAttribuut1
@ATTRIBUTE NaamAttribuut2 DataTypeAttribuut2
@ATTRIBUTE NaamAttribuut3 DataTypeAttribuut3

@DATA
WaardeAttribuut1Geval1, WaardeAttribuut2Geval1, WaardeAttribuut3Geval1
WaardeAttribuut1Geval2, WaardeAttribuut2Geval2, WaardeAttribuut3Geval2
WaardeAttribuut1Geval3, WaardeAttribuut2Geval3, WaardeAttribuut3Geval3
WaardeAttribuut1Geval4, WaardeAttribuut2Geval4, WaardeAttribuut3Geval4
WaardeAttribuut1Geval5, WaardeAttribuut2Geval5, WaardeAttribuut3Geval5
```

Figuur 4.3: Voorbeeld van een ARFF-bestand

Zoals men kan afleiden uit bovenstaande figuur, lijkt een ARFF-bestand te bestaan uit drie delen. Het bestand zal meestal beginnen met een commentaargedeelte. Daarna volgt het tweede gedeelte, de Header, met allerlei informatie over de dataset. Het derde deel, het data-gedeelte, bevat de eigenlijke data van de dataset. In hetgeen wat volgt geven we een korte beschrijving van de verschillende onderdelen van een ARFF-bestand. Voor een meer uitgebreide en gedetailleerde beschrijving van dit bestandsformaat verwijzen we naar de website, `http://www.cs.waikato.ac.nz/~ml/weka/arff.html`.

Header-gedeelte

Zoals gezien kan worden in figuur 4.4 bevat dit gedeelte van het ARFF-bestand de naam van de dataset of tabel en een lijst van de attributen tesamen met het datatype waartoe ze behoren. De naam van de dataset wordt weergegeven achter *@RELATION*. Het attribuuttype en de naam van het attribuut worden op de volgende manier vermeld *@ATTRIBUTE naam datatype*, dit is ook zichtbaar in figuur 4.4. Een overzicht van de mogelijke datatypes voor de attributen is terug te vinden op de hierboven vermelde website en in de verscheidene documenten op het Internet die handelen over het Attribute-Relation File Format.

```
@RELATION NaamDataSet

@ATTRIBUTE NaamAttribuut1 DataTypeAttribuut1
@ATTRIBUTE NaamAttribuut2 DataTypeAttribuut2
@ATTRIBUTE NaamAttribuut3 DataTypeAttribuut3
```

Figuur 4.4: Voorbeeld van de Header van een ARFF-bestand

Data-gedeelte

Het data-gedeelte van een ARFF-bestand begint altijd met *@DATA*. Daarna worden de instanties uit de dataset weergegeven waarbij iedere lijn één enkele instantie is. De attribuutwaarden voor de specifieke instantie worden gescheiden door komma's en ontbrekende attribuutwaarden worden aangegeven door vraagtekens. Er moet wel rekening mee gehouden worden dat de volgorde van de attribuutwaarden overeenkomt met de volgorde van de attributen in het Header-gedeelte van het ARFF-bestand.

```
@DATA
WaardeAttribuut1Geval1, WaardeAttribuut2Geval1, WaardeAttribuut3Geval1
WaardeAttribuut1Geval2, WaardeAttribuut2Geval2, WaardeAttribuut3Geval2
WaardeAttribuut1Geval3, WaardeAttribuut2Geval3, WaardeAttribuut3Geval3
```

Figuur 4.5: Voorbeeld van data in een ARFF-bestand

Commentaar in een ARFF-bestand

Zoals duidelijk wordt uit figuur 4.6 wordt commentaar in een ARFF-bestand aangeduid met behulp van een percentteken. Als er in het begin van de lijn een %-teken staat dan zal deze lijn als commentaar beschouwd worden door Weka.

```
%Titel of naam van de data set
%
%Extra informatie over de data set
%wordt in commentaar bovenaan het
%arff-bestand vermeld
%
```

Figuur 4.6: Voorbeeld van commentaar in een ARFF-bestand

4.2 Sensitiviteitsalgoritme

In deze paragraaf gaan we bespreken hoe het sensitiviteitsalgoritme wordt geïmplementeerd. Eerst geven we de betekenis van enkele veelgebruikte notaties en daarna zal er dieper ingegaan worden op de implementatie van het algoritme.

4.2.1 Gebruikte notaties

Bij het beschrijven van de implementatie van het sensitiviteitsalgoritme, maken we gebruik van pseudocode. Dit is een soort gestructureerde taal die aangewend wordt om de algemene opbouw van een programma weer te geven. In onderstaande opsomming geven we een beschrijving ter verduidelijking van de symbolen die we gebruikt hebben.

a
: Het aantal attributen van een geval.

d
: De afstand tussen twee punten.

$d_{Minimum}$
: De kleinst bestaande afstand tussen twee punten.

k
: Een knoop of een blad in de beslissingsboom.

k_p
: De knoop of het blad op positie p in de beslissingsboom.

n
: Het aantal kinderen van een bepaalde knoop in de beslissingsboom. In deze thesis kan dit maximaal twee zijn, omdat we met binaire beslissingsbomen werken.

p
: De positie van een knoop of blad in de beslissingsboom.

x
: Een geval uit de dataset.

$y = 5$
: Hiermee bedoelen we dat y gelijk is aan de waarde 5.

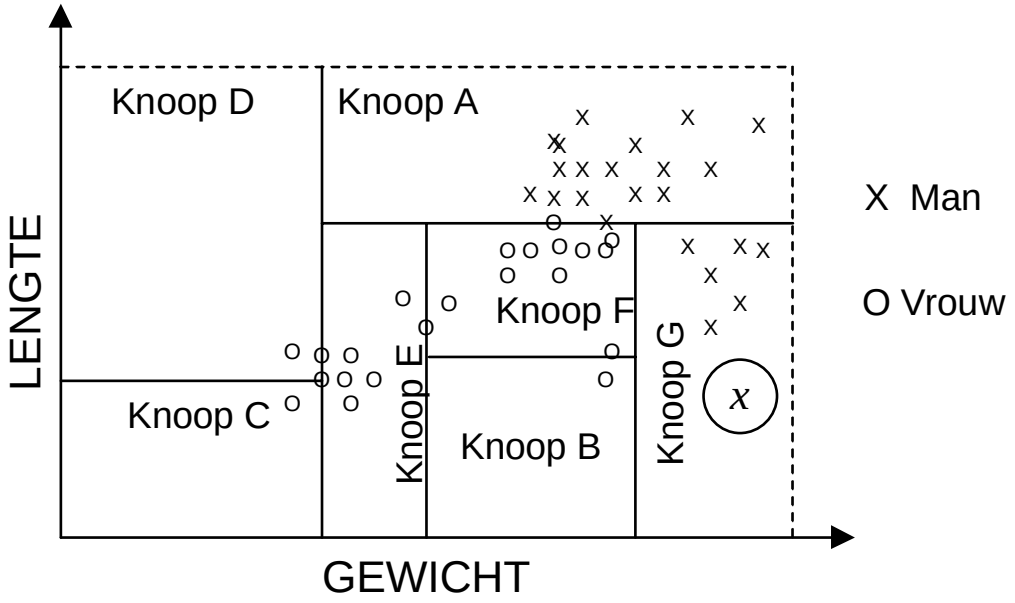
$y \leftarrow 5$
: Deze constructie geeft aan dat y de waarde 5 krijgt toegekend.

4.2.2 Sensitiviteitsalgoritme

Het doel van het sensitiviteitsalgoritme is het vinden van de kleinst bestaande afstand tussen een specifiek geval x uit de dataset en het daarbij horende beslissingsoppervlak. Figuur 4.7 op p 35 laat het geval x en het beslissingsoppervlak van een beslissingsboom gelijkaardig aan deze uit figuur 3.2 op p 18 zien.

Bij het toepassen van het sensitiviteitsalgoritme wordt voor ieder geval uit de dataset de functie $SensitivityAt(x, a)$ aangeroepen.

SensitivityAt$(\mathbf{x}, \mathbf{a})$
$d_{Minimum} \leftarrow PreorderAndGetMinimumDistance(x, a, 0)$
return $(d_{Minimum})$



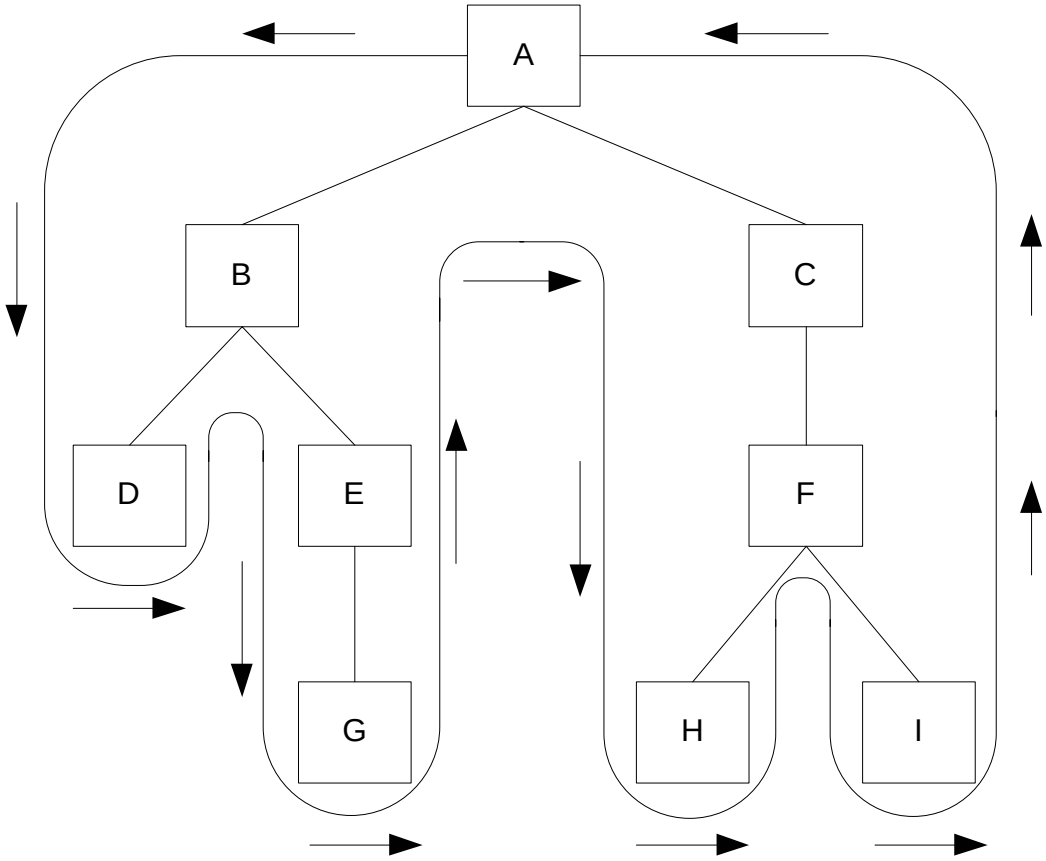
Figuur 4.7: Een vereenvoudigde voorstelling van een beslissingsoppervlak van een beslissingsboom

Om de minimale afstand te kunnen bepalen, is het nodig dat we de functie $PreorderAndGetMinimumDistance(x, a, p)$ aanroepen en hierbij de juiste waarden aan de parameters x, a en p meegeven. De parameter x zal het geval waarvoor we de minimale afstand willen bepalen bevatten. Parameter a komt overeen met het aantal attributen dat ieder geval in de dataset heeft. Aan parameter p geven we de waarde 0, wat betekent dat we de knoop bedoelen op positie 0 in de beslissingsboom. Dit wordt ook de wortelknoop of de wortel van de beslissingsboom genoemd.

4.2.3 Doorlopen en sorteren

Bij het uitvoeren van het sensitiviteitsalgoritme moeten we de beslissingsboom knoop per knoop kunnen doorlopen. Een bijkomende vereiste hierbij is dat alle knopen bezocht dienen te worden omdat we de klasse van geval x moeten vergelijken met de klasse van de bladeren van de beslissingsboom. Ook de test in iedere knoop moet opgevraagd kunnen worden met de verdere uitwerking van het sensitiviteitsalgoritme in het achterhoofd.

De Preorder sorteerwijze, geïllustreerd in figuur 4.8 op p 36, leek ons het



Figuur 4.8: Preorder

meest geschikt omdat hierbij telkens één tak volledig afgewerkt wordt tot in een blad alvorens er naar de volgende tak overgeschakeld wordt. Door deze manier van sorteren te gebruiken om de beslissingsboom te doorlopen zullen de benodigde gegevens van elke knoop of elk blad onmiddellijk gebruikt kunnen worden na de "eerste" ontmoeting (Gyssens 2001). Dit kan ook afgeleid worden uit het Preorder-algoritme 4.1. We zullen dus als volgt door de beslissingsboom bewegen: *A, B, D, E, G, C, F, H, I*. Dit is ook duidelijk af te leiden uit figuur 4.8.

Preorder(**k**) :
Bezoek knoop k;
Als knoop k kinderen heeft, doe dan voor ieder
kind c van knoop k : *Preorder*(*c*) (4.1)

We hebben een functie gecreëerd die het mogelijk maakt om op de Preorder manier door de beslissingsboom te bewegen. Daarbij kunnen de benodigde gegevens van de betreffende knoop of het specifieke blad onmiddellijk opgevraagd worden. Onderstaande functie illustreert dit.

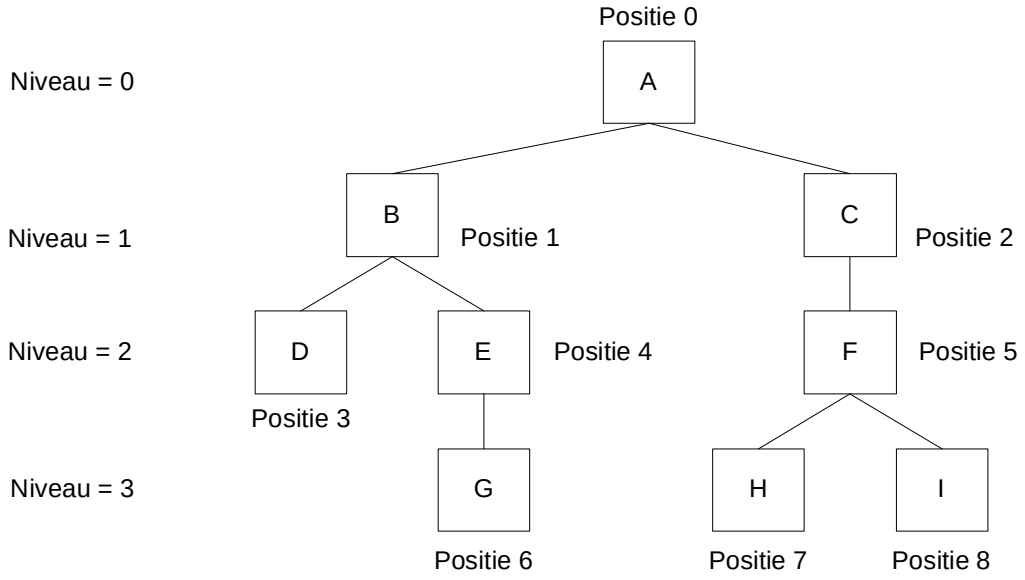
$$
\begin{array}{l}
\textbf{PreorderAndGetMinimumDistance}(\mathbf{x}, \mathbf{a}, \mathbf{p}) \\
d_{Minimum} \leftarrow +\infty \\
n \leftarrow NrChildren(k_p) \\
\textbf{if } n = 0 \\
\quad \textbf{then} \begin{cases} \textbf{if } Class(k_p) \neq Class(x) \\ \quad \textbf{then} \begin{cases} d \leftarrow ProjectAndCalculateDistance(x, a, p) \\ \textbf{if } d < d_{Minimum} \\ \quad \textbf{then } d_{Minimum} \leftarrow d \end{cases} \end{cases} \\
\textbf{for } i \leftarrow 1 \textbf{ to } n \\
\quad \textbf{do} \begin{cases} p_i \leftarrow GetChildPosition(k_p, i) \\ d \leftarrow PreorderAndGetMinimumDistance(x, a, p_i) \\ \textbf{if } d < d_{Minimum} \\ \quad \textbf{then } d_{Minimum} \leftarrow d \end{cases} \\
\textbf{return } (d_{Minimum})
\end{array}
$$

De eerste keer dat we voor een specifiek geval x met a attributen in deze functie terechtkomen, zal de minimale afstand $d_{Minimum}$ gelijkgesteld worden aan $+\infty$. De eerste afstand die berekend wordt voor het geval x zal hierdoor dus altijd beschouwd worden als de, voorlopige, minimale afstand. Na deze berekening wordt het aantal kinderen van de knoop k op de doorgegeven positie p bepaald.

Als knoop k_p geen kinderen heeft, dan weten we dat deze knoop een blad is. De klasse $Class(x)$ van het geval x moet dan vergeleken worden met de klasse $Class(k_p)$ van het betreffende blad. Wanneer beide klassen niet hetzelfde zijn, wat in formulevorm neerkomt op $Class(x) \neq Class(k_p)$, dan moeten we $ProjectAndCalculateDistance(x, a, p)$ aanroepen. Deze functie wordt aangeroepen zodat we de afstand d kunnen bepalen. De parameters die meegegeven worden bij aanroepen van deze functie zijn dezelfde als deze waarmee de huidige functie, $PreorderAndGetMinimumDistance(x, a, p)$ werd aangeroepen omdat de knoop k_p geen kinderen heeft. Als de berekende afstand d kleiner is dan de reeds bestaande kleinste afstand dan zal de mini-

mum afstand $d_{Minimum}$ gelijkgesteld worden aan de berekende afstand d.

Stel nu dat knoop k_p wel kinderen heeft, dan komen we in de for-lus van de functie $PreorderAndGetMinimumDistance(x, a, p)$ op p 37 terecht. In deze for-lus wordt voor elk i-de kind van knoop k_p de positie p_i bepaald. De afstand wordt berekend door $PreorderAndGetMinimumDistance(x, a, p_i)$ aan te roepen. Er zal dan opnieuw gecontroleerd worden als de berekende afstand d kleiner is dan de huidige minimale afstand. Indien dit het geval is, wordt $d_{Minimum}$ gelijkgesteld aan d. Zoals reeds meerdere malen aangehaald werd, werken we in deze thesis met binaire beslissingsbomen. Een beslissingsboom van dit soort zal maximaal twee kinderen per knoop kunnen hebben. Indien een knoop twee kinderen heeft dan zullen we tweemaal de bovenvermelde for-lus ingaan voor de betreffende knoop, zowel voor het eerste als voor het tweede kind. Figuur 4.9 werd opgenomen ter verduidelijking van de termen *positie p van een knoop*, *eerste kind* en *tweede kind*. De knoop B op positie 1 heeft als eerste kind blad D op positie 3 en als tweede kind knoop E op positie 5 in de beslissingsboom.



Figuur 4.9: Illustratie van niveau en positie van een blad in een beslissingsboom

Het resultaat van deze functie $PreorderAndGetMinimumDistance(x, a, p)$ is de minimale afstand tussen geval x en het bijhorende beslissingsoppervlak.

4.2.4 Allerlei berekeningen

De functie die in deze subparagraaf besproken wordt, dient vooral als intermediaire functie en zal het geval x projecteren op het bijhorende beslissingsoppervlak. Indien nodig zal zowel de projectie x_p van x als x zelf geconverteerd worden naar een andere metriek. De functie om de afstand tussen de al dan niet geconverteerde x_c en de al dan niet geconverteerde projectie x_{pc} te berekenen wordt eveneens aangeroepen in deze intermediaire functie $ProjectAndCalculateDistance(x, a, p)$.

ProjectAndCalculateDistance$(\mathbf{x}, \mathbf{a}, \mathbf{p})$
$x_p \leftarrow ProjectionOntoLeaf(x, p)$
$x_c \leftarrow Metric(x, a)$
$x_{pc} \leftarrow Metric(x_p, a)$
$d = EuclidianDistance(x_{pc}, x_c, a)$
return (d)

De functie $ProjectionOntoLeaf(x, p)$ wordt aangeroepen om het geval x op het blad met positie p te projecteren. Hieropvolgend zal dan zowel het geval x als de projectie x_p van x omgezet worden naar een andere metriek met behulp van de functie $Metric(x, a)$ respectievelijk $Metric(x, a)$. Als laatste wordt de afstand d tussen het geconverteerde geval x_c en de geconverteerde projectie x_{pc} berekend door de functie $EuclidianDistance(x_{pc}, x_c, a)$. Het resultaat van $ProjectAndCalculateDistance(x, a, p)$ is de afstand tussen het geval x en het blad op positie p in de beslissingsboom.

4.2.5 Projecteren

Zoals al duidelijk werd uit het voorgaande moet het geval x geprojecteerd worden op het bijhorende beslissingsoppervlak. Dit hebben we verwezenlijkt met behulp van onderstaande functie $ProjectionOntoLeaf(x, p)$.

ProjectionOntoLeaf$(\mathbf{x}, \mathbf{p})$
$x_p \leftarrow x$
$k \leftarrow k_p$**while** $HasParent(k)$
do $\begin{cases} k \leftarrow GetParent(k) \\ \textbf{if not } IsTestSatisfied(k, x) \\ \quad \textbf{then } x_p \leftarrow UpdateTestedAttribute(k, x) \end{cases}$
return (x_p)

In een eerste stap stellen we de projectie x_p van x gelijk aan x. Daarna stellen we de knoop k gelijk aan zijn ouder en dit blijven we doen tot we uiteindelijk in de wortelknoop van de beslissingsboom terechtkomen. Hierbij gaan we eveneens kijken of er voldaan wordt aan de test die zich bevindt in de ouderknoop van de huidige knoop op positie p. Indien dit niet het geval is, zal de waarde die x heeft voor het attribuut dat in de test van de ouder van de knoop k_p voorkomt, vervangen worden door de waarde waarop het attribuut getest wordt in deze ouderknoop.

Deze functie zal dus voor iedere test die x tegenkomt in de beslissingsboom gaan kijken of x voldoet aan de beperking opgelegd door de betreffende test. Als er niet voldaan wordt aan de beperking dan zal de waarde van x voor het attribuut waarop getest wordt gelijkgesteld worden aan de waarde van de test. Indien er wel voldaan wordt aan de beperking gesteld in de test dan zal de huidige waarde van x voor het attribuut behouden worden.We hebben hierbij de veronderstelling gemaakt dat iedere test in de beslissingsboom gebaseerd zal zijn op één enkel attribuut.

Het resultaat van deze $ProjectionOntoLeaf(x, p)$ is de projectie x_p van geval x op het blad dat zich op positie p in de beslissingsboom bevindt.

4.2.6 Toepassing metriek

Net zoals Isabelle Alvarez in *Sensitivity Analysis of the Result in Binary Decision Trees: Giving More Information to the End-user* maken we ook hier gebruik van twee verschillende metrieken om het sensitiviteitsalgoritme toe te passen, namelijk de MinMax-metriek en de Standaard-metriek.

Aan de hand van de if-testen in de functie $Metric(x, a)$ op p 41 geven we aan welke metriek we willen toepassen. Niet alleen de instantie of het geval

x zelf maar ook het aantal attributen a wordt doorgegeven omdat de waarden van alle attributen van het geval x geconverteerd moeten worden. Zoals men kan afleiden uit de eerste if-test in de functie $Metric(x, a)$ moet er niet noodzakelijk overgeschakeld worden naar een andere metriek. Er kan ook gewerkt worden met de huidige attribuutwaarden.

Metric$(\mathbf{x}, \mathbf{a})$
if $noMetric$
  then $x_c \leftarrow x$
if $standardMetric$
  then $x_c \leftarrow StandardMetric(x, a)$
if $minmaxMetric$
  then $x_c \leftarrow MinMaxMetric(x, a)$
return (x_c)

Zowel de Standaard-metriek als de MinMax-metriek zijn gedefinieerd met informatie die men gemakkelijk uit de training dataset kan afleiden. De Standaard-metriek werkt met een schatting van het gemiddelde E_i en een schatting van de standaardafwijking s_i. De MinMax-metriek wordt gebaseerd op een schatting van de meest uit elkaar gelegegen waarden van ieder attribuut i met andere woorden het minimum en maximum van het betreffende attribuut (Alvarez 2004).

Het converteren van de attribuutwaarden naar één van de metrieken doen we met behulp van onderstaande functies.

StandardMetric$(\mathbf{x}, \mathbf{a})$
$x_c \leftarrow x$
for $i \leftarrow 0$ **to** a
  do $x_{ci} \leftarrow \frac{x_i - E_i}{s_i}$
return (x_c)

MinMaxMetric$(\mathbf{x}, \mathbf{a})$
$x_c \leftarrow x$
for $i \leftarrow 0$ **to** a
  do $x_{ci} \leftarrow \frac{x_i - Min_i}{Max_i - Min_i}$
return (x_c)

Het resultaat van deze functies is de geconverteerde x, x_c, waarbij het converteren gebaseerd is op de Standaard-metriek of op de MinMax-metriek. We passen deze metrieken toe om de attribuutwaarden als het ware te standaardiseren. Bij de standaardmetriek kan dit letterlijk opgevat worden zoals duidelijk werd uit bovenstaande functie $StandardMetric(x, a)$. De waarde die x heeft voor het attribuut i wordt hier verminderd met de gemiddelde waarde voor dit attribuut. Dit verschil wordt dan gedeeld door de standaardafwijking van het betreffend attribuut. Als we de resultaten van de functie $MinMaxMetric(x, a)$ in onderstaand voorbeeld bekijken, dan zien we dat voor elk attribuut de waarden herleid zullen worden tot getallen tussen 0 en 1.

Voorbeeld MinMax-metriek:

Waarden attribuut 1: 0, 1, 2, 3, 5
Minimumwaarde attribuut 1: 0
Maximumwaarde attribuut 1: 5
Geconverteerde waarden attribuut 1:

$$\frac{(0-0)}{(5-0)} = 0;\ \frac{(1-0)}{(5-0)} = 0,2;\ \frac{(2-0)}{(5-0)} = 0,4;\ \frac{(3-0)}{(5-0)} = 0,6;\ \frac{(5-0)}{(5-0)} = 1$$

Waarden attribuut 2: 5000, 10000, 20000
Minimumwaarde attribuut 2: 5000
Maximumwaarde attribuut 2: 20000
Geconverteerde waarden attribuut 2:

$$\frac{(5000-5000)}{(20000-5000)} = 0;\ \frac{(10000-5000)}{(20000-5000)} = 0,33;\ \frac{(20000-5000)}{(20000-5000)} = 1$$

4.2.7 Afstandsberekening

De Euclidische afstand tussen twee punten x en y wordt met behulp van de formule op p 43 berekend. In het kader van het sensitiviteitsalgoritme zal dus de afstand tussen het al dan niet geconverteerde geval x en de al dan niet geconverteerde projectie x_p berekend worden. Dit hebben we gerealiseerd door de formule op p 43 te integreren in de functie $EuclidianDistance(x_1, x_2, a)$ op p 43 (Weisstein 2005c).

$$d = \|x - y\| = \sqrt{\sum_{i=1}^{a} (x_i - y_i)^2}$$

EuclidianDistance($\mathbf{x_1}, \mathbf{x_2}, \mathbf{a}$)
$d \leftarrow 0$
for $i \leftarrow 0$ **to** a
  do $d \leftarrow d + (x_{1_i} - x_{2_i})^2$
return $(\sqrt{d})$

Hoofdstuk 5

Gevalstudie

In dit hoofdstuk gaan we het geïmplementeerde sensitiviteitsalgoritme toepassen in een praktijksituatie, meer specifiek in het kader van een internationale data mining competitie. Eerst zullen we kort aanhalen wat de opzet was van deze competitie en daarna bespreken we hoe we de beslissingsboom, die nodig was voor het kunnen toepassen van het sensitiviteitsalgoritme, ontwikkeld hebben met behulp van Weka. Als laatste gaan we dan bekijken op welke manier de sensitiviteit gecombineerd kan worden met de waarschijnlijkheidsschattingen van de beslissingsboom om de resultaten te rangschikken.

5.1 Competitie

In het kader van deze thesis werd er deelgenomen aan een data mining competitie die gesponsord werd door *Deutsche Sparkassen- und Giroverband (DSGV)*. Deze wedstrijd werd georganiseerd in de aanloop van de data mining conferentie *The 29th Annual Conference of the German Classification Society (GfKl 2005): From Data and Information Analysis to Knowledge Engineering*. Het doel van de wedstrijd was het voorspellen of bedrijven een liquiditeitscrisis zouden ondergaan.

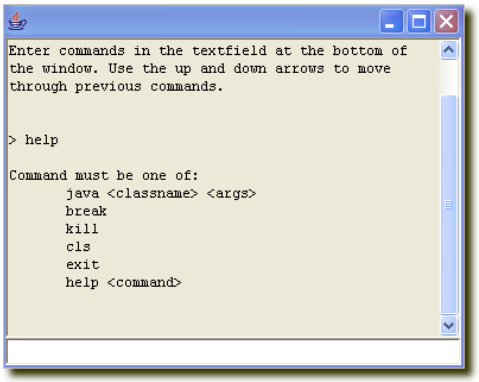
Deze voorspelling moest gebaseerd zijn op een model dat ontwikkeld werd met behulp van de training en test dataset van de competitie. De test dataset bevatte 26 variabelen met betekenisloze namen, de variabelen werden namelijk geïdentificeerd met de namen $var01$ tot en met $var26$. In het trainingsbestand was er dan ook nog een kolom aanwezig met het bedrijfsnummer, ID en een andere kolom waarin stond of het betreffende bedrijf op dat moment af te rekenen had met een liquiditeitscrisis of niet. Deze laatste kolom, *Ereignisflag* genaamd, was niet aanwezig in het testbestand omdat dit net de variabele was die aan de hand van het ontwikkelde model voorspeld moest worden. De 20000 gevallen in het bestand van de training dataset moesten gebruikt worden om een model te ontwikkelen. Daarna moest het ontwikkelde model toegepast worden op de 10000 gevallen in het bestand van de test dataset.

Een meer gedetailleerde beschrijving van de competitie en de hieraan verbonden taken is terug te vinden in bijlage C. Het rapport dat meegeleverd moest worden met de oplossing is opgenomen in bijlage D.

Aangezien de datasets van deze wedstrijd enkel numerieke attributen bevatten, konden zij dus gebruikt worden in het praktijkgedeelte van deze thesis. Het praktijkgedeelte bestaat namelijk uit het implementeren van het sensitiviteitsalgoritme en het kijken of de sensitiviteit gebruikt kan worden om de resultaten van binaire beslissingsbomen te beoordelen of te rangschikken. We hebben reeds eerder aangehaald dat het sensitiviteitsalgoritme enkel kan toegepast worden op datasets waarvan de attribuutwaarden numeriek zijn. De datasets van deze competitie kunnen dus gebruikt worden om het algoritme te testen.

5.2 Beslissingsboom

De beslissingsboom hebben we ontwikkeld met behulp van *Simple CLI* van Weka. In deze paragraaf halen we aan welke parameters gebruikt werden om deze beslissingsboom te ontwikkelen. Zoals duidelijk wordt in figuur 5.1 worden de commando's onderaan het scherm van de *Simple CLI* ingegeven. De beschrijving van de gebruikte parameters is gebaseerd op de informatie die beschikbaar is op de website van Weka, `http://www.cs.waikato.ac.nz/ml/weka/`.



Figuur 5.1: Simple CLI

5.2.1 Ontwikkeling van de beslissingsboom

Uit figuur 5.2 wordt duidelijk welke parameters gebruikt werden in het commando voor de *Simple CLI* om het gewenste model te kunnen ontwikkelen aan de hand van de training dataset *NaamFileTrainingData.arff*.

```
java weka.classifiers.trees.J48 -B -C 0.25 -M 2 -t NaamFileTrainingData.arff -d NaamFileModel
```

Figuur 5.2: Commando voor het ontwikkelen van een specifiek model

Verklaring van de gebruikte parameters

- *weka.classifiers.trees.J48*: Hiermee wordt het beslissingsboomalgoritme $J48$ aangeroepen. Dit algoritme is gebaseerd op het C4.5-beslissingsboomalgoritme dat beschreven wordt in bijlage A.
- *-B*: Deze parameter wordt gebruikt om aan te geven dat er een binaire beslissingsboom moet ontwikkeld worden.
- *-C 0.25*: Hiermee geven we de confidence threshold aan voor het snoeien van de beslissingsboom. De waarde 0.25 is de standaardwaarde die Weka gebruikt voor deze parameter.
- *-M 2*: Het minimum aantal instanties in een blad stellen we met deze parameter en de gegeven waarde gelijk aan twee, ook dit is de standaardwaarde van Weka.
- *-t NaamFileTrainingData.arff*: Deze parameter specificeert het ARFF-trainingbestand dat gebruikt moet worden om de beslissingsboom te ontwikkelen.
- *-d NaamFileModel*: Hiermee wordt duidelijk gemaakt onder welke naam en op welke plaats op de computer het ontwikkelde model, in dit geval een beslissingsboom, moet opgeslaan worden.

Doordat we in het commando van figuur 5.2 op p 46 geen testbestand opgegeven hebben, zichtbaar aan het feit dat er geen $-T$ aanwezig is, zal Weka automatisch ten-fold cross-validatie toepassen. Voor meer uitleg over cross-validatie verwijzen we naar bijlage A of naar de meer gespecialiseerde literatuur (Witten en Frank 2000, Quinlan 1993).

De beslissingsboom die ontstaan is door het commando uit figuur 5.2 op p 46 in de *Simple CLI* in te geven is terug te vinden in bijlage E.

5.2.2 Toepassen van de ontwikkelde beslissingsboom

Om de ontwikkelde beslissingsboom te kunnen toepassen op de test dataset *NaamFileTestData.arff*, moest het commando uit figuur 5.3 op p 48 ingegeven worden in de *Simple CLI* van Weka.

```
java weka.classifiers.trees.J48 -p 27 -l NaamFileModel -T NaamFileTestData.arff
```

Figuur 5.3: Commando voor het toepassen van een specifiek model

Verklaring van de gebruikte parameters

- *weka.classifiers.trees.J48*: Hiermee maken we duidelijk dat we de beslissingsboom willen aanmaken met behulp van het *J*48-algoritme.
- *-p 27*: Met behulp van deze parameter geven we door aan Weka dat we de zevenentwintigste attribuut uit het testbestand *NaamFileTestData.arff* willen laten voorspellen aan de hand van het ontwikkelde model.
- *-l NaamFileModel*: Deze parameter wordt gebruikt om aan te geven welk, eerder opgeslagen, model Weka moet toepassen op de test dataset.
- *NaamFileTestData.arff*: Hiermee geven we het ARFF-testbestand aan waarop de beslissingsboom moet toegepast worden.

De output die Weka genereert wanneer dit commando wordt ingegeven, zijn de gevallen in de test dataset tesamen met de voorspelde klasse en de kans dat de gedane voorspellingen juist zullen zijn. Het zijn deze gegevens, de voorspelde klasse en de waarschijnlijkheidsschattingen, die we gaan gebruiken om het sensitiviteitsalgoritme toe te passen.

5.3 Sortering van de resultaten

In deze paragraaf gaan we het sensitiviteitsalgoritme toepassen op de verkregen resultaten. Hierbij wordt er onderzocht op welke manier dit best kan gebeuren. In hetgeen wat volgt gaan we enkele mogelijke rangschikkingen van de resultaten bekijken tesamen met hun impact op de hoeveelheid juist geclassificeerde gevallen. Deze rangschikkingen zullen gebaseerd zijn op de waarschijnlijkheidsschattingen of op de sensitiviteit of op een combinatie van beide. Om de verschillende rangschikkingen te kunnen vergelijken hebben we besloten om, net zoals in de wedstrijd, slechts rekening te houden met de 2000 'beste' gevallen uit de dataset. Met 'beste' bedoelen we diegenen die volgens de gebruikte rangschikkingsmethode het meest waarschijnlijk zijn.

De beoordeling van de voorgestelde rangschikkingen gaat gebeuren door de 2000 meest waarschijnlijke gevallen van iedere toegepaste rangschikkingsmethode te vergelijken met het resultatenbestand dat ons werd opgestuurd door de verantwoordelijken van de wedstrijd.

Mogelijke manieren voor het selecteren van gevallen die we beschouwen in het kader van deze thesis:

- *Selectie volgens sortering op waarschijnlijkheidsschatting*
- *Selectie volgens sortering op sensitiviteit*
- *Selectie volgens sortering op het product van de waarschijnlijkheidsschatting en de sensitiviteit*
- *Selectie volgens sortering op gecorrigeerde waarschijnlijkheidsschatting*

5.3.1 Sortering op waarschijnlijkheidsschatting

Meestal zullen de waarschijnlijkheidsschattingen gegenereerd door een model gebruikt worden om de geclassificeerde gevallen te sorteren van meest waarschijnlijk naar minst waarschijnlijk. Zoals reeds verteld werd, genereert Weka bij het toepassen van het ontwikkelde model op de dataset per geclassificeerd geval een waarschijnlijkheidsschatting. De waarschijnlijkheidsschatting is de kans dat de voorspelling die Weka of het model gedaan heeft juist is, dus dat de voorspelde klasse ook de werkelijke klasse is van het specifieke geval.

Bij het ontwikkelen van de beslissingsboom voor de data mining competitie hebben we gebruik gemaakt van 10-fold cross-validation om overfitting zoveel mogelijk tegen te gaan. Maar om nog meer betrouwbare resultaten te bekomen werd dit cross-validation proces tien keer toegepast op de dataset (Witten en Frank 2000).

In tabel 5.1 op p 50 worden de resultaten van de selectie volgens deze waarschijnlijkheidsschattingen, 10-fold cross-validation en tien keer 10-fold cross-validation, getoond. De selectie is gebeurd op basis van dalende waarschijnlijkheidsschatting waarbij de 2000 gevallen met de hoogste waarschijnlijkheidsschatting werden geselecteerd.

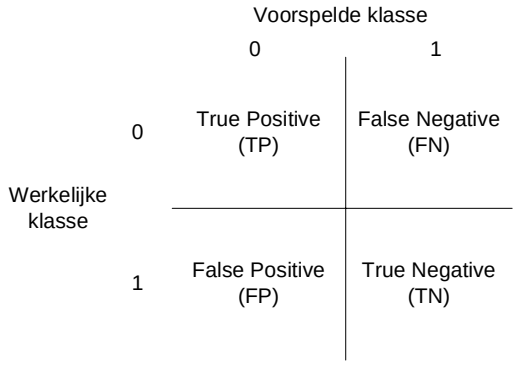
Zoals eerder vermeld, moet de beslissingsboom voor de competitie de gevallen van de test dataset classificeren als behorende tot klasse 0 of tot klasse 1. Klasse 0 komt overeen met bedrijven die niet in een liquiditeitscrisis zitten

en tot klasse 1 behoren de bedrijven die wel in een liquiditeitscrisis zitten.

Tabel 5.1: Resultaten van de selectie van 2000 gevallen volgens sortering op waarschijnlijkheidsschatting (100 % = 2000, cv = cross-validation)

Sorteer-	Aantal juist voorspeld			Aantal fout voorspeld		
methode	Klasse 0	Klasse 1	Totaal	Klasse 0	Klasse 1	Totaal
10-fold cv	1904	18	1922	54	24	78
10x10-fold cv Gemiddelde	1556	121	1678	140	182	322
Getransfor-meerde	940	471	1411	307	282	589
Sorteer-	Aantal juist voorspeld (%)			Aantal fout voorspeld (%)		
methode	Klasse 0	Klasse 1	Totaal	Klasse 0	Klasse 1	Totaal
10-fold cv	95,20	0,90	96,10	2,70	1,20	3,90
10x10-fold cv Gemiddelde	77,83	6,04	83,86	7,00	9,14	16,14
Getransfor-meerde	47,00	23,55	70,55	15,35	14,10	29,45

Een mogelijke manier om een beslissingsboom te evalueren is het gebruiken van een confusion matrix. Een confusion matrix toont het aantal gevallen dat terecht als klasse 0 en klasse 1 geclassificeerd is en het aantal gevallen dat onterecht als klasse 0 en klasse 1 geclassificeerd is. Doordat we maar twee mogelijke klassen hebben in onze gevalstudie zal de confusion matrix voorgesteld kunnen worden door een 2x2-matrix zoals in figuur 5.4.



Figuur 5.4: Confusion matrix

De classificaties die terechtkomen in het True Positive - en True Negative-kwadrant zijn correct gebeurde classificaties, de gevallen die geclassificeerd werden als klasse 0 en klasse 1 behoren ook werkelijk tot deze klassen. Een geval dat in het False Positive-kwadrant ligt, is een geval dat werkelijk klasse 1 heeft maar door de beslissingsboom als klasse 0 geclassificeerd werd. Indien een geval geclassificeerd werd als klasse 1 maar werkelijk tot klasse 0 behoort, dan wordt het in het False Negative-kwadrant geplaatst. Een goed classificerende beslissingsboom heeft hoge aantallen op de hoofddiagonaal en lage aantallen, liefst gelijk aan nul, op de andere diagonaal. In het kader van de data mining competitie waar we aan deelnamen, waren het de gevallen uit het True Negative-kwadrant waaraan veel belang werd gehecht. Het absoluut aantal gevallen in dit kwadrant is terug te vinden in tabel 5.1 op p 50 in de kolom *Aantal juist voorspeld, klasse 1* en het procentuele aantal vindt men terug in de confusion matrices in figuur 5.5.

	Klasse 0	Klasse 1
Klasse 0	86,375%	2,405%
Klasse 1	6,345%	4,875%

(a) Confusion matrix voor de ontwikkelde beslissingsboom van de gevalstudie (20000 gevallen)

	Klasse 0	Klasse 1
Klasse 0	86,07%	2,82%
Klasse 1	6,40%	4,71%

(b) Confusion matrix na het toepassen van de beslissingsboom op de test dataset van de gevalstudie (10000 gevallen)

Figuur 5.5: Confusion matrices gevalstudie sortering op waarschijnlijkheid

Zoals eerder vermeld, willen we vooral de gevallen die tot klasse 1 behoren juist voorspellen. Dit zijn de gevallen die tot het True Negative-kwadrant en het False Positive-kwadrant behoren. We veronderstellen dat de gevallen die tot het False Positive-kwadrant behoren diegenen zijn die voorspeld werden als klasse 0 maar die een lage waarschijnlijkheidsschatting hebben gekregen. Om deze gevallen te selecteren, kunnen we de waarschijnlijkheidsschattingen van de gevallen die tot klasse 0 behoren proberen om te zetten naar de kans dat ze toch tot klasse 1 zouden behoren. Dit kan gebeuren met onderstaande

formule:

$W_1 =$ *oorspronkelijke waarschijnlijkheidsschatting of kans dat het geval werkelijk behoort tot klasse* 1

$W_0 =$ *oorspronkelijke waarschijnlijkheidsschatting of kans dat het geval werkelijk behoort tot klasse* 0

$W_0^* =$ *getransformeerde waarschijnlijkheidsschatting*

$$W_1 = W_1 W_0^* = 1 - W_0$$

We kunnen gebruik maken van deze formule omdat de geclassificeerde gevallen ofwel tot klasse 0 ofwel tot klasse 1 behoren. Hierdoor mogen we veronderstellen dat als het geval niet tot klasse 0 behoort, het enkel nog tot klasse 1 kan behoren. De waarschijnlijkheidsschattingen van de gevallen die als klasse 1 geclassificeerd werden, gaan we niet transformeren. De resultaten die bekomen werden wanneer we de gevallen volgens deze dalende, getransformeerde en niet-getransformeerde, waarschijnlijkheidsschattingen sorteerden zijn terug te vinden in tabel 5.1 op p 50.

Voor onze gevalstudie kunnen we op basis van de confusion matrices in figuur 5.5 op p 51 en tabel 5.1 op p 50 besluiten dat de ontwikkelde beslissingsboom zeer goed presteert. We kunnen eveneens afleiden dat onze boom niet aan overfitting doet want bij het classificeren van de ongeziene data worden relatief gezien ongeveer evenveel gevallen juist en/of fout geclassificeerd, zie deel b van figuur 5.5 op p 51. Het valt echter wel op dat de foute classificaties vooral bij het toekennen van klasse 1 plaatsvinden. Dit kan te wijten zijn aan het feit dat er slechts weinig gevallen van deze klasse, 2244 gevallen ten opzichte van 17756 gevallen van klasse 0, in de training set aanwezig waren. Hierdoor kan de beslissingsboom niet zo'n goede voorspellingen maken omtrent de voorwaarden waaraan een geval moet voldoen om tot deze klasse 1 te kunnen behoren.

5.3.2 Sortering op sensitiviteit

Alvorens de resultaten van deze manier van rangschikken te tonen in tabel 5.2 op p 53 en in de confusion matrices in figuur 5.6 op p 54, herhalen we even wat sensitiviteit is en hoe sensitiviteit geïnterpreteerd moet worden. De sensitiviteit van een geval uit de dataset komt overeen met de kleinste afstand

tussen het geval en het bijhorende beslissingsoppervlak. Dit oppervlak wordt gecreëerd door de ontwikkelde beslissingsboom. We veronderstellen dat hoe groter deze afstand is, hoe kleiner de kans dat er een verkeerde voorspelling of classificatie is gebeurd. Voor het berekenen van de sensitiviteit hebben we gebruik gemaakt van de beslissingsoppervlakken van de beslissingsboom die gebruikt is voor de data mining competitie. De afstand werd dan berekend tussen de de geclassificeerde gevallen en de bijhorende oppervlakken.

Zoals reeds vermeld maken we gebruik van twee metrieken om de sensitiviteit te berekenen, de Standaard metriek en de Min-max metriek. De sensitiviteit zal bij beide metrieken op dezelfde manier berekend worden maar dezelfde attributen zullen voor dezelfde gevallen meestal andere waarden hebben door het toepassen van deze metrieken. De Min-max metriek zorgt er voor dat de attribuutwaarden tussen 0 en 1 liggen terwijl de Standaard metriek de waarden van de attributen voor de verschillende gevallen zal standaardiseren.

Tabel 5.2: Resultaten van de selectie van 2000 gevallen volgens sortering op sensitiviteit (100 % = 2000)

Toegepaste metriek	Aantal juist voorspeld			Aantal fout voorspeld		
	Klasse 0	Klasse 1	Totaal	Klasse 0	Klasse 1	Totaal
Standaard metriek	1366	386	1752	71	177	248
Min-max metriek	1425	383	1808	18	174	192
Toegepaste	Aantal juist voorspeld (%)			Aantal fout voorspeld (%)		
metriek	Klasse 0	Klasse 1	Totaal	Klasse 0	Klasse 1	Totaal
Standaard metriek	68,30	19,30	87,60	3,55	8,85	12,40
Min-max metriek	71,25	19,15	90,40	0,90	8,70	9,60

Indien we de 2000 gevallen met de hoogste sensitiviteit, berekend met behulp van de min-max metriek, selecteren dan zien we in tabel 5.2 dat er van deze geselecteerde gevallen zo'n 90 % juist geclassificeerd. Maar als we gebruik maken van de standaard metriek dan zien we dat er minder juist geclassificeerde gevallen worden geselecteerd, namelijk 87 %. De meeste foute classificaties zijn, zoals zichtbaar wordt in de confusion matrices in figuur 5.6 op p 54, nog steeds terug te vinden onder klasse 1 maar in mindere mate dan bij de sortering volgens waarschijnlijkheidsschatting.

	Klasse 0	Klasse 1
Klasse 0	68,30%	8,85%
Klasse 1	3,55%	19,30%

(a) Confusion matrix voor de resultaten van de selectie van 2000 gevallen volgens sortering op sensitiviteit (sensitiviteit berekend volgens de Standaard metriek)

	Klasse 0	Klasse 1
Klasse 0	71,25%	8,70%
Klasse 1	0,90%	19,15%

(b) Confusion matrix voor de resultaten van de selectie van 2000 gevallen volgens sortering op sensitiviteit (sensitiviteit berekend volgens de Min-max metriek)

Figuur 5.6: Confusion matrices gevalstudie sortering op sensitiviteit

5.3.3 Sortering op het product van waarschijnlijkheidsschatting en sensitiviteit

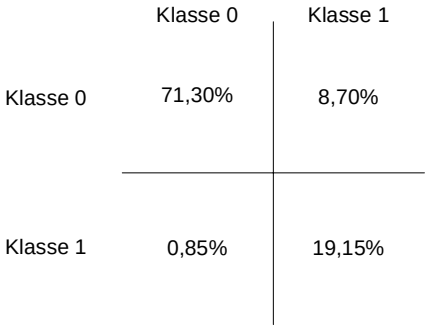
Een eerste sortering die gebaseerd is op zowel de waarschijnlijkheidsschatting als de sensitiviteit van een specifiek geval uit de dataset, is het product van deze beide factoren. De resultaten van deze rangschikkingsmethode zijn zichtbaar in tabel 5.3 op p 55 en in de confusion matrices in figuur 5.7 op p 55. De 2000 gevallen met het hoogste product werden geselecteerd.

We kunnen afleiden uit tabel 5.3 op p 55 dat van de 2000 geselecteerde gevallen zo'n 90 % juist geclassificeerd is. Deze vaststelling geldt zowel voor het sorteren volgens de sensitiviteit berekend met behulp van de standaard metriek als deze berekend volgens de min-max metriek. Opnieuw zijn de meeste foute classificaties terug te vinden onder klasse 1 maar weer in mindere mate dan bij de sortering volgens waarschijnlijkheidsschatting.

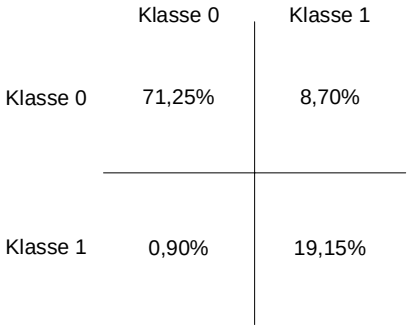
Hierbij vermelden we wel dat dit geen ideale sorteermethode is omdat, door het product te maken, de gevallen uit de dataset waarvan de sensitiviteit of de waarschijnlijkheidsschatting gelijk is aan nul meestal niet in aanmerking zullen komen voor selectie. Het product van de sensitiviteit en de waarschijnlijkheidsschatting zou dan gelijk zijn aan nul, ook al zou de waarschijnlijkheidsschatting of de sensitiviteit een hoge waarde hebben.

Tabel 5.3: Resultaten van de selectie van 2000 gevallen volgens sortering op het product van waarschijnlijkheidsschattigen en de sensitiviteit berekend (100 % = 2000)

Toegepaste	Aantal juist voorspeld			Aantal fout voorspeld		
metriek	Klasse 0	Klasse 1	Totaal	Klasse 0	Klasse 1	Totaal
Standaard metriek	1426	383	1809	17	174	191
Min-max metriek	1425	383	1808	18	174	192
Toegepaste	Aantal juist voorspeld (%)			Aantal fout voorspeld (%)		
metriek	Klasse 0	Klasse 1	Totaal	Klasse 0	Klasse 1	Totaal
Standaard metriek	71,30	19,15	90,45	0,85	8,70	9,55
Min-max metriek	71,25	19,15	90,40	0,90	8,70	9,60



(a) Confusion matrix voor de resultaten van de selectie van 2000 gevallen volgens sortering op product van sensitiviteit en waarschijnlijkheid (sensitiviteit berekend volgens de Standaard metriek)



(b) Confusion matrix voor de resultaten van de selectie van 2000 gevallen volgens sortering op product van sensitiviteit en waarschijnlijkheid (sensitiviteit berekend volgens de Min-max metriek)

Figuur 5.7: Confusion matrices gevalstudie sortering op product sensitiviteit en waarschijnlijkheid

5.3.4 Sortering op gecorrigeerde waarschijnlijkheid

In hetgeen wat volgt, trachten we een functie uit te werken waarmee de waarschijnlijkheidsschattingen op een beredeneerde manier aangepast kunnen worden. Aangezien de resulterende sorteermethode zowel rekening moet houden met de waarschijnlijkheidsschatting als met de sensitiviteit van het betreffende geval, moeten beide factoren in de functie voorkomen.

De aangepaste of gecorrigeerde waarschijnlijkheidsschatting wordt voorgesteld door het symbool W^* en de oorspronkelijke waarschijnlijkheidsschatting met het symbool W. Met behulp van de symbolen s en s_0 stellen we de sensitiviteit van het specifieke geval en de grenswaarde van de sensitiviteit voor. De symbolen c_1 en c_2 stellen beide de wijziging voor waarmee de oorspronkelijke waarschijnlijkheidsschatting W, indien nodig, zal gecorrigeerd worden.

$$W^* = \begin{cases} W + c_1 \quad als\ s > s_0 \\ W - c_2 \quad als\ s \leq s_0 \end{cases} \tag{5.1}$$

Zoals duidelijk werd in vergelijking 5.1, hebben we gekozen voor een stuksgewijze functie. Na het interpreteren van de definitie van sensitiviteit en het oorspronkelijke doel van het sensitiviteitsalgoritme volgens Isabelle Alvarez (Alvarez 2004), veronderstellen we dat bij een kleine sensitiviteit of afstand tot het beslissingsoppervlak er meer kans is op een verkeerde voorspelling of classificatie. Bij een grote afstand zou er daarentegen minder kans zijn op een verkeerde classificatie of voorspelling door de ontwikkelde beslissingsboom. Het leek ons dus logisch om te werken met een functie die zowel geschikt was voor sensitiviteitswaarden die boven of onder een bepaalde grenswaarde liggen. Bij het verder uitwerken van de functie zijn we er eveneens vanuit gegaan dat er misschien een andere correctie nodig was voor een waarschijnlijkheidsschatting waarvan de bijhorende sensitiviteit boven de grenswaarde ligt dan voor een waarschijnlijkheidsschatting waarvan de bijhorende sensitiviteit beneden de grenswaarde ligt. Op basis van deze uitgangspunten komen we tot vergelijking 5.1 voor de correctiefunctie.

Aangezien de Standaard metriek en de Min-max metriek verschillende sensitiviteiten per geval hebben maar ook omdat de maximale sensitiviteit bij beide metrieken verschilt, zullen we de metrieken apart behandelen wat betreft de verdere uitwerking van de correctiefunctie in vergelijking 5.1.

Na het selecteren van een grenswaarde voor de sensitiviteit, zullen we voor deze grenswaarde onderzoeken welke effecten de wijzigingen van de oorspronkelijke waarschijnlijkheidsschattingen zullen hebben op de resultaten van de sortering volgens de gecorrigeerde waarschijnlijkheidsschattingen. Deze wijzigingen variëren tussen de nul en de vijftig procent. De uiteindelijke keuze voor een specifieke waarde voor een parameter wordt bepaald door te kijken

welke waarden voor de parameters het meest gunstige effect hebben op de resultaten van de sortering volgens deze gecorrigeerde waarschijnlijkheid.

Standaard metriek

De sensitiviteit die berekend werd met de Standaard metriek ligt in het interval $[0; 1,375939965248]$. Een overzicht van de mogelijke waarden van deze sensitiviteit evenals het aantal gevallen met deze sensitiviteit, is terug te vinden in bijlage F. In tabel 5.4 geven we een overzicht van zowel het procentuele als het absolute aantal gevallen dat in een bepaald sensitiviteitsinterval ligt.

Tabel 5.4: Overzicht van het procentuele en absolute aantal gevallen van de 10000 gevallen in de test dataset dat in een bepaald sensitiviteitsinterval ligt (100% = 10000)

Sensitiviteitsinterval	Aantal gevallen (%)	Aantal gevallen (absoluut)
$[0,000000; 0,000000]$	47,72	4772
$[0,000000; 0,000011]$	54,36	5436
$[0,000000; 0,000012]$	94,37	9437
$[0,000000; 1,375940]$	100,00	10000

Voor elk van de intervallen in tabel 5.4 hebben we de waarschijnlijkheidsschattingen met de waarde van c_1 verhoogd als de sensitiviteit groter is dan de bovengrens van het interval. De waarschijnlijkheidsschattingen van de gevallen waarvan de sensitiviteit kleiner of gelijk is aan de bovengrens van het interval werden verlaagd met de waarde van c_2. De factoren c_1 en c_2 nemen waarden aan tussen nul en vijftig procent. In bijlage G worden de resultaten van deze correctiefunctie getoond. Tabel 5.5 op 58 toont met welke parameters de meest gunstige resultaten werden bereikt voor de Standaard metriek. Met gunstige resultaten bedoelen we zoveel mogelijk juist geclassificeerde gevallen, ongeacht de klasse waartoe ze behoren.

Tabel 5.5: Resultaten van de correctiefunctie bij de Standaard metriek

Parameters			Aantal juist voorspeld			Aantal fout voorspeld		
s_0	$c_1(\%)$	$c_2(\%)$	Klasse 0	Klasse 1	Totaal	Klasse 0	Klasse 1	Totaal
0,000001	0	5 - 50	1944	9	1953	36	11	47
0,000001	5 - 50	0 - 50	1944	9	1953	36	11	47
0,000011	0	5 - 50	1944	9	1953	36	11	47
0,000011	5 - 50	0 - 50	1944	9	1953	36	11	47
Parameters			Aantal juist voorspeld (%)			Aantal fout voorspeld (%)		
s_0	$c_1(\%)$	$c_2(\%)$	Klasse 0	Klasse 1	Totaal	Klasse 0	Klasse 1	Totaal
0,000001	0	5 - 50	97,20	0,45	97,65	1,80	0,55	2,35
0,000001	5 - 50	0 - 50	97,20	0,45	97,65	1,80	0,55	2,35
0,000011	0	5 - 50	97,20	0,45	97,65	1,80	0,55	2,35
0,000011	5 - 50	0 - 50	97,20	0,45	97,65	1,80	0,55	2,35

Onderstaande functies zijn voorbeelden van mogelijke stuksgewijze correctiefuncties voor de standaardmetriek.

$$W^* = \begin{cases} W & als\ s\ >\ 0,000001 \\ W\ -\ 0,15 & als\ s\ \leq\ 0,000001 \end{cases}$$

$$W^* = \begin{cases} W\ +\ 0,05 & als\ s\ >\ 0,000001 \\ W\ -\ 0,50 & als\ s\ \leq\ 0,000001 \end{cases}$$

$$W^* = \begin{cases} W\ +\ 0,15 & als\ s\ >\ 0,000001 \\ W & als\ s\ \leq\ 0,000001 \end{cases}$$

$$W^* = \begin{cases} W\ +\ 0,25 & als\ s\ >\ 0,000001 \\ W\ -\ 0,25 & als\ s\ \leq\ 0,000001 \end{cases}$$

$$\cdots$$

$$W^* = \begin{cases} W\ +\ 0,25 & als\ s\ >\ 0,000011 \\ W\ -\ 0,45 & als\ s\ \leq\ 0,000011 \end{cases}$$

$$W^* = \begin{cases} W\ +\ 0,30 & als\ s\ >\ 0,000011 \\ W\ -\ 0,30 & als\ s\ \leq\ 0,000011 \end{cases}$$

$$W^* = \begin{cases} W\ +\ 0,35 & als\ s\ >\ 0,000011 \\ W & als\ s\ \leq\ 0,000011 \end{cases}$$

$$W^* = \begin{cases} W & als\ s\ >\ 0,000011 \\ W\ -\ 0,10 & als\ s\ \leq\ 0,000011 \end{cases}$$

Indien we één van bovenstaande correctiefuncties toepassen op de dataset en daarna de geclassificeerde gevallen ordenen volgens dalende gecorrigeerde waarschijnlijkheid dan zal 97,65 procent van de 2000 geselecteerde gevallen juist geclassificeerd zijn en aan zo'n 2,35 procent van deze 2000 gevallen zal de verkeerde klasse toegewezen zijn. Dit is ook duidelijk af te leiden uit tabel 5.5 op p 58 en de confusion matrix voor de Standaard metriek in figuur 5.8 op p 60.

	Klasse 0	Klasse 1
Klasse 0	97,20%	0,55%
Klasse 1	0,45%	0,45%

(a) Confusion matrix voor de resultaten van de selectie van 2000 gevallen volgens sortering op gecorrigeerde waarschijnlijkheid (sensitiviteit berekend volgens de Standaard metriek)

	Klasse 0	Klasse 1
Klasse 0	95,20%	1,20%
Klasse 1	2,70%	0,90%

(b) Confusion matrix voor de resultaten van de selectie van 2000 gevallen volgens sortering op gecorrigeerde waarschijnlijkheid (sensitiviteit berekend volgens de Min-max metriek)

Figuur 5.8: Confusion matrices gevalstudie sortering op gecorrigeerde waarschijnlijkheid

Min-max metriek

Het interval $[0; 0,0235140007]$ bevat de mogelijke waarden voor de sensitiviteit berekend met behulp van de Min-max metriek. Een overzicht van alle mogelijke sensitiviteitswaarden voor de Min-max metriek is terug te vinden in bijlage H. Net zoals bij de Standaard-metriek hebben we ons gebaseerd op enkele sensitiviteitsintervallen, zie tabel 5.6, om een correctiefunctie te ontwikkelen.

Tabel 5.6: Overzicht van het procentuele en absolute aantal gevallen van de 10000 gevallen in de test dataset dat in een bepaald sensitiviteitsinterval ligt (100% = 10000)

Sensitiviteitsinterval	Aantal gevallen (%)	Aantal gevallen (absoluut)
$[0,000000; 0,000000]$	94,79	9479
$[0,000000; 0,000006]$	96,12	9612
$[0,000000; 0,0000100000]$	96,53	9653
$[0,000000; 0,000048999998]$	98,00	9800
$[0,000000; 0,0001830000]$	99,00	9900
$[0,000000; 0,023514000699]$	100,00	10000

Uit tabel 5.7 op p 62 wordt echter duidelijk dat het gebruiken van een correctiefunctie voor de waarschijnlijkheidsschattingen, niet echt voor een ver-

betering van de resultaten van de sortering zorgt. Het toepassen van de correctiefunctie brengt zelfs een verslechtering van de resultaten met zich mee. Het aantal juist geclassificeerde gevallen valt terug van 95,20 procent naar 95 of 94,75 procent terwijl het aantal fout geclassificeerde gevallen toeneemt van 4,80 procent naar 5 of 5,25 procent.

Als we de confusion matrices in figuur 5.8 op p 60, gebaseerd op de meest gunstige resultaten van beide metrieken, bekijken dan zien we dat de Standaard metriek beter presteert wat betreft het aantal gevallen in het True Positive-kwadrant. De Min-max metriek levert betere prestaties voor het aantal gevallen in het True Negative-kwadrant.

Uit dit hoofdstuk is het duidelijk geworden dat de manier waarop de geclassificeerde gevallen gesorteerd worden bepalend is voor de resultaten die men bekomt. Als we een sortering maken die gebaseerd is op de sensitiviteit dan zal ook de metriek gebruikt om de sensitiviteit te berekenen een invloed hebben op de resultaten.

Tabel 5.7: Resultaten van de correctiefunctie bij de Min-max metriek

Parameters			Aantal juist voorspeld			Aantal fout voorspeld		
s_0	$c_1(\%)$	$c_2(\%)$	Klasse 0	Klasse 1	Totaal	Klasse 0	Klasse 1	Totaal
0 - 0,023514000699	0	0	1904	18	1922	54	24	78
0,000048999998	0	5 - 10	1895	26	1921	54	25	79
0,000048999998	5	0 - 5	1895	26	1921	54	25	79
0,000048999998	10	0	1895	26	1921	54	25	79
0,000183000000	0	5	1900	21	1921	54	25	79
0,000183000000	0	5-10	1900	21	1921	54	25	79
0,000183000000	5	0-5	1900	21	1921	54	25	79
0,000183000000	10	0	1900	21	1921	54	25	79
0,023514000699	0	0 - 50	1904	18	1922	54	24	78
Parameters			Aantal juist voorspeld (%)			Aantal fout voorspeld (%)		
s_0	$c_1(\%)$	$c_2(\%)$	Klasse 0	Klasse 1	Totaal	Klasse 0	Klasse 1	Totaal
0 - 0,023514000699	0	0	95,20	0,90	96,10	2,70	1,20	3,90
0,000048999998	0	5 - 10	94,75	1,30	96,05	2,70	1,25	3,95
0,000048999998	5	0 - 5	94,75	1,30	96,05	2,70	1,25	3,95
0,000048999998	10	0	94,75	1,30	96,05	2,70	1,25	3,95
0,000183000000	0	5	95,00	1,05	96,05	2,70	1,25	3,95
0,000183000000	0	5-10	95,00	1,05	96,05	2,70	1,25	3,95
0,000183000000	5	0-5	95,00	1,05	96,05	2,70	1,25	3,95
0,000183000000	10	0	95,00	1,05	96,05	2,70	1,25	3,95
0,023514000699	0	0 - 50	95,20	0,90	96,10	2,70	1,20	3,90

Hoofdstuk 6

Conclusies

Dit hoofdstuk bevat het antwoord op de onderzoeksvraag uit de probleemstelling. Dit antwoord is gebaseerd op de conclusies omtrent het onderzoek en de resultaten van deze thesis. We geven ook enkele aanbevelingen en suggesties voor verder onderzoek.

6.1 Conclusies

Het doel van deze thesis was, aan de hand van een gevalstudie, een antwoord trachten te geven op onderstaande centrale onderzoeksvraag.

Op welke manier kan het sensitiviteitsalgoritme gebruikt worden als evaluatiemethode voor de resultaten van binaire beslissingsbomen?

Na het uitproberen en vergelijken van diverse sorteermethodes kunnen we besluiten dat er niet echt een 'beste' methode blijkt te zijn voor het rangschikken van de resultaten gegeven door een binaire beslissingsboom.

Afhankelijk van het doel waarop men zich richt, zal het sensitiviteitsalgoritme al dan niet op zichzelf gebruikt kunnen worden om de geclassificeerde gevallen te rangschikken. Dit wordt duidelijk wanneer we de resultaten uit het vorige hoofdstuk bekijken in de samenvattende tabel 6.1 op p 65.

In ons onderzoek is de correctiefunctie het meest effectief wanneer ze gebaseerd wordt op de sensitiviteit berekend met behulp van de Standaardmetriek. Indien de correctiefunctie daarentegen ontwikkeld wordt aan de hand van de sensitiviteit berekend volgens de Min-max metriek, dan waren er minder gunstige resultaten. De rangschikkingen waren minder goed dan wanneer ze gebaseerd zouden zijn op de oorspronkelijke waarschijnlijkheidsschattingen gegeven door de beslissingsboom. Onderstaande vergelijking is een algemene voorstelling van de gebruikte stuksgewijze correctiefunctie.

$$
\begin{aligned}
s = &\ \textit{sensitiviteit van een specifiek geval uit de dataset} \\
s_0 = &\ \textit{grenswaarde van de sensitiviteit} \\
c_1 = &\ \textit{wijziging van de oorspronkelijke} \\
&\ \textit{waarschijnlijkheidsschatting} \\
c_2 = &\ \textit{wijziging van de oorspronkelijke} \\
&\ \textit{waarschijnlijkheidsschatting} \\
W = &\ \textit{oorspronkelijke waarschijnlijkheidsschatting} \\
W^* = &\ \textit{aangepaste of gecorrigeerde waarschijnlijkheidsschatting}
\end{aligned}
$$

$$
W^* = \begin{cases} W + c_1 & als\ s > s_0 \\ W - c_2 & als\ s \leq s_0 \end{cases}
$$

Tabel 6.1: Resultaten van de selectie van 2000 gevallen volgens de in dit onderzoek geteste sorteringen (100 % = 2000, cv = cross-validation, ws = waarschijnlijkheidsschatting)

Sorteer-methode	Aantal juist voorspeld (%)			Aantal fout voorspeld (%)		
	Klasse 0	Klasse 1	Totaal	Klasse 0	Klasse 1	Totaal
10-fold cv	95,70	0,90	96,60	2,20	1,20	3,40
10x10-fold cv (gemiddelde)	77,83	6,04	83,86	7,00	9,14	16,14
Getransformeerde ws	47,00	23,55	70,55	15,35	14,10	29,45
Sensitiviteit Standaard metriek	68,30	19,30	87,60	3,55	8,85	12,40
Min-max metriek	71,25	19,15	90,40	0,90	8,70	9,60
Product Standaard metriek	71,30	19,15	90,45	0,85	8,70	9,55
Min-max metriek	71,25	19,15	90,40	0,90	8,70	9,60
Correctie-functie Standaard metriek	97,20	0,45	97,65	1,80	0,55	2,35
Min-max metriek	95,20 94,75	0,90 1,30	96,10 96,05	2,70 2,70	1,20 1,25	3,90 3,95

Wanneer we de geclassificeerde gevallen rangschikken op basis van de sensitiviteit of op het product van de sensitiviteit en de waarschijnlijkheidsschatting van een specifiek geval, dan zien we dat er meer juiste classificaties van klasse 1 gedaan worden. Het percentage juiste classificaties van klasse 1 is beduidend groter bij deze sorteermethodes dan bij de anderen. De enige sorteermethode die nog beter scoort wat betreft het juist classificeren van de gevallen die behoren tot klasse 1, is de methode gebaseerd op de getransformeerde waarschijnlijheidsschattingen. Maar indien we deze methode toepassen, moeten we ons wel tevreden stellen met een significant lager aandeel juiste classificaties voor klasse 0.

Als laatste willen we vermelden dat de conclusies niet algemeen geldend zijn. Zij werden geformuleerd in het kader van deze thesis en de daarbij gebruikte dataset. Verder onderzoek zal nodig zijn alvorens men de geformuleerde conclusies en resultaten kan veralgemenen.

6.2 Aanbevelingen en suggesties voor verder onderzoek

We kunnen onze aanbevelingen en suggesties richten op drie onderdelen van dit onderzoek. Een eerste onderdeel betreft de dataset die gebruikt wordt om de binaire beslissingsboom te ontwikkelen. Het tweede onderdeel waarvoor we enkele aanbevelingen geformuleerd hebben heeft betrekking tot de beslissingsboom die gebruikt wordt om het sensitiviteitsalgoritme te kunnen toepassen. Als laatste worden er dan enkele suggesties gegeven omtrent het sensitiviteitsalgoritme zelf.

6.2.1 Dataset

In onze gevalstudie werd er gebruik gemaakt van de dataset die ter beschikking werd gesteld in het kader van de data mining competitie. Om het aantal attributen, dat gebruikt zal worden om de beslissingsboom te ontwikkelen, zo beperkt mogelijk te houden zou men misschien gebruik kunnen maken van *atrribute selection* technieken. Zo zou men enkel de belangrijkste attributen in overweging kunnen nemen voor het ontwikkelen van de beslissingsboom.

6.2.2 Beslissingsboom

Aangezien het tijdsbestek van deze thesis beperkt was, zou er meer werk gestoken kunnen worden in het ontwikkelen van de beslissingsboom. Zoals in de vorige paragraaf al even aangehaald werd, zijn er veel attributen aanwezig in de gebruikte dataset. Door een uitgebreidere verkenning van de data zou de ontwikkelde beslissingsboom misschien verbeterd of kleiner gemaakt kunnen worden.

We hebben bij het ontwikkelen van de beslissingsboom meestal gebruik gemaakt van de standaardwaarden voor de diverse parameters die ingesteld kunnen worden bij de ontwikkeling van een beslissingsboom. Een aanbeveling die hieruit volgt is om bij de ontwikkeling van de beslissingsboom de parameters zoals bijvoorbeeld het minimum aantal gevallen dat in een blad moet zitten aan te passen om zo de optimale waarde ervan te zoeken.

6.2.3 Sensitiviteitsalgoritme

Een eerste opmerking omtrent het bestaande sensitiviteitsalgoritme, ontwikkeld door Isabelle Alvarez, is dat de naamgeving misschien niet echt geschikt is. Uit de definitie van sensitiviteit konden we immers afleiden dat hoe groter deze sensitiviteit of afstand is, hoe minder kans er is dat er een verkeerde voorspelling of classificatie heeft plaatsgevonden. Als men op de logische redenering afgaat dan lijkt het mekaar namelijk tegen te spreken dat indien de sensitiviteit of 'gevoeligheid' groter wordt, de kans op een verkeerde voorspelling kleiner wordt.

Aangezien dit onderzoek slechts gebaseerd was op één enkele dataset kunnen de resultaten en conclusies hieromtrent niet veralgemeend worden. Een voor de hand liggende suggestie is dan natuurlijk het sensitiviteitsalgoritme toe te passen op meerdere datasets. We dienen hierbij wel op te merken dat dit enkel datasets mogen zijn met numerieke attributen of attributen waarvan de waarden omgezet kunnen worden naar een metriek. Indien dit niet zo is zal het moeilijk, misschien zelfs onmogelijk worden om de afstanden tussen een specifiek geval en het bijhorende beslissingsoppervlak te berekenen. Een andere reden om het algoritme toe te passen op meerdere datasets bestaat erin dat niet alle ontwikkelde algoritmes en methodes geschikt zijn om op iedere dataset toegepast te worden.

Een laatste suggestie gaat over het selecteren van de grenswaarde van de

sensitiviteit bij het ontwikkelen van een correctiefunctie. Ook hier zou men meer uitgebreid kunnen testen welke waarde het meest geschikt of optimaal is naargelang de doelen die men voorop stelt. Er zal waarschijnlijk een andere grenswaarde nodig zijn wanneer men er veel belang aan hecht dat vooral de gevallen met een weinig frequente klasse juist geclassificeerd worden dan wanneer men hier niet veel belang aan hecht. Onder meer uitgebreid testen verstaan we bijvoorbeeld het gebruiken van meerdere beslissingsbomen om de ideale grenswaarde te bepalen. Een ander voorbeeld is alle mogelijke sensitiviteitswaarden van de dataset na te gaan op hun geschiktheid om gebruikt te worden als grenswaarde.

Bibliografie

Isabelle Alvarez, 2004. *Sensitivity Analysis of the result in Binary Decision Trees.* Technical report.

Michael J.A. Berry en Gordon Linoff, 1997. *Data Mining Techniques For Marketing, Sales and Customer Support.* John Wiley and Sons, Inc., first^{e} druk. ISBN 0-471-17980-9.

Michael J.A. Berry en Gordon Linoff, 2000. *Mastering Data Mining.* John Wiley and Sons Inc., second^{e} druk. ISBN 0-471-33123-6.

Eibe Frank, 2005. *Weka.* `http://www.cs.waikato.ac.nz/ml/weka/`.

Marc Gyssens, 2001. *Algoritme en datastructuren.* Limburgs Universitair Centrum.

David Hand, Heikki Mannila, en Padhraic Smyth, 2001. *Principles of Data Mining.* Bradford Book, first^{e} druk. ISBN 0-262-08290-X.

A. Ittner en M. Schlosser, 1998. *Non Linear Decision Trees - NDT.* Technical report, Department of Computer Science, AI Research Group Chemnitz University of Technology Germany and Department of Electrical Engineering, Fachhochschule Koblenz Germany.

D. D. Margineantu en T. G. Dietterich, 2001. *Improved Class Probability Estimates from Decision Tree Models.* Technical report, Department of Computer Science, Oregon State University.

T.M. Mitchell, 1997. *Machine Learning.* New-York, McGraw Hill. ISBN 0070428077.

S. K. Murthy, 1998. *Automatic Construction of Decision Trees from Data: A Multi-Disciplinary Survey.* Technical report, Siemens Corporate Research, Princeton.

Michael Negnevitsky, 2001. *Artificial Intelligence: A guide to intelligent systems*. Addison Wesley. ISBN 0201711591.

James A. O'Brien, 1998. *Leerboek ICT-toepassingen*. Academic Service, third^{e} druk. ISBN 90-395-0895.

F. Provost en P. Domingos, 2000. *Well-Trained PETs: Improving Probability Estimation Trees*. Technical report, Information Systems Department, Stern School of Business-New York University.

J.R. Quinlan, 1993. *C4.5: Programs for Machine Learning*. Morgan Kaufmann publishers. ISBN 1558602380.

Two-Crows-Corporation, 2005. *Data Mining Glossary*. `http://www.twocrows.com/glossary.htm`.

Sue Walsh, 2003. *Applying Data Mining Techniques using Enterprise Miner^{TM} Course Notes*. SAS Institute Inc., first^{e} druk. ISBN 1-59047-090-7.

Eric W. Weisstein, 2005a. *Conjunction*. From MathWorld–A Wolfram Web Resource: `http://mathworld.wolfram.com/Conjunction.html`.

Eric W. Weisstein, 2005b. *Disjunction*. From MathWorld–A Wolfram Web Resource: `http://mathworld.wolfram.com/Disjunction.html`.

Eric W. Weisstein, 2005c. *Distance*. From MathWorld–A Wolfram Web Resource: `http://mathworld.wolfram.com/Distance.html`.

Eric W. Weisstein, 2005d. *Element*. From MathWorld–A Wolfram Web Resource: `http://mathworld.wolfram.com/Element.html`.

Eric W. Weisstein, 2005e. *Euclidean space*. From MathWorld–A Wolfram Web Resource: `http://mathworld.wolfram.com/EuclideanSpace.html`.

Eric W. Weisstein, 2005f. *n-Tuple*. From MathWorld–A Wolfram Web Resource: `http://mathworld.wolfram.com/n-Tuple.html`.

Eric W. Weisstein, 2005g. *Open ball*. From MathWorld–A Wolfram Web Resource: `http://mathworld.wolfram.com/OpenBall.html`.

Eric W. Weisstein, 2005h. *Projection*. From MathWorld–A Wolfram Web Resource: `http://mathworld.wolfram.com/Projection.html`.

Eric W. Weisstein, 2005i. *Scalair*. From MathWorld–A Wolfram Web Resource: `http://mathworld.wolfram.com/Scalar.html`.

Eric W. Weisstein, 2005j. *Set*. From MathWorld–A Wolfram Web Resource: `http://mathworld.wolfram.com/Set.html`.

Eric W. Weisstein, 2005k. *Vector*. From MathWorld–A Wolfram Web Resource: `http://mathworld.wolfram.com/Vector.html`.

I.H. Witten en E. Frank, 2000. *Data Mining: practical machine learning tools and techniques with java implementations*. Morgan Kaufmann publishers. ISBN 1558605525.

B. Zadrozny en C. Elkan, 2001. *Obtaining calibrated probability estimates from decision trees and naive Bayesian classifiers*. Technical report, Department of Computer Science and Engineering, University of California.

Bijlage A

Het C4.5 algoritme

Deze bijlage is gebaseerd op het boek *C4.5: Programs for Machine Learning* van *J.R. Quinlan* (Quinlan 1993). Het C4.5 algoritme maakt classificatiemodellen door patronen te ontdekken in een verzameling records en deze patronen naderhand ook te analyseren. Talrijke, reeds vastgelegde, classificaties zullen onderzocht en gebruikt worden om zo een model te kunnen construeren door de specifieke voorbeelden te veralgemenen.

Sleutelvereisten om C4.5 toe te passen

Om deze inductieve benadering van het C4.5 algoritme te kunnen toepassen op een classificatieprobleem moet het probleem aan enkele belangrijke vereisten voldoen.

- Een eerste vereiste betreft de beschrijving van de attribuutwaarden. Ieder object, in de te analyseren data, moet uitgedrukt kunnen worden in termen van een vaste verzameling van kenmerken of attributen. Een attribuut mag zowel discrete als numerieke waarden aannemen maar de attributen die gebruikt worden om een geval te beschrijven mogen niet variëren tussen de verschillende gevallen.

- De categorieën of klassen waaraan de gevallen toegekend zullen worden, moeten vooraf bepaald en vastgelegd zijn.

- De klassen moeten goed afgelijnd zijn zodat het duidelijk is tot welke klasse een geval behoort. Er moeten ook veel meer gevallen dan klassen zijn.

- De aanwezigheid van voldoende data is ook belangrijk. Inductieve generalisatie gebeurt namelijk door het identificeren van patronen in data. Deze benadering zal mislukken als geldige, robuuste patronen niet ontdekt kunnen worden of onderscheiden kunnen worden van toeval. De hoeveelheid data die nodig is, wordt onder andere bepaald door het aantal attributen, het aantal klassen en de complexiteit van het classificatiemodel. Als deze factoren toenemen, zal er meer data nodig zijn om een betrouwbaar model te kunnen genereren.

- Als laatste moet de beschrijving van een klasse beperkt blijven tot een logische expressie bestaande uit veronderstellingen in verband met de waarden van specifieke attributen.

Constructie van beslissingsbomen

Het genereren van een beslissingsboom uit een set T van traininggevallen kan als volgt uitgelegd worden. Laat de diverse klassen voorgesteld worden door $\{K_1, K_2, \ldots, K_k\}$, dan zijn er drie mogelijkheden wat betreft T.

- *T bevat één of meerdere gevallen die allemaal behoren tot eenzelfde klasse K_j:* de beslissingsboom voor T is een blad dat klasse K_j zal identificeren.
- *T bevat geen gevallen:* de beslissingsboom zal opnieuw uit slechts één blad bestaan maar de klasse die geassocieerd moet worden met het blad moet bepaald worden met informatie die men niet uit T kan halen. C4.5 zal hiervoor dan de klasse gebruiken die het meest voorkomt bij de ouder van deze knoop.
- *T bevat gevallen die behoren tot een mengeling van klassen:* in dit geval moet T verder opgesplitst worden in subsets van gevallen die uiteindelijk zullen resulteren in verzamelingen van gevallen waar slechts één klasse aanwezig is. Er wordt een test gekozen die, gebaseerd op één enkel attribuut, één of meer wederzijds uitsluitende resultaten $\{R_1, R_2, \ldots, R_n\}$ heeft. T wordt verdeeld in subsets $T_1, T_2, \ldots, T_n$ waarbij T_i alle gevallen uit T bevat die R_i als resultaat van de gekozen test hebben. De beslissingsboom voor T bestaat uit een beslissingsknoop die de test identificeert en één tak voor ieder mogelijk resultaat. Dit mechanisme zal recursief toegepast worden op iedere subset van traininggevallen zodat de i-de tak leidt tot een beslissingsboom geconstrueerd uit de subset T_i van traininggevallen.

Zoals uit het voorgaande misschien al duidelijk werd, gaat het hier om een 'nonbacktracking' methode. Dit betekent dat zodra er een test geselecteerd is voor het splitsen van de gevallen in de training set er geen andere, alternatieve testen meer verkend of onderzocht zullen worden. Er zijn dus manieren nodig om de kwaliteit van de testen in de knopen van een beslissingsboom te beoordelen.

Evaluatie van testen in de knopen van een beslissingsboom

De evaluatie van testen in de knopen van een beslissingsboom kan onder andere gebeuren met behulp van het *Gain Criterium* en het *Gain Ratio*

Criterium. Beide evaluatiemethodes worden in deze subparagraaf besproken.

Gain Criterium

Veronderstel dat we een test met n resultaten hebben die de set T van trainginggevallen verdeeld in de subsets $T_1, T_2, \ldots, T_n$. Als deze test geëvalueerd moet worden zonder enige hieropvolgende verdelingen van de T_i's te verkennen, dan is de enige informatie die dienst kan doen als hulpmiddel, de verdeling van klassen in T en diens subsets.

Laat S een verzameling van gevallen zijn en $|S|$ het aantal gevallen dat zich in S bevindt. Het aantal gevallen in S dat behoort tot klasse K_i, wordt weergegeven door $freq(K_i, S)$.

De informatie overgedragen door een boodschap is afhankelijk van de kans dat de boodschap voorkomt en kan gemeten worden in bits als $-\log_2$ van die kans. Selecteer een willekeurig geval uit een set S en laat dit geval als klasse K_j hebben. De boodschap van dit geval heeft als waarschijnlijkheid:

$$\frac{freq(K_j, S)}{|S|} \tag{1}$$

en draagt de volgende hoeveelheid informatie over:

$$-\log_2 \left(\frac{freq(K_j, S)}{|S|} \right) bits \tag{2}$$

Om de verwachte informatie van deze boodschap met betrekking tot klasselidmaatschap te vinden, sommeren we over de klassen rekening houdend met hun frequenties in S:

$$info(S) = -\sum_{j=1}^{k} \frac{freq(K_j, S)}{|S|} \times -\log_2 \left(\frac{freq(K_j, S)}{|S|} \right) \tag{3}$$

Wanneer dit toegepast wordt op de training set T, dan meet $info(T)$ de gemiddelde hoeveelheid informatie die nodig is om de klasse van een geval in T te identificeren. Deze hoeveelheid wordt ook wel eens de entropie van de set S genoemd. Beschouw nu een gelijkaardige maatstaf nadat T verdeeld is in overeenkomst met de n resultaten van een test X. De verwachte informatievereiste kan gevonden worden als de gewogen som over de subsets:

$$info_X(T) = -\sum_{i=1}^{n} \frac{|T_i|}{|T|} \times info(T_i) \tag{4}$$

De hoeveelheid berekend in vergelijking 5 op p 76 meet de informatie die gewonnen wordt door T te verdelen in overeenkomst X.

$$gain(X) = info(T) - info_X(T) \tag{5}$$

Het *Gain Criterium* zal dan uiteindelijk de test selecteren die deze informatiegain zal maximaliseren. Een nadeel van dit criterium is echter dat het een sterke voorliefde heeft voor testen met veel mogelijke resultaten. Om deze afwijking recht te zetten werd het *Gain Ratio Criterium* ontwikkeld.

Gain Ratio Criterium

Het *Gain Ratio Criterium* maakt gebruik van een soort normalisatie waarbij de ogenschijnlijke gain, toewijsbaar aan testen met veel resultaten, aangepast wordt. Beschouw de informatie-inhoud van een boodschap toebehorend aan een geval dat niet de klasse aangeeft waartoe het geval behoort, maar wel het resultaat van de test. In analogie met de definitie van info(S) in vergelijking 3 op p 75 bekomen we nu:

$$split\ info(X) = -\sum_{i=1}^{n} \frac{|T_i|}{|T|} \times \log_2 \left(\frac{|T_i|}{|T|} \right) \tag{6}$$

Deze vergelijking 6 op p 76 stelt de potentiële informatie voor die gegenereerd wordt door T te verdelen in n subsets. De informatie gain daarentegen, meet de hoeveelheid informatie die relevant is voor de classificatie die uit dezelfde verdeling voortkomt. Vergelijking 7 op p 76 drukt dan de proportie van informatie, gegenereerd door de splitsing, uit die nuttig is voor het classificeren.

$$gain\ ratio(X) = \frac{gain(X)}{split\ info(X)} \tag{7}$$

Als de split onbelangrijk is, zal de splitinformatie uit vergelijking 6 op p 76 klein zijn en de ratio onstabiel zijn. Om dit te vermijden selecteert het *Gain Ratio Criterium* een test die de ratio uit vergelijking 7 op p 76 maximaliseert rekening houdend met de beperking dat de informatie gain groot moet zijn. Met groot bedoelen we minstens zo groot als de gemiddelde gain van alle onderzochte testen.

Mogelijke testen in de knopen van een beslissingsboom

Traditioneel bevat een test slechts één attribuut omdat dit de boom gemakkelijker te begrijpen maakt en omdat zo ook de combinatorische explosie vermeden wordt die kan ontstaan door het gebruiken van meerdere attributen in één enkele test. Het C4.5 algoritme bevat mechanismen voor het genereren van drie types van testen: de standaard test, een test gebaseerd op een discreet attribuut en tenslotte een test gebaseerd op een continu numeriek attribuut. Al deze testen worden op dezelfde manier getest en geëvalueerd, namelijk door het *Gain Criterium* of het *Gain Ratio Criterium* te gebruiken.

De standaard test

Deze test genereert slechts één resultaat en één tak voor iedere mogelijke waarde van het betreffend attribuut.

De test op basis van een discreet attribuut

Dit is een meer complexe test waarbij de mogelijke waarden gealloceerd worden aan een variabel aantal groepen met één resultaat en één tak voor iedere groep in plaats van voor iedere waarde.

De test op basis van een continu, numeriek attribuut

Bij het genereren van testen op continue attributen worden de traininggevallen in T eerst gesorteerd op basis van de waarden van het betreffend attribuut. Stel dat A een continu numeriek attribuut is. Er is slechts een eindig aantal van deze waarden mogelijk in de dataset, dus we kunnen deze in volgorde noteren als $\{w_1, w_2, \ldots, w_n\}$. Eender welke grenswaarde die tussen w_i en w_{i+1} ligt, zal hetzelfde effect hebben als het verdelen van de gevallen waarbij de waarde van attribuut A in $\{w_1, w_2, \ldots, w_i\}$ ligt en diegenen waarbij de waarde van attribuut A in $\{w_{i+1}, \ldots, w_n\}$ ligt. Er zijn dus slechts $m-1$ mogelijke splits voor A, waarbij elk ervan onderzocht zal worden. De uiteindelijke grenswaarde die door C4.5 gekozen zal worden, is de grootste waarde van A in de gehele training set die niet boven het punt $\left(\frac{w_i+w_{i+1}}{2}\right)$ ligt. Dit zorgt ervoor dat alle grenswaarden die in de boom zullen verschijnen ook

echt in de data voorkomen.

Aanpassingen nodig voor het opvangen van onbekende attribuutwaarden

Omdat iedere test gebaseerd is op minstens één attribuut zal het resultaat van een test enkel bepaald kunnen worden als de waarde van het attribuut gekend is. Aangezien het weglaten van deze data zou leiden tot een vermindering van de bekwaamheid om patronen te vinden, is er een oplossing in de vorm van aanpassingen aan het algoritme waardoor het met attributen met missende waarden kan omgaan. Om het algoritme te kunnen aanpassen moeten er drie vragen beantwoord worden:

- Het selecteren van de test, die gebruikt zal worden om de training set te verdelen, wordt gebaseerd op heuristische criteria zoals bijvoorbeeld het *Gain Criterium* of het *Gain Ratio Criterium*. Als twee testen attributen gebruiken met verschillende aantallen van ongekende waarden, hoe moet dit dan in rekening gebracht worden wanneer hun relatieve gewenstheid afgewogen wordt?

- Eens een test geselecteerd is, kunnen traininggevallen met onbekende waarden niet geassocieerd worden met een bepaald resultaat van de betreffende test waardoor ze dus ook niet toegewezen kunnen worden aan een specifieke subset T_i. Hoe moeten deze gevallen behandeld worden bij het verdelen?

- Wanneer de beslissingsboom gebruikt wordt om een ongezien geval te classificeren, hoe moet het systeem dan handelen wanneer het geval een ongekende waarde heeft voor het attribuut dat in de huidige beslissingsknoop getest wordt?

In hetgeen wat volgt, wordt beschreven hoe het C4.5-algoritme werd aangepast om een antwoord te kunnen geven op deze drie vragen.

Evaluatie van testen

Het ligt voor de hand dat een test geen informatie kan voorzien over het lidmaatschap tot een klasse van een geval als de waarde van het attribuut dat getest wordt ongekend is. Laat T de training set zijn en X een test gebaseerd

op attribuut A. Veronderstel ook dat de waarde van A gekend is in een fractie F van de gevallen in de training set. Zowel $info(T)$ als $info_X(T)$ kunnen berekend worden als voorheen behalve dat enkel de gevallen met gekende waarden voor A in rekening gebracht mogen worden. De definitie van *gain* kan dan tot de formule in vergelijking 8 op p 79 omgevormd worden:

$$
\begin{aligned}
&P(A_{gekend}) = waarschijnlijkheid\ dat\ de\ waarde\ van\ A\ gekend\ is \\
&P(A_{niet\ gekend}) = waarschijnlijkheid\ dat\ de\ waarde\ van\ A\ niet\ gekend\ is
\end{aligned}
$$

$$
\begin{aligned}
gain(X) &= P(A_{gekend}) \times (info(T) - info_X(T)) + P(A_{niet\ gekend}) \times 0 \\
&= F \times (info(T) - info_X(T))
\end{aligned}
\tag{8}
$$

De definitie van *split* $info(X)$ in vergelijking 6 op p 76 kan op een gelijkaardige manier aangepast worden door de gevallen met ongekende waarden als een additionele groep te bekijken. Als een test n resultaten heeft, dan zal de *split informatie* berekend worden alsof de test $n + 1$ resultaten heeft.

Verdeling van de training set

Wanneer het resultaat niet gekend is, is het moeilijk om het betreffende traininggeval toe te wijzen aan een specifieke subset T_i. Daarom kent C4.5 een gewicht toe aan elk geval in elke subset T_i dat de kans voorstelt dat het geval behoort tot iedere subset. Als het geval een gekend resultaat heeft dan is dit gewicht 1, maar als het geval een onbekend resultaat heeft dan is het gewicht de kans dat resultaat R_i voorkomt. Iedere subset T_i is dus nu een verzameling van mogelijke fractionele gevallen zodat $|T_i|$ nu geïnterpreteerd moet worden als de som van de fractionele gewichten van de gevallen in de set. Dus in het algemeen wordt een geval uit T met gewicht w, wiens resultaat niet gekend is, toegewezen aan iedere subset T_i met gewicht $w \times P(R_i)$. De kans op resultaat R_i of $P(R_i)$ wordt geschat als de som van de gewichten van de gevallen in T, waarvan geweten is dat ze als resultaat R_i hebben, gedeeld door de som van de gewichten van de gevallen die een resultaat hebben voor deze test.

Classificatie van een ongezien geval

Wanneer er een blad bereikt wordt waarbij de attribuutwaarde voor het specifieke geval niet gekend is, zodat het resultaat van de test niet bepaald kan worden dan zal het systeem alle mogelijke resultaten verkennen en de resulterende classificaties rekenkundig combineren. Omdat er verschillende paden

vanaf de wortel van een beslissingsboom of subboom mogelijk zijn, zal een classificatie eerder een klasseverdeling zijn dan één enkele klasse. Wanneer de totale klasseverdeling voor het ongekende geval op deze manier werd samengesteld, dan zal de klasse met de hoogste kans op voorkomen toegewezen worden als de voorspelde klasse.

Het snoeien van beslissingsbomen

De recursieve verdeelmethode voor het construeren van beslissingsbomen zal de set van traininggevallen blijven herverdelen totdat iedere subset in de verdeling gevallen van dezelfde klasse bevat of totdat geen enkele test nog enige verbetering kan veroorzaken. Hierdoor ontstaat er dikwijls een complexe boom die de data *overfit* door meer structuur af te leiden dan gerechtvaardigd is door de gevallen in de training set. Hier komt ook bij dat deze complexe boom niet noodzakelijk een lagere foutenratio dan een eventuele simpelere boom zal hebben.

Het simplificeren van een beslissingsboom zal gebeuren door delen van de boom te verwijderen die niet bijdragen aan de classificatienauwkeurigheid betreffende ongeziene gevallen. Hierdoor zal de boom minder complex en beter begrijpbaar worden. Er zijn twee manieren waarop deze simplificatie doorgevoerd kan worden:

- *Pre-snoeien:* er is besloten om een set van traininggevallen niet verder te verdelen. Een voordeel bij deze methode is dat er geen tijd verspild wordt met een structuur op te bouwen die naderhand niet gebruikt wordt in de uiteindelijke gesnoeide boom. De typische benadering hierbij is dat men gaat kijken naar de beste manier om de subset te splitsen en dan deze gaat beoordelen aan de hand van bijvoorbeeld de $informatie\ gain$. Als deze beoordeling beneden een voorafbepaalde grenswaarde valt, dan zal een verdere verdeling afgewezen worden en dan wordt de boom voor deze subset het meest geschikte blad. Een nadeel bij deze methode is dat zowel een te hoge als een te lage grenswaarde de prestaties van de boom sterk kunnen verminderen.

- *Post-snoeien:* na het bouwen van de boom gaat men delen van de structuur terug verwijderen. Bij deze methode laat men het verdeel-en-heers-principe volledig zijn gang gaan en gaat men achteraf de overfitte beslissingsboom snoeien. Deze methode heeft als nadeel dat er meer tijd nodig is omdat eerst de volledige boom gebouwd zal worden

maar dit tijdverlies weegt niet op tegen de voordelen die de meer gedetailleerde verkenning van mogelijke verdelingen met zich meebrengt. Beslissingsbomen eerst volledig samenstellen en dan snoeien is dus langzamer maar wel betrouwbaarder dan de bomen te snoeien terwijl men ze construeert.

Beslissingsbomen zullen dus meestal vereenvoudigd worden door een subboom te vervangen door een blad. De klasse geassocieerd aan een blad wordt net als voorheen gevonden door de traininggevallen in een blad te onderzoeken en de meest voorkomende klasse te kiezen. Het C4.5 algoritme staat ook toe dat een subboom vervangen wordt door één van de eigen takken.

Als het verwijderen van takken en/of subbomen een daling van de verwachte foutenratio in die subboom tot gevolg heeft dan zal ook de foutenratio van de beslissingsboom afnemen. Het kunnen voorspellen van de foutenratio van de beslissingsboom, de subbomen en bladeren van de beslissingsboom zou het mogelijk maken om onderaan de boom te beginnen en iedere subboom die niet samenvalt met een blad of eindknoop te onderzoeken. Als vervanging van deze subboom met een blad of een eigen meest gebruikte tak tot een lagere foutenratio zou leiden, dan moet de boom gesnoeid worden. Er zijn twee soorten technieken om de foutenratio te voorspellen:

- *Voorspelling van de foutenratio door gebruik te maken van een nieuwe set van gevallen die verschillend is van de training set:* omdat deze gevallen niet onderzocht werden toen de boom geconstrueerd werd, zijn de schattingen die van deze gevallen bekomen zijn zeker onbevooroordeeld en als er genoeg zijn, ook betrouwbaar. Het nadeel van methodes die deze techniek gebruiken, is dat een deel van de beschikbare data gereserveerd moet worden voor de aparte set waardoor de beslissingsboom uit een kleinere dataset gebouwd moet worden. Dit kan problemen geven wanneer men niet over een voldoende grote dataset beschikt. Een mogelijke oplossing hiervoor is cross-validatie. Vooraleer we cross-validatie kunnen toepassen moeten we beslissen in hoeveel delen we de dataset willen verdelen. Stel dat we kiezen voor een hoeveelheid van a delen, dan zal de dataset verdeeld worden in a ongeveer gelijke partities. Iedere partitie zal dan om de beurt gebruikt worden om te testen terwijl de overblijvende partities gebruikt worden om het model te trainen. Deze procedure wordt a keer herhaald zodat uiteindelijk iedere instantie exact één keer gebruikt werd voor het testen. Meer specifiek is *Stratified Ten Fold Cross-validation* de standaard manier geworden om de foutenratio van een leertechniek te voorspellen. Met

stratified bedoelen we dat de cross-validatie op zo'n manier is gebeurd dat iedere klasse voldoende vertegenwoordigd werd in elk van de 10 partities waarin de dataset verdeeld werd (Witten en Frank 2000).

- *Voorspelling van de foutenratio door enkel gebruik te maken van de training set waarmee de boom gebouwd werd:* de kans op fouten kan niet exact bepaald worden maar heeft wel een kansverdeling die samengevat kan worden door een paar betrouwbaarheidslimieten. Voor een gegeven betrouwbaarheidsniveau CF, kan de bovenlimiet van deze kans gevonden worden met behulp van de betrouwbaarheidslimieten van de binomiale verdeling [1], deze bovenlimiet zullen we noteren als $U_{CF}(E, N)$. C4.5 zal dan de voorspelde foutenratio in een blad gelijkstellen aan deze bovenlimiet, met het argument dat de boom geconstrueerd werd om de geobserveerde of waargenomen foutenratio te minimaliseren. Hierbij moeten we wel in het achterhoofd houden dat deze beschrijving niet volledig overeenkomt met de statistische opvattingen omtrent betrouwbaarheidsintervallen. Dus een blad dat N traininggevallen bevat met een voorspelde foutenratio van $U_{CF}(E, N)$ veroorzaakt $N \times U_{CF}(E, N)$ fouten. Het aantal voorspelde fouten geassocieerd aan een (sub)boom is dan de som van de voorspelde fouten van de eigen takken. De som van de voorspelde fouten in de bladeren, gedeeld door het aantal gevallen in de training set kan gebruikt worden als een schatting van de foutenratio van de gesnoeide boom met betrekking tot ongeziene gevallen.

[1] zie bijlage B voor meer uitleg over de binomiale verdeling

Bijlage B

De binomiale verdeling

De uitleg over de binomiale verdeling in deze bijlage werd overgenomen van `http://mmc.et.tudelft.nl/presan/node23.html`.

Wanneer we n onafhankelijke experimenten doen en elk experiment heeft kans p op 'succes' dan heeft de stochastische variabele X, die het totaal aantal 'successen' voorstelt, een binomiale verdeling met parameters n en p.

De kansen van de binomiale verdeling worden gegeven door:

$$p_i = \binom{n}{i} (1-p)^{n-i} p^i \qquad 0 \leq i \leq n \tag{9}$$

De factor p_i staat voor de i 'successen' die moeten optreden, de factor $(1-p)^{n-i}$ voor de voor de $n-i$ keer 'falen' en de binomiaal cofficint '*noveri*' of

$$\binom{n}{i} \tag{10}$$

staat voor het aantal mogelijkheden om i objecten uit een verzameling van n te kiezen (zonder onderscheid te maken in de volgorde).

De binomiaal coëfficiënt wordt als volgt berekend:

$$= \binom{n}{k} = \frac{n}{k(n-k)} = \frac{n}{1}\frac{n-1}{2}\cdots\frac{n-k+1}{k} \tag{11}$$

Voor de binomiale verdeling worden het gemiddelde en de variantie gegeven door:

$$\begin{aligned} \mu &= np \\ \sigma^2 &= np(1-p) \end{aligned} \tag{12}$$

Bijlage C

Data mining competitie

Data Mining Competition

In the run up to the conference a data mining competition takes place, which is sponsored by the "Deutsche Sparkassen- und Giroverband" (DSGV).

Prediction of the Liquidity Crisis of Companies

Companies getting into a liquidity crisis is one of the major problems banks have to face today. The goal of the competition is to predict a liquidity crisis based on (a subset of) 26 variables describing attributes of companies.

The data:

For the participation in the competition you need the file "data_company.zip". After extraction you will find training and test data in the files "data_company_train.txt" and "data_company_test.txt" with a semicolon separating the values.

The training data file has 20,001 lines:

a. The column names are given in the first line.
b. The following 20,000 lines contain the companies' data.

The training data file has 28 columns:

a. The first column is a unique identifier ("ID").
b. The second column ("Ereignisflag") encodes whether the company got into a liquidity crisis, where "1" indicates a crisis and "0" no crisis.
c. The following columns ("VAR01" to "VAR26") contain the values of the 26 variables that may explain the crisis. All 26 variables are metric.

The test data file has 10,001 lines and 27 columns:

a. The first column is a unique identifier ("ID").
b. The following columns ("VAR01" to "VAR26") contain the values of the 26 variables that may explain the crisis.
The test data set has no binary flag indicating the class label.

The task:

Build a model to predict the liquidity crisis for the companies.

Best model:

Each participant has to deliver a file containing a list with the IDs of the first 2,000 companies (of the 10,000 test data sets) with the highest measure (e.g. probability, membership) of liquidity crisis together with the predicted value of this measure, as well as a short report which describes the method used to obtain the classification result.
Submitted models will be ranked according to the number of correctly

classified companies having a liquidity crisis. If more than one model correctly classifies the same number of companies, the model using fewer variables wins the competition.

Awards:

For the best solutions the following awards are provided by the sponsor:

- 1st winner: 1,000 Euro
- 2nd winner: 500 Euro

Additionally, the winners are invited to present their results at the conference (the registration fees are waived and a subsidy for travel expenses will be provided).

Time schedule:

- Start of the competition: October 4, 2004
- Deadline for submission of solutions: January 15, 2005
- Announcement of winners: February 1, 2005
- Conference and presentation of results: March 9-11, 2005

Contact:

If you have any question concerning the competition, please contact Jens Strackeljan by email: jens.strackeljan@tu-clausthal.de

sponsored by

Bijlage D

Rapport data mining competitie

Report Data Mining Competition: Prediction of the Liquidity Crisis of Companies

Master 's thesis student: Sabrina Noblesse[1]
Promotor: Prof. dr. Koen Vanhoof[2]

[1] Limburgs Universitair Centrum, 3590 Diepenbeek, Belgium
[2] Departement BEDR/VEB, Limburgs Universitair Centrum
Universitaire Campus D, 3590 Diepenbeek, Belgium

Abstract. This report will give some additional information about the included file that contains a list with the IDs of the first 2000 companies together with the predicted value of liquidity crisis. The method used to obtain the classification result will also be described.

1 File contents: additional information

In this section, we will give some extra information about the data file `DataMiningCompetition.txt` that accompanies this report. This file contains 2001 lines. The first row contains the names of the different columns as shown below.

- *ID:* unique identifier
- *Confidence:* measure of Weka for the certainty by which the prediction of the classvalue is made
- *Ereignisflag:* predicted classvalue ("1" indicates a crisis and "0" indicates no crisis)
- *Sensitivity:* minimum distance between an instance and its projections
- *SensitivityConfidence:* product of the sensitivity of the instance and the confidence for prediction of the classvalue
- *MaxSensitivity:* maximum sensitivity that exists for the 10000 instances in the test data file

The next 2000 lines of the data file contain the values of these columns for the first 2000 companies selected by the classification method described in the next section.

2 Classification method

The method used to obtain the classification result is based on the Weka j48 (Witten I. and Frank E. (2000)) algorithm and the sensitivity algorithm

described in the paper "Sensitivity Analysis of the Result in Binary Decision Trees" by Isabelle Alvarez (2004).

First, the training data file was transformed into an ARFF-file. An ARFF, Attribute-Relation File Format, file is an ASCII text file that contains a list of instances together with their attributes and the values of those attributes for each instance . This file format was developed by the Machine Learning Project at the Department of Computer Science of The University of Waikato for use with the Weka machine learning software (Witten I. and Frank E. (2000)).

Then Weka j48 was used to grow a binary decision tree on the transformed training data file. From the resulting model, we could derive the different decision surfaces within the tree. A decision surface can be seen as the boundary between regions with different class labels. So, every leaf within the tree has its own decision surface. When this model was applied to the transformed test data file, Weka also provided a confidence value for each classified instance.

Once Weka had predicted the classvalue for each instance, we could apply the earlier mentioned sensitivity algorithm. This algorithm projects a given case onto every leaf that has a different classvalue than the classvalue predicted for this case. It then computes and ranks the distance between the case and its different projections.

We know that when the distance between a case and the decision surface of a leaf becomes smaller, the risk of a misclassified instance becomes more realistic. That is why we have multiplied the confidence provided by the Weka j48 algorithm for a specific case with the minimum of the distances provided by the sensitivity algorithm for the same case. Let us call this product the "SensitivityConfidence" for now. In this way we could base our classification on the confidence or certainty by which the prediction of the classvalue was made and on the greatest smallest distance between a case and its projections. The final step in selecting the 2000 first companies was ranking the different classified instances by descending order of SensitiviyConfidence.

References

ALVAREZ, I. (2004): Sensitivity Analysis of the Result in Binary Decision Trees. In: Proc. of the 15th European Conference on Machine Learning Vol 3201: *Lecture Notes in Artificial Intelligence*. Springer-Verslag, 51–62.

WITTEN I. and FRANK E. (2000): *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementation*. Morgan Kaufmann.

Bijlage E

Beslissingsboom data mining competitie

```
=== Run information ===

Scheme:       weka.classifiers.trees.J48 -C 0.25 -B -M 2
Relation:     liquidity
Instances:    20000
Attributes:   27
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree
------------------

VAR26 <= 0.0
|  VAR11 <= -1.17
|  |  VAR15 <= 14.38
|  |  |  VAR15 <= 8.9
|  |  |  |  VAR03 <= 0.0
|  |  |  |  |  VAR05 <= 384.0
|  |  |  |  |  |  VAR08 <= 1.27
|  |  |  |  |  |  |  VAR02 <= 64.32: 0 (26.0/6.0)
|  |  |  |  |  |  |  VAR02 > 64.32
|  |  |  |  |  |  |  |  VAR22 <= -86.6
|  |  |  |  |  |  |  |  |  VAR22 <= -213.92: 0 (5.44/1.56)
|  |  |  |  |  |  |  |  |  VAR22 > -213.92: 1 (9.12/1.94)
|  |  |  |  |  |  |  |  VAR22 > -86.6: 0 (7.6/0.98)
|  |  |  |  |  |  VAR08 > 1.27
|  |  |  |  |  |  |  VAR22 <= -132.99: 0 (2.56/0.19)
|  |  |  |  |  |  |  VAR22 > -132.99: 1 (5.28/0.19)
|  |  |  |  |  VAR05 > 384.0: 1 (5.0)
|  |  |  |  VAR03 > 0.0: 1 (124.0/43.0)
|  |  |  VAR15 > 8.9: 0 (640.0/149.0)
|  |  VAR15 > 14.38
|  |  |  VAR21 <= -314.65
|  |  |  |  VAR20 <= 2530.61
|  |  |  |  |  VAR03 <= 29.45
|  |  |  |  |  |  VAR09 <= 4.63
|  |  |  |  |  |  |  VAR13 <= -105.23: 1 (19.17/3.9)
|  |  |  |  |  |  |  VAR13 > -105.23
|  |  |  |  |  |  |  |  VAR04 <= -214.37
|  |  |  |  |  |  |  |  |  VAR06 <= -9733.33: 1 (19.73/1.27)
|  |  |  |  |  |  |  |  |  VAR06 > -9733.33
```

```
| | | | | | | | | | VAR21 <= -6207.01: 0 (4.12)
| | | | | | | | | | VAR21 > -6207.01: 1 (4.95/1.68)
| | | | | | | | VAR04 > -214.37
| | | | | | | | | VAR14 <= 53.28
| | | | | | | | | | VAR16 <= 2.44
| | | | | | | | | | | VAR13 <= -77.91: 1 (10.86/2.36)
| | | | | | | | | | | VAR13 > -77.91
| | | | | | | | | | | | VAR02 <= 32.16: 1 (3.36/0.87)
| | | | | | | | | | | | VAR02 > 32.16: 0 (12.09/4.86)
| | | | | | | | | | VAR16 > 2.44: 0 (4.06)
| | | | | | | | | VAR14 > 53.28: 0 (49.43/16.91)
| | | | | | VAR09 > 4.63: 1 (16.23/1.0)
| | | | | VAR03 > 29.45: 1 (64.0/12.0)
| | | | VAR20 > 2530.61: 0 (12.0/1.0)
| | | VAR21 > -314.65
| | | | VAR12 <= 55.93: 0 (18.0)
| | | | VAR12 > 55.93
| | | | | VAR09 <= 0.93: 1 (4.67/0.67)
| | | | | VAR09 > 0.93: 0 (2.33)
| VAR11 > -1.17
| | VAR03 <= 0.0
| | | VAR23 <= 21.05
| | | | VAR15 <= 14.38: 0 (13316.69/466.47)
| | | | VAR15 > 14.38
| | | | | VAR20 <= -1008.16
| | | | | | VAR13 <= -87.21
| | | | | | | VAR13 <= -94.77: 0 (6.95/1.01)
| | | | | | | VAR13 > -94.77
| | | | | | | | VAR17 <= 6.82: 1 (32.76/10.89)
| | | | | | | | VAR17 > 6.82: 0 (5.94/0.99)
| | | | | | VAR13 > -87.21
| | | | | | | VAR17 <= 0.0
| | | | | | | | VAR22 <= -31.96
| | | | | | | | | VAR20 <= -1932.65: 0 (294.83/50.26)
| | | | | | | | | VAR20 > -1932.65
| | | | | | | | | | VAR20 <= -1225.17
| | | | | | | | | | | VAR20 <= -1695.92: 1 (3.96)
| | | | | | | | | | | VAR20 > -1695.92: 0 (13.86/2.97)
| | | | | | | | | | VAR20 > -1225.17: 1 (4.95)
| | | | | | | | VAR22 > -31.96
| | | | | | | | | VAR06 <= -150419.26: 1 (2.97/0.99)
| | | | | | | | | VAR06 > -150419.26: 0 (55.59/0.14)
| | | | | | | VAR17 > 0.0
```

```
| | | | | | | | VAR16 <= 2.44
| | | | | | | | | VAR24 <= 2.42
| | | | | | | | | | VAR01 <= -6335.52: 1 (2.97/0.99)
| | | | | | | | | | VAR01 > -6335.52: 0 (10.89)
| | | | | | | | | VAR24 > 2.42: 1 (4.95/0.99)
| | | | | | | | VAR16 > 2.44: 1 (2.97)
| | | | | VAR20 > -1008.16: 0 (2356.59/103.97)
| | | VAR23 > 21.05
| | | | VAR24 <= 0.0
| | | | | VAR13 <= -61.63: 1 (48.7/8.84)
| | | | | VAR13 > -61.63: 0 (5.94/0.49)
| | | | VAR24 > 0.0: 0 (105.5/21.18)
| | VAR03 > 0.0
| | | VAR15 <= 14.38
| | | | VAR15 <= 13.7
| | | | | VAR23 <= 18.05
| | | | | | VAR03 <= 58.9
| | | | | | | VAR16 <= 0.0: 0 (131.53/28.91)
| | | | | | | VAR16 > 0.0
| | | | | | | | VAR07 <= -621.39: 1 (5.47)
| | | | | | | | VAR07 > -621.39
| | | | | | | | | VAR16 <= 2.44
| | | | | | | | | | VAR07 <= -95.95: 0 (34.72/4.47)
| | | | | | | | | | VAR07 > -95.95
| | | | | | | | | | | VAR24 <= 0.81
| | | | | | | | | | | | VAR05 <= 217.0: 1 (5.36)
| | | | | | | | | | | | VAR05 > 217.0
| | | | | | | | | | | | | VAR15 <= 10.96: 0 (5.0)
| | | | | | | | | | | | | VAR15 > 10.96: 1 (3.89/1.0)
| | | | | | | | | | | VAR24 > 0.81: 0 (39.12/15.51)
| | | | | | | | | VAR16 > 2.44
| | | | | | | | | | VAR15 <= 7.53: 1 (8.25/1.79)
| | | | | | | | | | VAR15 > 7.53: 0 (3.89)
| | | | | | VAR03 > 58.9: 1 (35.23/14.36)
| | | | | VAR23 > 18.05: 1 (32.53/13.47)
| | | | VAR15 > 13.7
| | | | | VAR20 <= -82632.65
| | | | | | VAR23 <= 14.29
| | | | | | | VAR03 <= 29.45: 0 (24.09/2.85)
| | | | | | | VAR03 > 29.45
| | | | | | | | VAR07 <= -173.99: 0 (8.24/1.24)
| | | | | | | | VAR07 > -173.99: 1 (7.85/2.0)
| | | | | | VAR23 > 14.29
```

```
| | | | | | | VAR14 <= 53.28: 1 (12.67/1.76)
| | | | | | | VAR14 > 53.28
| | | | | | | | VAR23 <= 15.79: 0 (2.21/0.21)
| | | | | | | | VAR23 > 15.79
| | | | | | | | | VAR07 <= -132.95: 0 (4.03/1.03)
| | | | | | | | | VAR07 > -132.95: 1 (5.91)
| | | | | VAR20 > -82632.65
| | | | | | VAR07 <= -301.73
| | | | | | | VAR07 <= -1313.87: 1 (8.05/1.03)
| | | | | | | VAR07 > -1313.87
| | | | | | | | VAR04 <= -507.47: 0 (33.17/4.08)
| | | | | | | | VAR04 > -507.47
| | | | | | | | | VAR14 <= 53.28
| | | | | | | | | | VAR08 <= 2.55
| | | | | | | | | | | VAR09 <= 0.0
| | | | | | | | | | | | VAR13 <= -94.19: 0 (3.06/1.06)
| | | | | | | | | | | | VAR13 > -94.19: 1 (6.03)
| | | | | | | | | | | VAR09 > 0.0: 0 (9.16/2.0)
| | | | | | | | | | VAR08 > 2.55: 1 (5.0)
| | | | | | | | | VAR14 > 53.28
| | | | | | | | | | VAR16 <= 0.0: 0 (4.0)
| | | | | | | | | | VAR16 > 0.0
| | | | | | | | | | | VAR05 <= 270.0: 1 (3.0)
| | | | | | | | | | | VAR05 > 270.0: 0 (5.0)
| | | | | VAR07 > -301.73
| | | | | | | VAR23 <= 18.05: 0 (714.06/78.58)
| | | | | | | VAR23 > 18.05
| | | | | | | | VAR23 <= 21.05: 0 (41.93/4.32)
| | | | | | | | VAR23 > 21.05
| | | | | | | | | VAR08 <= 0.0: 0 (5.87/1.68)
| | | | | | | | | VAR08 > 0.0: 1 (5.69/0.45)
| | | VAR15 > 14.38
| | | | VAR03 <= 58.9
| | | | | VAR21 <= -20641.4
| | | | | | VAR02 <= 64.32: 1 (4.0)
| | | | | | VAR02 > 64.32
| | | | | | | VAR06 <= -21698.52
| | | | | | | VAR06 <= -24338.52
| | | | | | | | VAR07 <= -207.51: 1 (9.0)
| | | | | | | | VAR07 > -207.51
| | | | | | | | | VAR06 <= -298765.93: 1 (3.0)
| | | | | | | | | VAR06 > -298765.93
| | | | | | | | | | VAR21 <= -56547.77: 0 (8.0)
```

```
| | | | | | | | | | VAR21 > -56547.77
| | | | | | | | | | | VAR04 <= -105.17: 0 (2.0)
| | | | | | | | | | | VAR04 > -105.17: 1 (7.0)
| | | | | | | VAR06 > -24338.52: 0 (7.0)
| | | | | | | VAR06 > -21698.52: 1 (7.0)
| | | | | VAR21 > -20641.4
| | | | | | VAR03 <= 29.45
| | | | | | | VAR14 <= 27.01
| | | | | | | | VAR17 <= 0.0: 0 (51.0/6.0)
| | | | | | | | VAR17 > 0.0
| | | | | | | | | VAR04 <= 90.23: 0 (2.0)
| | | | | | | | | VAR04 > 90.23: 1 (2.0)
| | | | | | | VAR14 > 27.01
| | | | | | | | VAR12 <= 63.56
| | | | | | | | | VAR19 <= 56.2: 0 (40.21/8.92)
| | | | | | | | | VAR19 > 56.2: 1 (3.89/0.41)
| | | | | | | | VAR12 > 63.56: 1 (4.9/0.3)
| | | | | | VAR03 > 29.45
| | | | | | | VAR09 <= 0.93
| | | | | | | | VAR17 <= 8.33
| | | | | | | | | VAR02 <= 32.16
| | | | | | | | | | VAR18 <= 69.63: 1 (4.4)
| | | | | | | | | | VAR18 > 69.63: 0 (6.6/1.6)
| | | | | | | | | VAR02 > 32.16: 0 (36.0/7.0)
| | | | | | | | VAR17 > 8.33: 1 (3.0)
| | | | | | | VAR09 > 0.93
| | | | | | | | VAR01 <= -5602.73: 0 (2.0)
| | | | | | | | VAR01 > -5602.73: 1 (13.0/1.0)
| | | | VAR03 > 58.9: 1 (46.0/11.0)
VAR26 > 0.0
| VAR03 <= 0.0
| | VAR08 <= 1.27
| | | VAR08 <= 0.0: 0 (39.0/2.0)
| | | VAR08 > 0.0
| | | | VAR07 <= -23.12
| | | | | VAR06 <= -7479.26: 0 (7.0/1.0)
| | | | | VAR06 > -7479.26: 1 (3.4)
| | | | VAR07 > -23.12: 0 (15.6/0.6)
| | VAR08 > 1.27: 1 (7.0/2.0)
| VAR03 > 0.0
| | VAR11 <= -15.79: 1 (869.0/231.0)
| | VAR11 > -15.79
| | | VAR06 <= -291.85
```

```
| | | | VAR15 <= 1.37
| | | | | VAR01 <= 12449.73
| | | | | | VAR15 <= 0.0
| | | | | | | VAR14 <= 53.28
| | | | | | | | VAR06 <= -29345.19: 0 (2.0)
| | | | | | | | VAR06 > -29345.19: 1 (3.0)
| | | | | | | VAR14 > 53.28: 0 (2.0)
| | | | | | VAR15 > 0.0
| | | | | | | VAR05 <= 255.0
| | | | | | | | VAR26 <= 34.24
| | | | | | | | | VAR14 <= 53.28: 1 (2.0)
| | | | | | | | | VAR14 > 53.28
| | | | | | | | | | VAR04 <= -271.26: 1 (5.0)
| | | | | | | | | | VAR04 > -271.26
| | | | | | | | | | | VAR06 <= -10975.56: 0 (8.0)
| | | | | | | | | | | VAR06 > -10975.56
| | | | | | | | | | | | VAR24 <= 0.0: 1 (4.0)
| | | | | | | | | | | | VAR24 > 0.0
| | | | | | | | | | | | | VAR01 <= 47.54: 0 (2.0)
| | | | | | | | | | | | | VAR01 > 47.54: 1 (4.0/1.0)
| | | | | | | | VAR26 > 34.24
| | | | | | | | | VAR08 <= 2.55
| | | | | | | | | | VAR07 <= -135.84: 0 (10.0)
| | | | | | | | | | VAR07 > -135.84: 1 (5.0/1.0)
| | | | | | | | | VAR08 > 2.55: 1 (2.0)
| | | | | | | VAR05 > 255.0: 1 (32.0/5.0)
| | | | | VAR01 > 12449.73: 0 (6.0)
| | | | VAR15 > 1.37
| | | | | VAR18 <= 48.89: 0 (42.0/14.0)
| | | | | VAR18 > 48.89
| | | | | | VAR02 <= 64.32: 1 (6.5/1.5)
| | | | | | VAR02 > 64.32: 0 (35.5/14.0)
| | | VAR06 > -291.85: 0 (65.0/11.0)

Number of Leaves  :  118
```

Bijlage F

Mogelijke waarden voor de sensitiviteit berekend met behulp van de Standaard metriek en het cumulatief aantal gevallen

Aantal gevallen	Sensitiviteit
4772	0,000000000000
5383	0,000001000000
5386	0,000002000000
5392	0,000003000000
5394	0,000004000000
5398	0,000005000000
5399	0,000006000000
5403	0,000007000000
5404	0,000008000000
5416	0,000009000000
5429	0,000010000000
5436	0,000011000000
9437	0,000012000000
9441	0,000014000000
9442	0,000015000000
9443	0,000018000001
9444	0,000018999999
9445	0,000019999999
9447	0,000021000000
9451	0,000022000000
9453	0,000024000001
9455	0,000029000001
9456	0,000029999999
9458	0,000034000001
9459	0,000037000002
9460	0,000038999999
9461	0,000039999999
9462	0,000042000000
9468	0,000043000000
9469	0,000048000002
9470	0,000048999998
9471	0,000053000000
9473	0,000054000000
9474	0,000056000001
9475	0,000057000001
9476	0,000062999999

Aantal gevallen	Sensitiviteit
9478	0,000065000000
9479	0,000073000003
9480	0,000077999997
9481	0,000080999998
9483	0,000086000000
9484	0,000088000001
9485	0,000092000002
9486	0,000100999998
9487	0,000108000000
9488	0,000115000003
9489	0,000117000003
9490	0,000121999998
9491	0,000123000005
9492	0,000123999998
9494	0,000125999999
9495	0,000127000007
9497	0,000128000000
9499	0,000132000001
9500	0,000138999996
9501	0,000144999998
9502	0,000149000000
9503	0,000155000002
9504	0,000186999998
9505	0,000188000005
9506	0,000195999994
9507	0,000201000003
9508	0,000203000003
9509	0,000209000005
9511	0,000226999997
9512	0,000232000006
9513	0,000236000007
9514	0,000237999993
9515	0,000243999995
9516	0,000247999997
9517	0,000249000004
9518	0,000254999992

Aantal gevallen	Sensitiviteit
9519	0,000265000010
9520	0,000274999999
9521	0,000279000000
9522	0,000280999986
9523	0,000284999987
9524	0,000291000004
9525	0,000292000012
9526	0,000308999995
9528	0,000321000000
9529	0,000322000007
9531	0,000323999993
9532	0,000325000001
9533	0,000329000002
9535	0,000337000005
9536	0,000339999999
9537	0,000345000008
9538	0,000348000001
9539	0,000353000010
9540	0,000357000012
9541	0,000357999990
9542	0,000387999986
9543	0,000392999995
9544	0,000394000002
9545	0,000395999989
9546	0,000396999996
9547	0,000399000011
9548	0,000400999998
9549	0,000410999986
9550	0,000445000012
9551	0,000475000008
9553	0,000483999989
9554	0,000488999998
9555	0,000494999986
9557	0,000499000016
9558	0,000509999983
9559	0,000520000001

Aantal gevallen	Sensitiviteit
9560	0,000526999997
9561	0,000531000027
9562	0,000532999984
9563	0,000534999999
9564	0,000545000017
9565	0,000553000020
9566	0,000574000005
9567	0,000593999983
9568	0,000598000013
9569	0,000612000003
9570	0,000621000014
9571	0,000624999986
9572	0,000625999994
9573	0,000637000019
9575	0,000645000022
9576	0,000652000017
9577	0,000654999982
9578	0,000661000027
9579	0,000667000015
9580	0,000668999972
9581	0,000671999995
9582	0,000684999977
9583	0,000685999985
9584	0,000693999988
9585	0,000699999975
9587	0,000708999985
9588	0,000711000001
9589	0,000731999986
9590	0,000734000001
9591	0,000739999989
9592	0,000743000011
9593	0,000745999976
9594	0,000746999984
9595	0,000753999979
9596	0,000762999989
9597	0,000773000007

Aantal gevallen	Sensitiviteit
9598	0,000796000008
9599	0,000801999995
9600	0,000821000023
9601	0,000821999973
9602	0,000837999978
9603	0,000853999984
9604	0,000865000009
9605	0,000867999974
9606	0,000877999992
9607	0,000882000022
9608	0,000890999974
9609	0,000898000028
9610	0,000904000015
9611	0,000905000023
9612	0,000933000003
9613	0,000941000006
9614	0,000942000013
9615	0,000953999988
9616	0,000954999996
9617	0,000969999994
9618	0,000976999989
9619	0,000985000050
9620	0,000990999979
9621	0,000992999994
9622	0,000996000017
9623	0,000997000025
9624	0,001002999954
9625	0,001011000015
9626	0,001018999959
9627	0,001037999988
9628	0,001038999995
9629	0,001043000026
9630	0,001050999970
9631	0,001052999985
9632	0,001056000008
9633	0,001065999968

Aantal gevallen	Sensitiviteit
9634	0,001088000019
9636	0,001097999979
9637	0,001121000038
9638	0,001151000033
9639	0,001157999970
9640	0,001162000000
9641	0,001163000008
9642	0,001165000023
9643	0,001182000036
9644	0,001246999949
9645	0,001256000018
9646	0,001323999953
9647	0,001324999961
9648	0,001364000025
9650	0,001396000036
9651	0,001429000054
9652	0,001444000052
9653	0,001445999951
9654	0,001453000004
9655	0,001507999958
9657	0,001508999965
9658	0,001516000018
9659	0,001524999971
9660	0,001527999993
9661	0,001534000039
9662	0,001545000006
9663	0,001565000042
9664	0,001567000058
9665	0,001592000015
9666	0,001596000046
9667	0,001692999969
9668	0,001705999952
9669	0,001715000020
9670	0,001731000026
9671	0,001743000001
9672	0,001752999960

Aantal gevallen	Sensitiviteit
9673	0,001767999958
9674	0,001797999954
9675	0,001815999974
9676	0,001828999957
9677	0,001881000004
9678	0,001904999954
9679	0,001919999951
9680	0,001922999974
9681	0,001949999947
9682	0,001990000019
9683	0,001997000072
9684	0,001998000080
9685	0,002011999954
9686	0,002013999969
9687	0,002021000022
9688	0,002046999987
9689	0,002077999990
9690	0,002094999887
9691	0,002105999971
9692	0,002115000039
9693	0,002128999913
9694	0,002182000084
9695	0,002199999988
9696	0,002200999996
9697	0,002222999930
9698	0,002276000101
9699	0,002326000016
9700	0,002329000039
9701	0,002350999974
9702	0,002358000027
9703	0,002370999893
9704	0,002378999954
9705	0,002381999977
9706	0,002409999957
9707	0,002417000011
9708	0,002422000049

Aantal gevallen	Sensitiviteit
9709	0,002430000110
9710	0,002499999944
9711	0,002501999959
9713	0,002576000057
9714	0,002582000103
9715	0,002648999915
9716	0,002659999998
9717	0,002663000021
9718	0,002677999903
9719	0,002700000070
9720	0,002708999906
9721	0,002710999921
9722	0,002732000081
9723	0,002750999993
9724	0,002824000083
9725	0,002829999896
9726	0,002830999903
9727	0,002842999995
9728	0,002846000018
9729	0,002898999956
9730	0,002920999890
9731	0,002927999943
9732	0,002936000004
9733	0,002963999985
9734	0,002964999992
9735	0,002973000053
9736	0,002985999919
9737	0,003003000049
9738	0,003025999991
9739	0,003028000006
9740	0,003108999925
9741	0,003130000085
9742	0,003170999931
9743	0,003175999969
9744	0,003289999906
9745	0,003306000028

Aantal gevallen	Sensitiviteit
9746	0,003445999930
9747	0,003466000082
9748	0,003476999933
9749	0,003635999979
9750	0,003642000025
9751	0,003673000028
9752	0,003706000047
9753	0,003838999895
9754	0,003871999914
9755	0,003880999982
9756	0,003928999882
9757	0,004009999800
9758	0,004052999895
9759	0,004135000054
9760	0,004158999771
9761	0,004195000045
9762	0,004213999957
9763	0,004242000170
9764	0,004255999811
9765	0,004259999841
9767	0,004277000204
9768	0,004385000095
9769	0,004397999961
9770	0,004445999861
9771	0,004490000196
9772	0,004492999986
9773	0,004509999882
9774	0,004575999919
9775	0,004592999816
9776	0,004761999939
9777	0,004825999960
9778	0,004856999964
9779	0,004993000068
9780	0,005123000126
9781	0,005175999831
9782	0,005202000029

Aantal gevallen	Sensitiviteit
9783	0,005381000228
9784	0,005413999781
9785	0,005423000082
9786	0,005433000159
9787	0,005537999794
9788	0,005667999852
9789	0,005745999981
9790	0,005789999850
9791	0,005865999963
9792	0,006010999903
9793	0,006095999852
9794	0,006109999958
9795	0,006111000199
9796	0,006390000228
9797	0,006595000159
9798	0,006794000044
9799	0,006891000085
9800	0,006936000194
9801	0,006949999835
9802	0,006957999896
9803	0,007251999807
9804	0,007333000191
9805	0,007484000176
9806	0,007629999891
9807	0,007739000022
9808	0,007774000056
9809	0,007865999825
9810	0,008086999878
9811	0,008181000128
9812	0,008191999979
9813	0,008228000253
9814	0,008276999928
9815	0,008290000260
9816	0,008338999934
9817	0,008476000279
9818	0,008553000167

Aantal gevallen	Sensitiviteit
9819	0,008716999553
9820	0,008790999651
9821	0,008911999874
9822	0,009046000428
9823	0,009293000214
9824	0,009359000251
9825	0,009480999783
9826	0,009488999844
9827	0,009672000073
9828	0,009692000225
9829	0,009740999900
9830	0,009859000333
9831	0,010320000350
9832	0,010455000214
9833	0,010459000245
9834	0,010499999858
9835	0,010580999777
9836	0,010591000319
9837	0,010629000142
9838	0,010658999905
9839	0,010755999945
9840	0,011014999822
9841	0,011187000200
9842	0,011540999636
9843	0,011591999792
9844	0,011656000279
9845	0,011827999726
9846	0,012082999572
9848	0,013395000249
9849	0,013542000204
9850	0,013762000017
9851	0,013838999905
9852	0,013872999698
9853	0,013927999884
9854	0,013960000128
9855	0,014050999656

Aantal gevallen	Sensitiviteit
9856	0,014607000165
9857	0,014825999737
9858	0,015089999884
9859	0,015115999617
9860	0,015197999775
9861	0,015394999646
9862	0,015522000380
9863	0,015607000329
9864	0,015690000728
9865	0,015800999478
9866	0,015984000638
9867	0,016000000760
9868	0,016405999660
9869	0,016419000924
9870	0,016656000167
9871	0,016679000109
9872	0,016697999090
9873	0,016956999898
9874	0,016964999959
9875	0,016993999481
9876	0,017113000154
9877	0,017215000466
9878	0,017563000321
9879	0,017595000565
9880	0,017687000334
9881	0,018303999677
9882	0,018517000601
9883	0,018688000739
9884	0,019063999876
9885	0,019129000604
9886	0,019230000675
9887	0,019666999578
9888	0,019864000380
9889	0,019904000685
9890	0,019946999848
9891	0,020073000342

Aantal gevallen	Sensitiviteit
9892	0,020222999156
9893	0,020291000605
9894	0,020604999736
9895	0,021051000804
9896	0,021743999794
9897	0,022653000429
9898	0,022716000676
9899	0,023109000176
9900	0,023552000523
9901	0,024171000347
9902	0,024435000494
9903	0,025392999873
9904	0,025652000681
9905	0,025894999504
9906	0,026189999655
9907	0,026350000873
9908	0,026411000639
9909	0,026645999402
9910	0,027044000104
9911	0,027496999130
9912	0,028707999736
9913	0,029030000791
9915	0,029231999069
9916	0,029795000330
9917	0,029834000394
9918	0,030128000304
9919	0,030263999477
9920	0,030556999147
9921	0,030801000074
9922	0,030822999775
9923	0,031120000407
9924	0,032400000840
9925	0,032529998571
9926	0,034956000745
9927	0,035176999867
9928	0,035286001861

Aantal gevallen	Sensitiviteit
9929	0,035704001784
9930	0,036042001098
9931	0,037484999746
9932	0,037574000657
9933	0,037627998739
9934	0,039136998355
9935	0,039211001247
9936	0,039473999292
9937	0,039666999131
9938	0,041678998619
9939	0,043014001101
9940	0,043067999184
9941	0,044879999012
9942	0,045850001276
9943	0,046326000243
9944	0,047545999289
9945	0,047800000757
9946	0,048044998199
9947	0,049424998462
9948	0,051382001489
9949	0,054490000010
9950	0,054786998779
9951	0,055034000427
9952	0,057824000716
9953	0,058713998646
9954	0,058961998671
9955	0,067174002528
9956	0,068346001208
9957	0,068479001522
9958	0,070183999836
9959	0,076467998326
9960	0,077454000711
9961	0,078111998737
9962	0,083320997655
9963	0,083664000034
9964	0,085101000965

Aantal gevallen	Sensitiviteit
9965	0,088464997709
9966	0,095270000398
9967	0,099632002413
9968	0,099753998220
9969	0,100004002452
9970	0,102702997625
9971	0,107643999159
9972	0,109205998480
9973	0,112157002091
9974	0,113294996321
9975	0,114568002522
9976	0,137564003468
9977	0,138380005956
9978	0,138907998800
9979	0,173518002033
9980	0,218951001763
9981	0,223370999098
9982	0,234657004476
9983	0,242987006903
9984	0,312766999006
9985	0,329376995564
9990	0,352084010839
9991	0,354647994041
9992	0,362109988928
9993	0,428824990988
9994	0,574212014675
9995	0,628647029400
9996	0,949575006962
9997	0,953732013702
9998	0,961419999599
9999	0,985988020897
10000	1,375939965248

Bijlage G

Resultaten gecorrigeerde waarschijnlijkheidsschattingen Standaard metriek

Grenswaarde Sensitiviteit (s0)	c1	c2	juist klasse 0 top 2000	juist klasse 1 top 2000	juist totaal top 2000	fout voorspeld als klasse 0	fout voorspeld als klasse 1	fout voorspeld totaal
0	0	0	1904	18	1922	54	24	78
		0,05	1930	12	1942	44	14	58
		0,1	1930	12	1942	44	14	58
		0,15	1930	12	1942	44	14	58
		0,2	1930	12	1942	44	14	58
		0,25	1930	12	1942	44	14	58
		0,3	1930	12	1942	44	14	58
		0,35	1930	12	1942	44	14	58
		0,4	1930	12	1942	44	14	58
		0,45	1930	12	1942	44	14	58
		0,5	1930	12	1942	44	14	58
0,000001000000	0	0	1904	18	1922	54	24	78
		0,05	1944	9	1953	36	11	47
		0,1	1944	9	1953	36	11	47
		0,15	1944	9	1953	36	11	47
		0,2	1944	9	1953	36	11	47
		0,25	1944	9	1953	36	11	47
		0,3	1944	9	1953	36	11	47
		0,35	1944	9	1953	36	11	47
		0,4	1944	9	1953	36	11	47
		0,45	1944	9	1953	36	11	47
		0,5	1944	9	1953	36	11	47
	0,05	0	1944	9	1953	36	11	47
		0,05	1944	9	1953	36	11	47
		0,1	1944	9	1953	36	11	47
		0,15	1944	9	1953	36	11	47
		0,2	1944	9	1953	36	11	47
		0,25	1944	9	1953	36	11	47
		0,3	1944	9	1953	36	11	47
		0,35	1944	9	1953	36	11	47
		0,4	1944	9	1953	36	11	47
		0,45	1944	9	1953	36	11	47
		0,5	1944	9	1953	36	11	47

Grenswaarde Sensitiviteit (s0)	c1	c2	juist klasse 0 top 2000	juist klasse 1 top 2000	juist totaal top 2000	fout voorspeld als klasse 0	fout voorspeld als klasse 1	fout voorspeld totaal
0,000001000000	0,1	0	1944	9	1953	36	11	47
		0,05	1944	9	1953	36	11	47
		0,1	1944	9	1953	36	11	47
		0,15	1944	9	1953	36	11	47
		0,2	1944	9	1953	36	11	47
		0,25	1944	9	1953	36	11	47
		0,3	1944	9	1953	36	11	47
		0,35	1944	9	1953	36	11	47
		0,4	1944	9	1953	36	11	47
		0,45	1944	9	1953	36	11	47
		0,5	1944	9	1953	36	11	47
	0,15	0	1944	9	1953	36	11	47
		0,05	1944	9	1953	36	11	47
		0,1	1944	9	1953	36	11	47
		0,15	1944	9	1953	36	11	47
		0,2	1944	9	1953	36	11	47
		0,25	1944	9	1953	36	11	47
		0,3	1944	9	1953	36	11	47
		0,35	1944	9	1953	36	11	47
		0,4	1944	9	1953	36	11	47
		0,45	1944	9	1953	36	11	47
		0,5	1944	9	1953	36	11	47
	0,2	0	1944	9	1953	36	11	47
		0,05	1944	9	1953	36	11	47
		0,1	1944	9	1953	36	11	47
		0,15	1944	9	1953	36	11	47
		0,2	1944	9	1953	36	11	47
		0,25	1944	9	1953	36	11	47
		0,3	1944	9	1953	36	11	47
		0,35	1944	9	1953	36	11	47
		0,4	1944	9	1953	36	11	47
		0,45	1944	9	1953	36	11	47
		0,5	1944	9	1953	36	11	47

Grenswaarde Sensitiviteit (s0)	c1	c2	juist klasse 0 top 2000	juist klasse 1 top 2000	juist totaal top 2000	fout voorspeld als klasse 0	fout voorspeld als klasse 1	fout voorspeld totaal
0,000001000000	0,25	0	1944	9	1953	36	11	47
		0,05	1944	9	1953	36	11	47
		0,1	1944	9	1953	36	11	47
		0,15	1944	9	1953	36	11	47
		0,2	1944	9	1953	36	11	47
		0,25	1944	9	1953	36	11	47
		0,3	1944	9	1953	36	11	47
		0,35	1944	9	1953	36	11	47
		0,4	1944	9	1953	36	11	47
		0,45	1944	9	1953	36	11	47
		0,5	1944	9	1953	36	11	47
	0,3	0	1944	9	1953	36	11	47
		0,05	1944	9	1953	36	11	47
		0,1	1944	9	1953	36	11	47
		0,15	1944	9	1953	36	11	47
		0,2	1944	9	1953	36	11	47
		0,25	1944	9	1953	36	11	47
		0,3	1944	9	1953	36	11	47
		0,35	1944	9	1953	36	11	47
		0,4	1944	9	1953	36	11	47
		0,45	1944	9	1953	36	11	47
		0,5	1944	9	1953	36	11	47
	0,35	0	1944	9	1953	36	11	47
		0,05	1944	9	1953	36	11	47
		0,1	1944	9	1953	36	11	47
		0,15	1944	9	1953	36	11	47
		0,2	1944	9	1953	36	11	47
		0,25	1944	9	1953	36	11	47
		0,3	1944	9	1953	36	11	47
		0,35	1944	9	1953	36	11	47
		0,4	1944	9	1953	36	11	47
		0,45	1944	9	1953	36	11	47
		0,5	1944	9	1953	36	11	47

Grenswaarde Sensitiviteit (s0)	c1	c2	juist klasse 0 top 2000	juist klasse 1 top 2000	juist totaal top 2000	fout voorspeld als klasse 0	fout voorspeld als klasse 1	fout voorspeld totaal
0,000001000000	0,1	0	1944	9	1953	36	11	47
		0,05	1944	9	1953	36	11	47
		0,1	1944	9	1953	36	11	47
		0,15	1944	9	1953	36	11	47
		0,2	1944	9	1953	36	11	47
		0,25	1944	9	1953	36	11	47
		0,3	1944	9	1953	36	11	47
		0,35	1944	9	1953	36	11	47
		0,4	1944	9	1953	36	11	47
		0,45	1944	9	1953	36	11	47
		0,5	1944	9	1953	36	11	47
	0,15	0	1944	9	1953	36	11	47
		0,05	1944	9	1953	36	11	47
		0,1	1944	9	1953	36	11	47
		0,15	1944	9	1953	36	11	47
		0,2	1944	9	1953	36	11	47
		0,25	1944	9	1953	36	11	47
		0,3	1944	9	1953	36	11	47
		0,35	1944	9	1953	36	11	47
		0,4	1944	9	1953	36	11	47
		0,45	1944	9	1953	36	11	47
		0,5	1944	9	1953	36	11	47
	0,2	0	1944	9	1953	36	11	47
		0,05	1944	9	1953	36	11	47
		0,1	1944	9	1953	36	11	47
		0,15	1944	9	1953	36	11	47
		0,2	1944	9	1953	36	11	47
		0,25	1944	9	1953	36	11	47
		0,3	1944	9	1953	36	11	47
		0,35	1944	9	1953	36	11	47
		0,4	1944	9	1953	36	11	47
		0,45	1944	9	1953	36	11	47
		0,5	1944	9	1953	36	11	47

Grenswaarde Sensitiviteit (s0)	c1	c2	juist klasse 0 top 2000	juist klasse 1 top 2000	juist totaal top 2000	fout voorspeld als klasse 0	fout voorspeld als klasse 1	fout voorspeld totaal
0,000011	0	0	1904	18	1922	54	24	78
		0,05	1944	9	1953	36	11	47
		0,1	1944	9	1953	36	11	47
		0,15	1944	9	1953	36	11	47
		0,2	1944	9	1953	36	11	47
		0,25	1944	9	1953	36	11	47
		0,3	1944	9	1953	36	11	47
		0,35	1944	9	1953	36	11	47
		0,4	1944	9	1953	36	11	47
		0,45	1944	9	1953	36	11	47
		0,5	1944	9	1953	36	11	47
	0,05	0	1944	9	1953	36	11	47
		0,05	1944	9	1953	36	11	47
		0,1	1944	9	1953	36	11	47
		0,15	1944	9	1953	36	11	47
		0,2	1944	9	1953	36	11	47
		0,25	1944	9	1953	36	11	47
		0,3	1944	9	1953	36	11	47
		0,35	1944	9	1953	36	11	47
		0,4	1944	9	1953	36	11	47
		0,45	1944	9	1953	36	11	47
		0,5	1944	9	1953	36	11	47
	0,1	0	1944	9	1953	36	11	47
		0,05	1944	9	1953	36	11	47
		0,1	1944	9	1953	36	11	47
		0,15	1944	9	1953	36	11	47
		0,2	1944	9	1953	36	11	47
		0,25	1944	9	1953	36	11	47
		0,3	1944	9	1953	36	11	47
		0,35	1944	9	1953	36	11	47
		0,4	1944	9	1953	36	11	47
		0,45	1944	9	1953	36	11	47
		0,5	1944	9	1953	36	11	47

Grenswaarde Sensitiviteit (s0)	c1	c2	juist klasse 0 top 2000	juist klasse 1 top 2000	juist totaal top 2000	fout voorspeld als klasse 0	fout voorspeld als klasse 1	fout voorspeld totaal
0,000011	0,15	0	1944	9	1953	36	11	47
		0,05	1944	9	1953	36	11	47
		0,1	1944	9	1953	36	11	47
		0,15	1944	9	1953	36	11	47
		0,2	1944	9	1953	36	11	47
		0,25	1944	9	1953	36	11	47
		0,3	1944	9	1953	36	11	47
		0,35	1944	9	1953	36	11	47
		0,4	1944	9	1953	36	11	47
		0,45	1944	9	1953	36	11	47
		0,5	1944	9	1953	36	11	47
	0,2	0	1944	9	1953	36	11	47
		0,05	1944	9	1953	36	11	47
		0,1	1944	9	1953	36	11	47
		0,15	1944	9	1953	36	11	47
		0,2	1944	9	1953	36	11	47
		0,25	1944	9	1953	36	11	47
		0,3	1944	9	1953	36	11	47
		0,35	1944	9	1953	36	11	47
		0,4	1944	9	1953	36	11	47
		0,45	1944	9	1953	36	11	47
		0,5	1944	9	1953	36	11	47
	0,25	0	1944	9	1953	36	11	47
		0,05	1944	9	1953	36	11	47
		0,1	1944	9	1953	36	11	47
		0,15	1944	9	1953	36	11	47
		0,2	1944	9	1953	36	11	47
		0,25	1944	9	1953	36	11	47
		0,3	1944	9	1953	36	11	47
		0,35	1944	9	1953	36	11	47
		0,4	1944	9	1953	36	11	47
		0,45	1944	9	1953	36	11	47
		0,5	1944	9	1953	36	11	47

Grenswaarde Sensitiviteit (s0)	c1	c2	juist klasse 0 top 2000	juist klasse 1 top 2000	juist totaal top 2000	fout voorspeld als klasse 0	fout voorspeld als klasse 1	fout voorspeld totaal
0,000011	0,3	0	1944	9	1953	36	11	47
		0,05	1944	9	1953	36	11	47
		0,1	1944	9	1953	36	11	47
		0,15	1944	9	1953	36	11	47
		0,2	1944	9	1953	36	11	47
		0,25	1944	9	1953	36	11	47
		0,3	1944	9	1953	36	11	47
		0,35	1944	9	1953	36	11	47
		0,4	1944	9	1953	36	11	47
		0,45	1944	9	1953	36	11	47
		0,5	1944	9	1953	36	11	47
	0,35	0	1944	9	1953	36	11	47
		0,05	1944	9	1953	36	11	47
		0,1	1944	9	1953	36	11	47
		0,15	1944	9	1953	36	11	47
		0,2	1944	9	1953	36	11	47
		0,25	1944	9	1953	36	11	47
		0,3	1944	9	1953	36	11	47
		0,35	1944	9	1953	36	11	47
		0,4	1944	9	1953	36	11	47
		0,45	1944	9	1953	36	11	47
		0,5	1944	9	1953	36	11	47
	0,4	0	1944	9	1953	36	11	47
		0,05	1944	9	1953	36	11	47
		0,1	1944	9	1953	36	11	47
		0,15	1944	9	1953	36	11	47
		0,2	1944	9	1953	36	11	47
		0,25	1944	9	1953	36	11	47
		0,3	1944	9	1953	36	11	47
		0,35	1944	9	1953	36	11	47
		0,4	1944	9	1953	36	11	47
		0,45	1944	9	1953	36	11	47
		0,5	1944	9	1953	36	11	47

Grenswaarde Sensitiviteit (s0)	c1	c2	juist klasse 0 top 2000	juist klasse 1 top 2000	juist totaal top 2000	fout voorspeld als klasse 0	fout voorspeld als klasse 1	fout voorspeld totaal
0,000011	0,45	0	1944	9	1953	36	11	47
		0,05	1944	9	1953	36	11	47
		0,1	1944	9	1953	36	11	47
		0,15	1944	9	1953	36	11	47
		0,2	1944	9	1953	36	11	47
		0,25	1944	9	1953	36	11	47
		0,3	1944	9	1953	36	11	47
		0,35	1944	9	1953	36	11	47
		0,4	1944	9	1953	36	11	47
		0,45	1944	9	1953	36	11	47
		0,5	1944	9	1953	36	11	47
	0,5	0	1944	9	1953	36	11	47
		0,05	1944	9	1953	36	11	47
		0,1	1944	9	1953	36	11	47
		0,15	1944	9	1953	36	11	47
		0,2	1944	9	1953	36	11	47
		0,25	1944	9	1953	36	11	47
		0,3	1944	9	1953	36	11	47
		0,35	1944	9	1953	36	11	47
		0,4	1944	9	1953	36	11	47
		0,45	1944	9	1953	36	11	47
		0,5	1944	9	1953	36	11	47
0,000012	0	0	1904	18	1922	54	24	78
		0,05	1880	36	1916	54	30	84
		0,1	1880	36	1916	54	30	84
		0,15	1865	45	1910	54	36	90
		0,2	1828	72	1900	52	48	100
		0,25	1439	354	1793	45	162	207
		0,3	1434	357	1791	45	164	209
		0,35	1382	387	1769	44	187	231
		0,4	1369	394	1763	44	193	237
		0,45	1366	396	1762	44	194	238
		0,5	1366	396	1762	44	194	238

Grenswaarde Sensitiviteit (s0)	c1	c2	juist klasse 0 top 2000	juist klasse 1 top 2000	juist totaal top 2000	fout voorspeld als klasse 0	fout voorspeld als klasse 1	fout voorspeld totaal
0,000011	0	0	1904	18	1922	54	24	78
		0,05	1944	9	1953	36	11	47
		0,1	1944	9	1953	36	11	47
		0,15	1944	9	1953	36	11	47
		0,2	1944	9	1953	36	11	47
		0,25	1944	9	1953	36	11	47
		0,3	1944	9	1953	36	11	47
		0,35	1944	9	1953	36	11	47
		0,4	1944	9	1953	36	11	47
		0,45	1944	9	1953	36	11	47
		0,5	1944	9	1953	36	11	47
	0,05	0	1944	9	1953	36	11	47
		0,05	1944	9	1953	36	11	47
		0,1	1944	9	1953	36	11	47
		0,15	1944	9	1953	36	11	47
		0,2	1944	9	1953	36	11	47
		0,25	1944	9	1953	36	11	47
		0,3	1944	9	1953	36	11	47
		0,35	1944	9	1953	36	11	47
		0,4	1944	9	1953	36	11	47
		0,45	1944	9	1953	36	11	47
		0,5	1944	9	1953	36	11	47
	0,1	0	1944	9	1953	36	11	47
		0,05	1944	9	1953	36	11	47
		0,1	1944	9	1953	36	11	47
		0,15	1944	9	1953	36	11	47
		0,2	1944	9	1953	36	11	47
		0,25	1944	9	1953	36	11	47
		0,3	1944	9	1953	36	11	47
		0,35	1944	9	1953	36	11	47
		0,4	1944	9	1953	36	11	47
		0,45	1944	9	1953	36	11	47
		0,5	1944	9	1953	36	11	47

Grenswaarde Sensitiviteit (s0)	c1	c2	juist klasse 0 top 2000	juist klasse 1 top 2000	juist totaal top 2000	fout voorspeld als klasse 0	fout voorspeld als klasse 1	fout voorspeld totaal
0,000012	0,35	0	1382	387	1769	44	187	231
		0,05	1369	394	1763	44	193	237
		0,1	1366	396	1762	44	194	238
		0,15	1366	396	1762	44	194	238
		0,2	1366	396	1762	44	194	238
		0,25	1366	396	1762	44	194	238
		0,3	1366	396	1762	44	194	238
		0,35	1366	396	1762	44	194	238
		0,4	1366	396	1762	44	194	238
		0,45	1366	396	1762	44	194	238
		0,5	1366	396	1762	44	194	238
	0,4	0	1369	394	1763	44	193	237
		0,05	1366	396	1762	44	194	238
		0,1	1366	396	1762	44	194	238
		0,15	1366	396	1762	44	194	238
		0,2	1366	396	1762	44	194	238
		0,25	1366	396	1762	44	194	238
		0,3	1366	396	1762	44	194	238
		0,35	1366	396	1762	44	194	238
		0,4	1366	396	1762	44	194	238
		0,45	1366	396	1762	44	194	238
		0,5	1366	396	1762	44	194	238
	0,45	0	1366	396	1762	44	194	238
		0,05	1366	396	1762	44	194	238
		0,1	1366	396	1762	44	194	238
		0,15	1366	396	1762	44	194	238
		0,2	1366	396	1762	44	194	238
		0,25	1366	396	1762	44	194	238
		0,3	1366	396	1762	44	194	238
		0,35	1366	396	1762	44	194	238
		0,4	1366	396	1762	44	194	238
		0,45	1366	396	1762	44	194	238
		0,5	1366	396	1762	44	194	238

Grenswaarde Sensitiviteit (s0)	c1	c2	juist klasse 0 top 2000	juist klasse 1 top 2000	juist totaal top 2000	fout voorspeld als klasse 0	fout voorspeld als klasse 1	fout voorspeld totaal
0,000012	0,35	0	1382	387	1769	44	187	231
		0,05	1369	394	1763	44	193	237
		0,1	1366	396	1762	44	194	238
		0,15	1366	396	1762	44	194	238
		0,2	1366	396	1762	44	194	238
		0,25	1366	396	1762	44	194	238
		0,3	1366	396	1762	44	194	238
		0,35	1366	396	1762	44	194	238
		0,4	1366	396	1762	44	194	238
		0,45	1366	396	1762	44	194	238
		0,5	1366	396	1762	44	194	238
	0,4	0	1369	394	1763	44	193	237
		0,05	1366	396	1762	44	194	238
		0,1	1366	396	1762	44	194	238
		0,15	1366	396	1762	44	194	238
		0,2	1366	396	1762	44	194	238
		0,25	1366	396	1762	44	194	238
		0,3	1366	396	1762	44	194	238
		0,35	1366	396	1762	44	194	238
		0,4	1366	396	1762	44	194	238
		0,45	1366	396	1762	44	194	238
		0,5	1366	396	1762	44	194	238
	0,45	0	1366	396	1762	44	194	238
		0,05	1366	396	1762	44	194	238
		0,1	1366	396	1762	44	194	238
		0,15	1366	396	1762	44	194	238
		0,2	1366	396	1762	44	194	238
		0,25	1366	396	1762	44	194	238
		0,3	1366	396	1762	44	194	238
		0,35	1366	396	1762	44	194	238
		0,4	1366	396	1762	44	194	238
		0,45	1366	396	1762	44	194	238
		0,5	1366	396	1762	44	194	238

Grenswaarde Sensitiviteit (s0)	c1	c2	juist klasse 0 top 2000	juist klasse 1 top 2000	juist totaal top 2000	fout voorspeld als klasse 0	fout voorspeld als klasse 1	fout voorspeld totaal
0,000012	0,5	0	1366	396	1762	44	194	238
		0,05	1366	396	1762	44	194	238
		0,1	1366	396	1762	44	194	238
		0,15	1366	396	1762	44	194	238
		0,2	1366	396	1762	44	194	238
		0,25	1366	396	1762	44	194	238
		0,3	1366	396	1762	44	194	238
		0,35	1366	396	1762	44	194	238
		0,4	1366	396	1762	44	194	238
		0,45	1366	396	1762	44	194	238
		0,5	1366	396	1762	44	194	238
1,37594	0	0	1904	18	1922	54	24	78
		0,05	1904	18	1922	54	24	78
		0,1	1904	18	1922	54	24	78
		0,15	1904	18	1922	54	24	78
		0,2	1904	18	1922	54	24	78
		0,25	1904	18	1922	54	24	78
		0,3	1904	18	1922	54	24	78
		0,35	1904	18	1922	54	24	78
		0,4	1904	18	1922	54	24	78
		0,45	1904	18	1922	54	24	78
		0,5	1904	18	1922	54	24	78

Bijlage H

Mogelijke waarden voor de sensitiviteit berekend met behulp van de Min-max metriek en het cumulatief aantal gevallen

Aantal gevallen	Sensitiviteit
9479	0,000000000000
9511	0,000001000000
9539	0,000002000000
9556	0,000003000000
9576	0,000004000000
9597	0,000005000000
9612	0,000006000000
9631	0,000007000000
9641	0,000008000000
9646	0,000009000000
9653	0,000010000000
9664	0,000011000000
9671	0,000012000000
9677	0,000013000000
9689	0,000014000000
9696	0,000015000000
9701	0,000016000000
9707	0,000017000000
9712	0,000018000001
9721	0,000018999999
9728	0,000019999999
9737	0,000021000000
9742	0,000022000000
9746	0,000023000001
9747	0,000024000001
9750	0,000024999999
9753	0,000026000000
9755	0,000027000000
9760	0,000028000000
9762	0,000029000001
9766	0,000029999999
9770	0,000031000000
9773	0,000032000000
9775	0,000033000000
9776	0,000034000001

Aantal gevallen	Sensitiviteit
9778	0,000036000001
9781	0,000037000002
9784	0,000037999998
9787	0,000038999999
9788	0,000039999999
9790	0,000040999999
9791	0,000042000000
9793	0,000043000000
9794	0,000045000001
9796	0,000048000002
9800	0,000048999998
9802	0,000050999999
9803	0,000051999999
9805	0,000053000000
9806	0,000054000000
9809	0,000055000000
9810	0,000056000001
9811	0,000057000001
9813	0,000058000001
9815	0,000059000002
9817	0,000059999998
9819	0,000061999999
9821	0,000062999999
9822	0,000064000000
9823	0,000065000000
9826	0,000066000001
9828	0,000067000001
9829	0,000070000002
9832	0,000074000003
9834	0,000075000004
9835	0,000075999997
9836	0,000076999997
9837	0,000077999997
9838	0,000078999998
9841	0,000081999999

Aantal gevallen	Sensitiviteit
9842	0,000083999999
9844	0,000085000000
9845	0,000086000000
9846	0,000091000002
9847	0,000092000002
9848	0,000095000003
9849	0,000096000003
9850	0,000097999997
9851	0,000098999997
9852	0,000102999998
9853	0,000107000000
9854	0,000108000000
9855	0,000109000001
9859	0,000110000001
9860	0,000111000001
9861	0,000112000002
9862	0,000113000002
9863	0,000116000003
9865	0,000118000004
9866	0,000118999997
9867	0,000119999997
9868	0,000120999997
9869	0,000121999998
9870	0,000123999998
9871	0,000128000000
9873	0,000130000000
9874	0,000132000001
9875	0,000134000002
9876	0,000134999995
9879	0,000136000002
9880	0,000136999995
9881	0,000138000003
9882	0,000138999996
9885	0,000140999997
9887	0,000142000004

Aantal gevallen	Sensitiviteit
9888	0,000146000006
9889	0,000148000006
9890	0,000149000000
9893	0,000153999994
9894	0,000159999996
9895	0,000161999997
9896	0,000165999998
9899	0,000182000003
9900	0,000182999996
9902	0,000184999997
9903	0,000190999999
9904	0,000193999993
9905	0,000195999994
9906	0,000197000001
9907	0,000205999997
9909	0,000207000005
9910	0,000207999998
9911	0,000211000006
9913	0,000211999999
9914	0,000214000000
9915	0,000216000000
9916	0,000226000004
9917	0,000237999993
9918	0,000241999995
9919	0,000243000002
9921	0,000246999989
9922	0,000250000012
9923	0,000253000006
9924	0,000260000001
9925	0,000261000008
9927	0,000265000010
9928	0,000277999992
9930	0,000280999986
9931	0,000281999994
9932	0,000283000001

Aantal gevallen	Sensitiviteit
9933	0,000288999989
9934	0,000300999993
9935	0,000302000000
9936	0,000304999994
9937	0,000316999998
9938	0,000318000006
9939	0,000327999995
9940	0,000333000004
9941	0,000338000013
9942	0,000345000008
9943	0,000349999988
9944	0,000372000010
9945	0,000384000014
9946	0,000387999986
9947	0,000413000002
9948	0,000427999999
9949	0,000437000010
9950	0,000440000003
9951	0,000441999990
9952	0,000464000012
9953	0,000471000007
9954	0,000472999993
9955	0,000483999989
9956	0,000497000001
9957	0,000538999971
9958	0,000549999997
9959	0,000553000020
9960	0,000586999988
9961	0,000590000011
9962	0,000614000019
9963	0,000621000014
9964	0,000625999994
9965	0,000683000020
9966	0,000705000013
9967	0,000707999978

Aantal gevallen	Sensitiviteit
9968	0,000745999976
9969	0,000766000012
9970	0,000800999987
9971	0,000801999995
9972	0,000811000005
9973	0,000824999996
9974	0,000890999974
9975	0,000910000002
9976	0,000980000012
9977	0,000983000034
9978	0,001102000009
9979	0,001215999946
9980	0,001550000045
9981	0,001581000048
9982	0,001660999958
9983	0,001950999955
9984	0,002498999937
9985	0,002644999884
9986	0,002907000016
9987	0,003957000095
9988	0,004064999986
9989	0,004085000139
9990	0,004251999781
9991	0,004451000132
9992	0,004639000166
9993	0,006723000202
9994	0,006752000190
9995	0,006806999911
9996	0,006980999839
9997	0,009696999565
9998	0,012337000109
9999	0,015265000053
10000	0,023514000699

Bijlage I

Resultaten gecorrigeerde waarschijnlijkheidsschattingen Min-max metriek

Grenswaarde Sensitiviteit (s0)	c1	c2	juist klasse 0 top 2000	juist klasse 1 top 2000	juist totaal top 2000	fout voorspeld als klasse 0	fout voorspeld als klasse 1	fout voorspeld totaal
0	0	0	1904	18	1922	54	24	78
		0,05	1880	36	1916	54	30	84
		0,1	1880	36	1916	54	30	84
		0,15	1865	45	1910	54	36	90
		0,2	1829	71	1900	52	48	100
		0,25	1477	328	1805	46	149	195
		0,3	1472	331	1803	46	151	197
		0,35	1422	361	1783	44	173	217
		0,4	1410	368	1778	44	178	222
		0,45	1407	370	1777	44	179	223
		0,5	1407	370	1777	44	179	223
0,0000060000	0	0	1904	18	1922	54	24	78
		0,05	1882	34	1916	54	30	84
		0,1	1882	34	1916	54	30	84
		0,15	1867	43	1910	54	36	90
		0,2	1839	65	1904	52	44	96
		0,25	1591	248	1839	48	113	161
		0,3	1586	251	1837	48	115	163
		0,35	1543	276	1819	48	133	181
		0,4	1532	282	1814	48	138	186
		0,45	1530	283	1813	48	139	187
		0,5	1530	283	1813	48	139	187
	0,05	0	1882	34	1916	54	30	84
		0,05	1882	34	1916	54	30	84
		0,1	1867	43	1910	54	36	90
		0,15	1839	65	1904	52	44	96
		0,2	1591	248	1839	48	113	161
		0,25	1586	251	1837	48	115	163
		0,3	1543	276	1819	48	133	181
		0,35	1532	282	1814	48	138	186
		0,4	1530	283	1813	48	139	187
		0,45	1530	283	1813	48	139	187
		0,5	1530	283	1813	48	139	187

Grenswaarde Sensitiviteit (s0)	c1	c2	juist klasse 0 top 2000	juist klasse 1 top 2000	juist totaal top 2000	fout voorspeld als klasse 0	fout voorspeld als klasse 1	fout voorspeld totaal
	0,1	0	1882	34	1916	54	30	84
		0,05	1867	43	1910	54	36	90
		0,1	1839	65	1904	52	44	96
		0,15	1591	248	1839	48	113	161
		0,2	1586	251	1837	48	115	163
		0,25	1543	276	1819	48	133	181
		0,3	1532	282	1814	48	138	186
		0,35	1530	283	1813	48	139	187
		0,4	1530	283	1813	48	139	187
		0,45	1530	283	1813	48	139	187
		0,5	1530	283	1813	48	139	187
	0,15	0	1867	43	1910	54	36	90
		0,05	1839	65	1904	52	44	96
		0,1	1591	248	1839	48	113	161
		0,15	1586	251	1837	48	115	163
		0,2	1543	276	1819	48	133	181
		0,25	1532	282	1814	48	138	186
		0,3	1530	283	1813	48	139	187
		0,35	1530	283	1813	48	139	187
		0,4	1530	283	1813	48	139	187
		0,45	1530	283	1813	48	139	187
		0,5	1530	283	1813	48	139	187
	0,2	0	1839	65	1904	52	44	96
		0,05	1591	248	1839	48	113	161
		0,1	1586	251	1837	48	115	163
		0,15	1543	276	1819	48	133	181
		0,2	1532	282	1814	48	138	186
		0,25	1530	283	1813	48	139	187
		0,3	1530	283	1813	48	139	187
		0,35	1530	283	1813	48	139	187
		0,4	1530	283	1813	48	139	187
		0,45	1530	283	1813	48	139	187
		0,5	1530	283	1813	48	139	187

Grenswaarde Sensitiviteit (s0)	c1	c2	juist klasse 0 top 2000	juist klasse 1 top 2000	juist totaal top 2000	fout voorspeld als klasse 0	fout voorspeld als klasse 1	fout voorspeld totaal
	0,25	0	1591	248	1839	48	113	161
		0,05	1586	251	1837	48	115	163
		0,1	1543	276	1819	48	133	181
		0,15	1532	282	1814	48	138	186
		0,2	1530	283	1813	48	139	187
		0,25	1530	283	1813	48	139	187
		0,3	1530	283	1813	48	139	187
		0,35	1530	283	1813	48	139	187
		0,4	1530	283	1813	48	139	187
		0,45	1530	283	1813	48	139	187
		0,5	1530	283	1813	48	139	187
	0,3	0	1586	251	1837	48	115	163
		0,05	1543	276	1819	48	133	181
		0,1	1532	282	1814	48	138	186
		0,15	1530	283	1813	48	139	187
		0,2	1530	283	1813	48	139	187
		0,25	1530	283	1813	48	139	187
		0,3	1530	283	1813	48	139	187
		0,35	1530	283	1813	48	139	187
		0,4	1530	283	1813	48	139	187
		0,45	1530	283	1813	48	139	187
		0,5	1530	283	1813	48	139	187
	0,35	0	1543	276	1819	48	133	181
		0,05	1532	282	1814	48	138	186
		0,1	1530	283	1813	48	139	187
		0,15	1530	283	1813	48	139	187
		0,2	1530	283	1813	48	139	187
		0,25	1530	283	1813	48	139	187
		0,3	1530	283	1813	48	139	187
		0,35	1530	283	1813	48	139	187
		0,4	1530	283	1813	48	139	187
		0,45	1530	283	1813	48	139	187
		0,5	1530	283	1813	48	139	187

Grenswaarde Sensitiviteit (s0)	c1	c2	juist klasse 0 top 2000	juist klasse 1 top 2000	juist totaal top 2000	fout voorspeld als klasse 0	fout voorspeld als klasse 1	fout voorspeld totaal
	0,4	0	1532	282	1814	48	138	186
		0,05	1530	283	1813	48	139	187
		0,1	1530	283	1813	48	139	187
		0,15	1530	283	1813	48	139	187
		0,2	1530	283	1813	48	139	187
		0,25	1530	283	1813	48	139	187
		0,3	1530	283	1813	48	139	187
		0,35	1530	283	1813	48	139	187
		0,4	1530	283	1813	48	139	187
		0,45	1530	283	1813	48	139	187
		0,5	1530	283	1813	48	139	187
	0,45	0	1530	283	1813	48	139	187
		0,05	1530	283	1813	48	139	187
		0,1	1530	283	1813	48	139	187
		0,15	1530	283	1813	48	139	187
		0,2	1530	283	1813	48	139	187
		0,25	1530	283	1813	48	139	187
		0,3	1530	283	1813	48	139	187
		0,35	1530	283	1813	48	139	187
		0,4	1530	283	1813	48	139	187
		0,45	1530	283	1813	48	139	187
		0,5	1530	283	1813	48	139	187
	0,5	0	1530	283	1813	48	139	187
		0,05	1530	283	1813	48	139	187
		0,1	1530	283	1813	48	139	187
		0,15	1530	283	1813	48	139	187
		0,2	1530	283	1813	48	139	187
		0,25	1530	283	1813	48	139	187
		0,3	1530	283	1813	48	139	187
		0,35	1530	283	1813	48	139	187
		0,4	1530	283	1813	48	139	187
		0,45	1530	283	1813	48	139	187
		0,5	1530	283	1813	48	139	187

Grenswaarde Sensitiviteit (s0)	c1	c2	juist klasse 0 top 2000	juist klasse 1 top 2000	juist totaal top 2000	fout voorspeld als klasse 0	fout voorspeld als klasse 1	fout voorspeld totaal
0,000048999998	0	0	1904	18	1922	54	24	78
		0,05	1895	26	1921	54	25	79
		0,1	1895	26	1921	54	25	79
		0,15	1884	34	1918	54	28	82
		0,2	1870	44	1914	54	32	86
		0,25	1757	130	1887	51	62	113
		0,3	1754	131	1885	51	64	115
		0,35	1723	151	1874	51	75	126
		0,4	1714	155	1869	51	80	131
		0,45	1712	156	1868	51	81	132
		0,5	1712	156	1868	51	81	132
	0,05	0	1895	26	1921	54	25	79
		0,05	1895	26	1921	54	25	79
		0,1	1884	34	1918	54	28	82
		0,15	1870	44	1914	54	32	86
		0,2	1757	130	1887	51	62	113
		0,25	1754	131	1885	51	64	115
		0,3	1723	151	1874	51	75	126
		0,35	1714	155	1869	51	80	131
		0,4	1712	156	1868	51	81	132
		0,45	1712	156	1868	51	81	132
		0,5	1712	156	1868	51	81	132
	0,1	0	1895	26	1921	54	25	79
		0,05	1884	34	1918	54	28	82
		0,1	1870	44	1914	54	32	86
		0,15	1757	130	1887	51	62	113
		0,2	1754	131	1885	51	64	115
		0,25	1723	151	1874	51	75	126
		0,3	1714	155	1869	51	80	131
		0,35	1712	156	1868	51	81	132
		0,4	1712	156	1868	51	81	132
		0,45	1712	156	1868	51	81	132
		0,5	1712	156	1868	51	81	132

Grenswaarde Sensitiviteit (s0)	c1	c2	juist klasse 0 top 2000	juist klasse 1 top 2000	juist totaal top 2000	fout voorspeld als klasse 0	fout voorspeld als klasse 1	fout voorspeld totaal
	0,15	0	1884	34	1918	54	28	82
		0,05	1870	44	1914	54	32	86
		0,1	1757	130	1887	51	62	113
		0,15	1754	131	1885	51	64	115
		0,2	1723	151	1874	51	75	126
		0,25	1714	155	1869	51	80	131
		0,3	1712	156	1868	51	81	132
		0,35	1712	156	1868	51	81	132
		0,4	1712	156	1868	51	81	132
		0,45	1712	156	1868	51	81	132
		0,5	1712	156	1868	51	81	132
	0,2	0	1870	44	1914	54	32	86
		0,05	1757	130	1887	51	62	113
		0,1	1754	131	1885	51	64	115
		0,15	1723	151	1874	51	75	126
		0,2	1714	155	1869	51	80	131
		0,25	1712	156	1868	51	81	132
		0,3	1712	156	1868	51	81	132
		0,35	1712	156	1868	51	81	132
		0,4	1712	156	1868	51	81	132
		0,45	1712	156	1868	51	81	132
		0,5	1712	156	1868	51	81	132
	0,25	0	1757	130	1887	51	62	113
		0,05	1754	131	1885	51	64	115
		0,1	1723	151	1874	51	75	126
		0,15	1714	155	1869	51	80	131
		0,2	1712	156	1868	51	81	132
		0,25	1712	156	1868	51	81	132
		0,3	1712	156	1868	51	81	132
		0,35	1712	156	1868	51	81	132
		0,4	1712	156	1868	51	81	132
		0,45	1712	156	1868	51	81	132
		0,5	1712	156	1868	51	81	132

Grenswaarde Sensitiviteit (s0)	c1	c2	juist klasse 0 top 2000	juist klasse 1 top 2000	juist totaal top 2000	fout voorspeld als klasse 0	fout voorspeld als klasse 1	fout voorspeld totaal
	0,3	0	1754	131	1885	51	64	115
		0,05	1723	151	1874	51	75	126
		0,1	1714	155	1869	51	80	131
		0,15	1712	156	1868	51	81	132
		0,2	1712	156	1868	51	81	132
		0,25	1712	156	1868	51	81	132
		0,3	1712	156	1868	51	81	132
		0,35	1712	156	1868	51	81	132
		0,4	1712	156	1868	51	81	132
		0,45	1712	156	1868	51	81	132
		0,5	1712	156	1868	51	81	132
	0,35	0	1723	151	1874	51	75	126
		0,05	1714	155	1869	51	80	131
		0,1	1712	156	1868	51	81	132
		0,15	1712	156	1868	51	81	132
		0,2	1712	156	1868	51	81	132
		0,25	1712	156	1868	51	81	132
		0,3	1712	156	1868	51	81	132
		0,35	1712	156	1868	51	81	132
		0,4	1712	156	1868	51	81	132
		0,45	1712	156	1868	51	81	132
		0,5	1712	156	1868	51	81	132
	0,4	0	1714	155	1869	51	80	131
		0,05	1712	156	1868	51	81	132
		0,1	1712	156	1868	51	81	132
		0,15	1712	156	1868	51	81	132
		0,2	1712	156	1868	51	81	132
		0,25	1712	156	1868	51	81	132
		0,3	1712	156	1868	51	81	132
		0,35	1712	156	1868	51	81	132
		0,4	1712	156	1868	51	81	132
		0,45	1712	156	1868	51	81	132
		0,5	1712	156	1868	51	81	132

Grenswaarde Sensitiviteit (s0)	c1	c2	juist klasse 0 top 2000	juist klasse 1 top 2000	juist totaal top 2000	fout voorspeld als klasse 0	fout voorspeld als klasse 1	fout voorspeld totaal
	0,45	0	1712	156	1868	51	81	132
		0,05	1712	156	1868	51	81	132
		0,1	1712	156	1868	51	81	132
		0,15	1712	156	1868	51	81	132
		0,2	1712	156	1868	51	81	132
		0,25	1712	156	1868	51	81	132
		0,3	1712	156	1868	51	81	132
		0,35	1712	156	1868	51	81	132
		0,4	1712	156	1868	51	81	132
		0,45	1712	156	1868	51	81	132
		0,5	1712	156	1868	51	81	132
	0,5	0	1712	156	1868	51	81	132
		0,05	1712	156	1868	51	81	132
		0,1	1712	156	1868	51	81	132
		0,15	1712	156	1868	51	81	132
		0,2	1712	156	1868	51	81	132
		0,25	1712	156	1868	51	81	132
		0,3	1712	156	1868	51	81	132
		0,35	1712	156	1868	51	81	132
		0,4	1712	156	1868	51	81	132
		0,45	1712	156	1868	51	81	132
		0,5	1712	156	1868	51	81	132
0,000183	0	0	1904	18	1922	54	24	78
		0,05	1900	21	1921	54	25	79
		0,1	1900	21	1921	54	25	79
		0,15	1894	25	1919	54	27	81
		0,2	1887	30	1917	54	29	83
		0,25	1832	75	1907	52	41	93
		0,3	1830	76	1906	52	42	94
		0,35	1815	89	1904	52	44	96
		0,4	1809	91	1900	52	48	100
		0,45	1808	91	1899	52	49	101
		0,5	1808	91	1899	52	49	101

Grenswaarde Sensitiviteit (s0)	c1	c2	juist klasse 0 top 2000	juist klasse 1 top 2000	juist totaal top 2000	fout voorspeld als klasse 0	fout voorspeld als klasse 1	fout voorspeld totaal
	0,05	0	1900	21	1921	54	25	79
		0,05	1900	21	1921	54	25	79
		0,1	1894	25	1919	54	27	81
		0,15	1887	30	1917	54	29	83
		0,2	1832	75	1907	52	41	93
		0,25	1830	76	1906	52	42	94
		0,3	1815	89	1904	52	44	96
		0,35	1809	91	1900	52	48	100
		0,4	1808	91	1899	52	49	101
		0,45	1808	91	1899	52	49	101
		0,5	1808	91	1899	52	49	101
	0,1	0	1900	21	1921	54	25	79
		0,05	1894	25	1919	54	27	81
		0,1	1887	30	1917	54	29	83
		0,15	1832	75	1907	52	41	93
		0,2	1830	76	1906	52	42	94
		0,25	1815	89	1904	52	44	96
		0,3	1809	91	1900	52	48	100
		0,35	1808	91	1899	52	49	101
		0,4	1808	91	1899	52	49	101
		0,45	1808	91	1899	52	49	101
		0,5	1808	91	1899	52	49	101
	0,15	0	1894	25	1919	54	27	81
		0,05	1887	30	1917	54	29	83
		0,1	1832	75	1907	52	41	93
		0,15	1830	76	1906	52	42	94
		0,2	1815	89	1904	52	44	96
		0,25	1809	91	1900	52	48	100
		0,3	1808	91	1899	52	49	101
		0,35	1808	91	1899	52	49	101
		0,4	1808	91	1899	52	49	101
		0,45	1808	91	1899	52	49	101
		0,5	1808	91	1899	52	49	101

Grenswaarde Sensitiviteit (s0)	c1	c2	juist klasse 0 top 2000	juist klasse 1 top 2000	juist totaal top 2000	fout voorspeld als klasse 0	fout voorspeld als klasse 1	fout voorspeld totaal
	0,2	0	1887	30	1917	54	29	83
		0,05	1832	75	1907	52	41	93
		0,1	1830	76	1906	52	42	94
		0,15	1815	89	1904	52	44	96
		0,2	1809	91	1900	52	48	100
		0,25	1808	91	1899	52	49	101
		0,3	1808	91	1899	52	49	101
		0,35	1808	91	1899	52	49	101
		0,4	1808	91	1899	52	49	101
		0,45	1808	91	1899	52	49	101
		0,5	1808	91	1899	52	49	101
	0,25	0	1832	75	1907	52	41	93
		0,05	1830	76	1906	52	42	94
		0,1	1815	89	1904	52	44	96
		0,15	1809	91	1900	52	48	100
		0,2	1808	91	1899	52	49	101
		0,25	1808	91	1899	52	49	101
		0,3	1808	91	1899	52	49	101
		0,35	1808	91	1899	52	49	101
		0,4	1808	91	1899	52	49	101
		0,45	1808	91	1899	52	49	101
		0,5	1808	91	1899	52	49	101
	0,3	0	1830	76	1906	52	42	94
		0,05	1815	89	1904	52	44	96
		0,1	1809	91	1900	52	48	100
		0,15	1808	91	1899	52	49	101
		0,2	1808	91	1899	52	49	101
		0,25	1808	91	1899	52	49	101
		0,3	1808	91	1899	52	49	101
		0,35	1808	91	1899	52	49	101
		0,4	1808	91	1899	52	49	101
		0,45	1808	91	1899	52	49	101
		0,5	1808	91	1899	52	49	101

Grenswaarde Sensitiviteit (s0)	c1	c2	juist klasse 0 top 2000	juist klasse 1 top 2000	juist totaal top 2000	fout voorspeld als klasse 0	fout voorspeld als klasse 1	fout voorspeld totaal
	0,35	0	1815	89	1904	52	44	96
		0,05	1809	91	1900	52	48	100
		0,1	1808	91	1899	52	49	101
		0,15	1808	91	1899	52	49	101
		0,2	1808	91	1899	52	49	101
		0,25	1808	91	1899	52	49	101
		0,3	1808	91	1899	52	49	101
		0,35	1808	91	1899	52	49	101
		0,4	1808	91	1899	52	49	101
		0,45	1808	91	1899	52	49	101
		0,5	1808	91	1899	52	49	101
	0,4	0	1809	91	1900	52	48	100
		0,05	1808	91	1899	52	49	101
		0,1	1808	91	1899	52	49	101
		0,15	1808	91	1899	52	49	101
		0,2	1808	91	1899	52	49	101
		0,25	1808	91	1899	52	49	101
		0,3	1808	91	1899	52	49	101
		0,35	1808	91	1899	52	49	101
		0,4	1808	91	1899	52	49	101
		0,45	1808	91	1899	52	49	101
		0,5	1808	91	1899	52	49	101
	0,45	0	1808	91	1899	52	49	101
		0,05	1808	91	1899	52	49	101
		0,1	1808	91	1899	52	49	101
		0,15	1808	91	1899	52	49	101
		0,2	1808	91	1899	52	49	101
		0,25	1808	91	1899	52	49	101
		0,3	1808	91	1899	52	49	101
		0,35	1808	91	1899	52	49	101
		0,4	1808	91	1899	52	49	101
		0,45	1808	91	1899	52	49	101
		0,5	1808	91	1899	52	49	101

Grenswaarde Sensitiviteit (s0)	c1	c2	juist klasse 0 top 2000	juist klasse 1 top 2000	juist totaal top 2000	fout voorspeld als klasse 0	fout voorspeld als klasse 1	fout voorspeld totaal
	0,5	0	1808	91	1899	52	49	101
		0,05	1808	91	1899	52	49	101
		0,1	1808	91	1899	52	49	101
		0,15	1808	91	1899	52	49	101
		0,2	1808	91	1899	52	49	101
		0,25	1808	91	1899	52	49	101
		0,3	1808	91	1899	52	49	101
		0,35	1808	91	1899	52	49	101
		0,4	1808	91	1899	52	49	101
		0,45	1808	91	1899	52	49	101
		0,5	1808	91	1899	52	49	101
0,023514001	0	0	1904	18	1922	54	24	78
		0,05	1904	18	1922	54	24	78
		0,1	1904	18	1922	54	24	78
		0,15	1904	18	1922	54	24	78
		0,2	1904	18	1922	54	24	78
		0,25	1904	18	1922	54	24	78
		0,3	1904	18	1922	54	24	78
		0,35	1904	18	1922	54	24	78
		0,4	1904	18	1922	54	24	78
		0,45	1904	18	1922	54	24	78
		0,5	1904	18	1922	54	24	78