Risico-Analyse KLIC meldingen

Meer grip op de kabelschadepreventie van (graaf)werkzaamheden

Stagebedrijf:
KPN
Afdeling VM&B
Stationsstraat 115
Amersfoort

Opleidingsinstituut:
Vrije Universiteit
Faculteit der Wiskunde & Informatica
Studierichting Bedrijfswiskunde & Informatica
Boelelaan 1081
Amsterdam

Hugo Huijser
Nijkerk, 26 februari 2003

SAMENVATTING

De afdeling Voorraadmanagement & Beheer (VM&B) van KPN wil meer inzicht krijgen in de risico's van meldingen die gedaan worden bij het uitvoeren van graafwerkzaamheden. Wanneer er een goed beeld is van de risicovolle meldingen, kunnen preventieve maatregelen worden getroffen om de schadekans zoveel mogelijk te verlagen.

Binnen de stage is onderzocht hoe men meer inzicht kan krijgen in het schaderisico van de meldingen. Met behulp van Data Mining technieken is een model ontwikkeld dat voor elke nieuwe melding een risicofactor berekent.

Voordat dit proces kon plaatsvinden, was het nodig om bij elke kabelschade uit het verleden de melding terug te vinden van de desbetreffende (graaf)werkzaamheden. Met behulp van een koppelingsmethode zijn deze meldingen teruggevonden.

1. INLEIDING EN OPBOUW

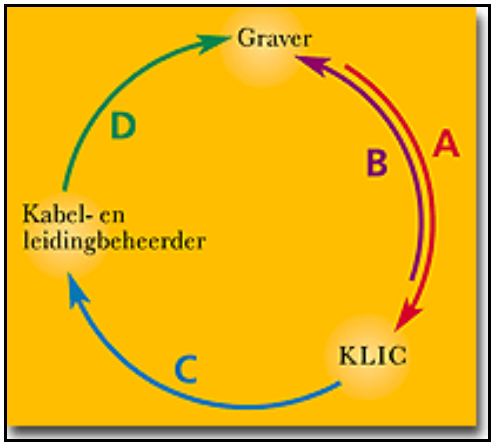
1.1 Inleiding

Wanneer er gesproken wordt over transport en distributie, gaat de gedachte al snel naar vrachtverkeer, naar vervoer over land, water en via de lucht, maar een belangrijk deel van het transport en distributie in Nederland blijft voor het oog verborgen. Dit deel betreft het ondergrondse verkeer via kabels en leidingen. Alle kabels en leidingen in Nederland bij elkaar hebben in totaal een lengte van circa 1¾ miljoen kilometer. Deze ondergrondse infrastructuur heeft een vervangingswaarde van méér dan 70 miljard euro.

Wanneer je ergens graaft, loop je het risico dat je één van deze leidingen of kabels beschadigt. Een beschadiging aan een belangrijke kabel of leiding kan verstrekkende gevolgen hebben. De eigenaren van kabels en leidingen weten precies waar deze gelegd zijn. Ze geven deze kennis graag door, in ieders belang. Hiervoor is het KLIC opgericht, het Kabels en Leidingen Informatie Centrum. Het KLIC beheert alle informatie over wat er onder de grond ligt.

Een graver dient minimaal drie dagen voor de graafwerkzaamheden een ‘KLIC melding’ te doen.

Figuur 1.1 geeft aan hoe het KLIC proces in zijn werk gaat.



Figuur 1.1 Het KLIC proces

De graver neemt op met KLIC per telefoon, e-mail of fax (**A**), KLIC bevestigt de graafmelding (**B**), KLIC geeft de melding door aan de kabel- en leidingbeheerder(s), waarvan KPN er één is (**C**), en deze neemt contact op met de graver (**D**).

Per jaar komen bij KPN zo'n 120.000 van deze KLIC-meldingen binnen.
Elke melding bevat informatie over de aanstaande grondroering, zoals:

- de (graaf)locatie
- de uitvoerende partij
- de aard van de werkzaamheden
- de startdatum van de werkzaamheden

KPN reageert op een melding, door de grondroerder, veelal een aannemer, te voorzien van tekeningen van de situatie ter plaatse zodat deze rekening kan houden met de kabels van KPN. Deze tekeningen komen uit een geautomatiseerd systeem waar de ligging van bijna alle kabels van KPN in opgeslagen staat.

Ondanks deze procedure verloopt het toch niet altijd zonder kleerscheuren. Per jaar wordt circa 12.000 keer één of meerdere kabels van KPN beschadigd.
Middels Adviseurs heeft KPN contact met plannenmakers rondom grondroeringen: gemeenten en projectontwikkelaars. Het werk van een Adviseur is erop gericht de kosten van die plannen voor KPN te minimaliseren.
KPN heeft Toezicht Werken Derden (TWD's) in dienst die KLIC meldingen bezoeken om toezicht te houden.
Er is besloten met minder Toezicht Werken Derden te gaan werken, waardoor het onmogelijk is om iedere melding te bezoeken. Om de beschikbare capaciteit zo efficiënt mogelijk in te zetten, is het wenselijk om alleen de meest risicovolle meldingen te bezoeken.

Deze wens is de aanleiding voor de stage-opdracht. De opdracht is uitgevoerd in opdracht van de afdeling Voorraadmanagement & Beheer (VM&B) in Amersfoort. Deze afdeling is landelijk verantwoordelijk voor de beheersprocessen van het netwerk. Leidingbeheer is daarbij één van de belangrijkste.

Bij de opdracht is gebruik gemaakt van Data Mining technieken. Data Mining is 'schatgraven' in grote hoeveelheden data. Vooral wanneer de informatie niet van te voren bekend is, maar wel impliciet in de data ligt opgeslagen, kan Data Mining uitkomst bieden.

1.2 Opbouw

De verslaggeving van de stage-opdracht is als volgt opgebouwd:

Allereerst worden de probleem- en doelstelling uitgewerkt. Hierna wordt dieper ingegaan op de huidige en gewenste situatie binnen KPN. Vervolgens wordt de aanpak van de opdracht beschreven. Hierna volgt de uitwerking hiervan. Er wordt afgesloten met conclusies en aanbevelingen.

2. PROBLEEM- EN DOELSTELLING

2.1 Probleemstelling

In het verleden werd veel toezicht gehouden bij grondroeringen om de schadekans te beperken. Sinds de laatste reorganisatie is in verband met bezuinigingen het aantal Toezicht Werken Derden sterk gereduceerd.
Door deze beperkte capaciteit is het niet mogelijk om toezicht te houden bij alle (risicovolle) grondroeringen. Slechts bij een beperkt deel van de grondroeringen zal toezicht gehouden kunnen worden om zo het schaderisico te verminderen.

Momenteel zijn het aantal KLIC meldingen in verhouding tot het aantal Toezicht Werken Derden zo groot dat de Toezicht Werken Derden vooral toezicht houden op basis van verzoeken van Adviseurs of aannemers.

2.2 Doelstelling

Uiteraard is het zinvol om díe grondroeringen te bezoeken, zodanig dat de kosten voor KPN minimaal zijn.
Voor het oplossen van de probleemstelling is de volgende doelstelling opgesteld:

Het maken van een risicoprofiel van KLIC meldingen zodat de Toezicht Werken Derden toezicht kunnen houden bij grondroeringen (m.b.v. prioriteitsstelling) zodat de totale schadepost voor KPN geminimaliseerd wordt.

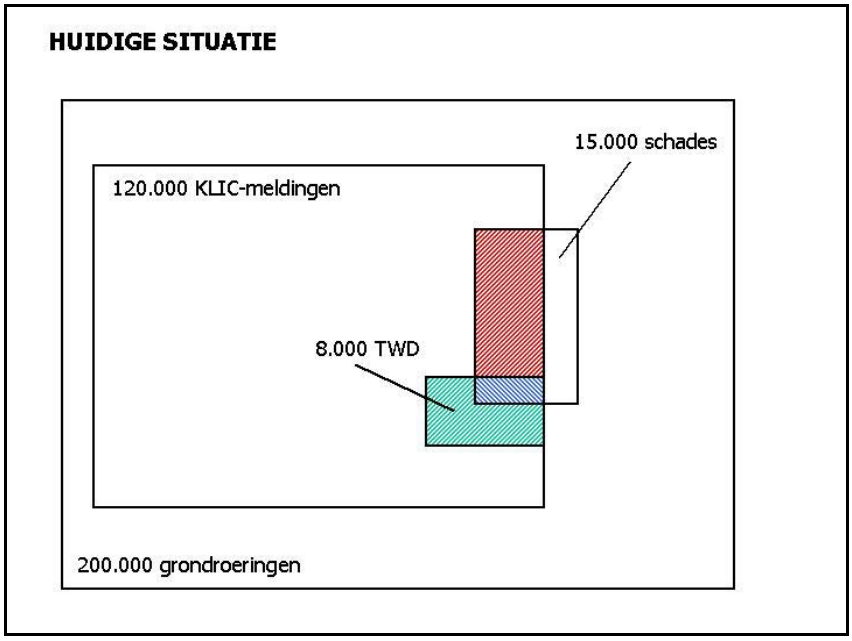
3. HUIDIGE EN GEWENSTE SITUATIE BINNEN KPN

3.1 Huidige situatie

Voor een goed beeld van de huidige situatie, wat betreft de KLIC meldingen, bezoeken van Toezicht Werken Derden (TWD) en kabelschades, zijn de volgende gegevens van belang:

- Jaarlijks worden 200.000 grondroeringen gedaan.
- Jaarlijks worden 120.000 KLIC meldingen gedaan.
- Jaarlijks zijn er 15.000 kabelschades t.g.v. een grondroering.
- Jaarlijks bezoeken de TWD 8.000 grondroeringen.

Dit geeft de volgende situatie:



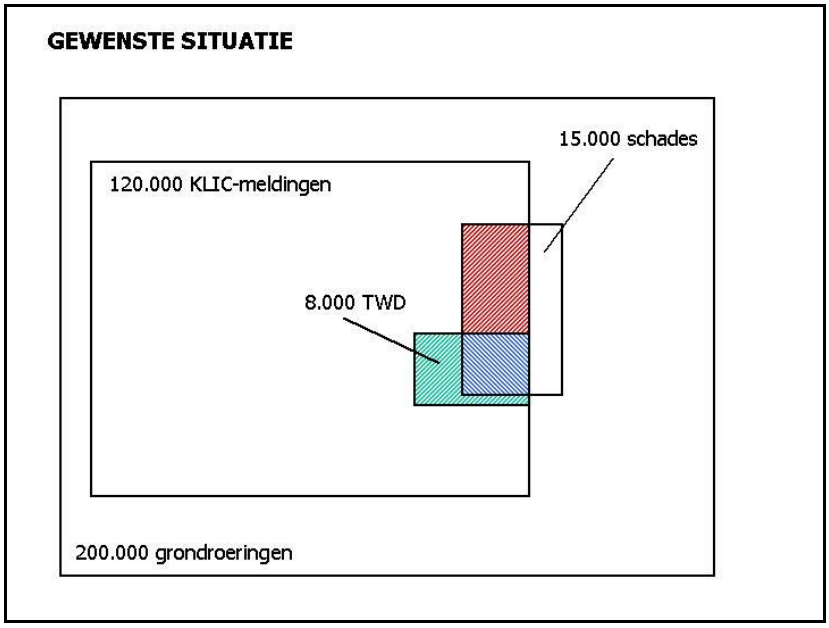
Figuur 3.1 Huidige situatie binnen KPN

In deze opdracht wordt enkel gekeken naar het gebied binnen de 120.000 KLIC meldingen. De grondroeringen waar geen KLIC melding wordt gedaan vormen een geheel ander probleem en zullen buiten beschouwing worden gelaten.
De TWD's bezoeken jaarlijks zo'n 8.000 grondroeringen en willen met deze bezoeken zoveel mogelijk kabelschades voorkomen.
Het rood gearceerde gebied bevat de KLIC meldingen van grondroeringen waarbij een kabel van KPN beschadigd werd.
Het groen gearceerde gebied staat voor de KLIC meldingen die bezocht werden door een TWD. Het overlappende gedeelte van deze twee gebieden (blauw gearceerd) zou dan moeten staan voor grondroeringen waar een TWD op bezoek is geweest en waar schade aan een kabel is gemaakt.
Nu doen zich een aantal onduidelijkheden voor: ten eerste is de grootte van dit blauw gearceerde gebied onduidelijk omdat er weinig informatie beschikbaar is omtrent de bezochte KLIC meldingen.

Verder is het niet duidelijk in hoeverre de TWD's verantwoordelijk zijn geweest voor het 'schadevrij-zijn' van de KLIC meldingen binnen het groene gebied. In hoeverre heeft de TWD invloed gehad op de werkwijze van de grondroerder zodanig dat de werkzaamheden géén schade tot gevolg hadden. Oftwel: was er bij deze grondroeringen wél schade geweest wanneer er geen TWD langs was geweest?

3.2 Gewenste situatie

Al weten we niet wat het precieze effect van een bezoek van een TWD is op de schadekans van een grondroering, nemen we wel aan dat deze kans in ieder geval kleiner wordt met een bezoek van een TWD.
In de gewenste situatie zullen de TWD's ingezet worden bij díe grondroeringen waar met grote kans een (kabel)schade wordt verwacht.



Figuur 3.2 Gewenste situatie

Wanneer we dit vertalen naar het figuur betekent dit dat we het blauw gearceerde gebied willen maximaliseren. We verwachten dat een deel van deze grondroeringen door toedoen van de TWD zonder schade zal zijn.

4. DATAMINING

In dit hoofdstuk zal het begrip Data Mining worden gedefinieerd en uitgelegd. Belangrijke begrippen en ideeën omtrent Data Mining zullen aan de orde komen.

4.1 Inleiding

De hoeveelheid informatie in de wereld verdubbelt iedere 20 maanden. Het aantal en de grootte van databases groeit waarschijnlijk nog sneller. Iedere dag worden enorme hoeveelheden gegevens verzameld en opgeslagen. De ontwikkelingen op het gebied van Internet en Intranet dragen hier ook aan bij. Simpele handelingen zoals het gebruik van een creditcard of het plegen van een telefoontje zijn zaken die een database worden opgeslagen. In deze gigantische hoeveelheid gegevens ligt vaak waardevolle informatie verborgen.
Informatie definiëren we hierbij als volgt:

Informatie is data waaraan een betekenis is toegekend (Schreiber 1998).

Belangrijk hierbij is dat de informatie bruikbaar is en nut heeft voor de ontvanger van deze informatie. Het zal duidelijk zijn dat een groot deel van alle opgeslagen informatie nooit door mensenogen gezien zal worden. Wanneer er al überhaupt iets uit deze gigantische hoeveelheden dat gehaald zal worden, zullen computers gebruikt moeten worden voor de analyse hiervan.
Simpele statistische technieken voor het analyseren van data bestaan al lange tijd, maar blijken niet toereikend te zijn om uit de enorme databases alle mogelijke informatie te halen. Nieuwe technieken in combinatie met de steeds krachtiger en snellere computers hebben er toe geleid dat het 'schatgraven' in databases steeds beter mogelijk is. De voordelen die behaald kunnen worden zijn groot, de informatie is aanwezig, maar moet wel gevonden worden.

4.2 Enkele definities

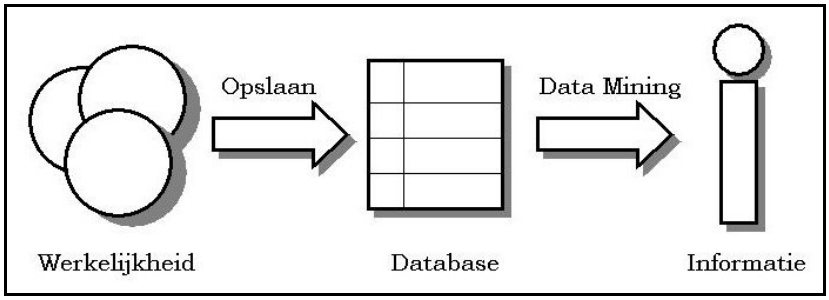
Om een omschrijving of definitie van het begrip *Data Mining* te kunnen geven, is het eerst nodig om duidelijk te hebben wat het begrip *database* precies inhoudt.

Een database is een methode om gegevens op een gestructureerde manier te bewaren, zodat de gegevens later eenvoudig terug te vinden zijn.

Er zijn verschillende manier om dit te doen. Een eenvoudige manier is de simpele kaartenbak, op elk kaartje staan gegevens die bij elkaar horen. Dankzij de opkomst van de computer zijn er methoden bedacht die veel meer mogelijkheden bieden. De meest gebruikte methode is de relationele database. In een relationele database worden gegevens in meerdere tabellen bewaard samen met de onderlinge relaties tussen de tabellen. Alle gegevens die bij elkaar horen komen in één tabel te staan.
Nu kan er een definitie van het begrip Data Mining gegeven worden, waarvan er veel verschillende te vinden zijn. Veel definities in de literatuur hebben betrekking op Data Mining in een commerciële omgeving. De volgende definitie is een ruime omschrijving die ook wetenschappelijke en persoonlijke toepassingen omvat.

Data Mining is het extraheren van voorheen onbekende informatie uit (vaak grote hoeveelheden) data (Meij 2002)

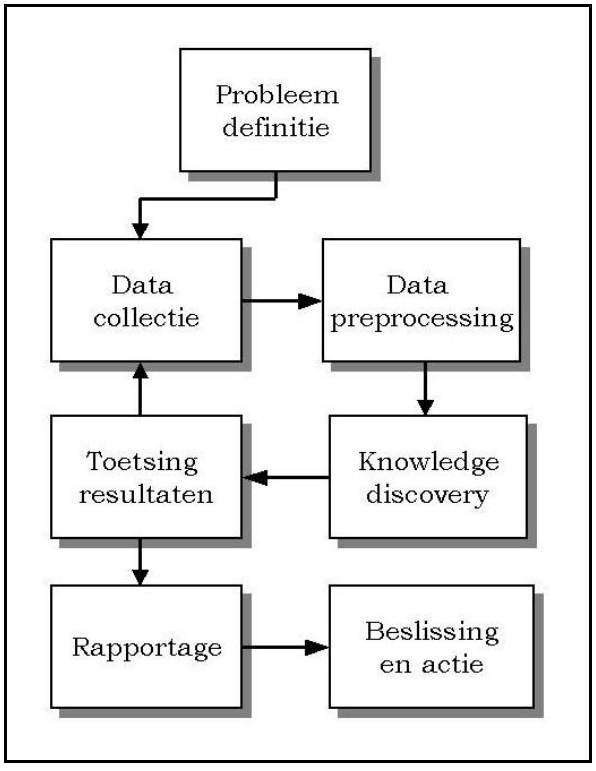
De informatie is niet van tevoren bekend, maar ligt wel impliciet in de data opgeslagen. In figuur 4.1 wordt de kern van Data Mining schematisch weergegeven; het vinden van de schat (aan informatie) in de database.



Figuur 4.1. Data Mining in vogelvlucht

4.3 Het Data Mining Proces

Het opzetten van een Data Mining omgeving heeft wel wat voeten in de aarde. De centrale database moet een soort constante informatiebron zijn waaruit steeds nieuwe informatie kan worden geput.
Het Data Minen zal moeten gebeuren volgens een specifiek proces. In figuur 4.2 wordt het Data Mining proces weergegeven. De verschillende fasen uit dit proces zullen in de volgende paragrafen nader toegelicht worden.



Figuur 4.2 Data Mining Proces

4.3.1 Probleem definitie

De eerste stap van het proces is het definiëren van de doelen, dat wil zeggen het aangeven van de informatie die men wil afleiden uit de gegevens in de database. Het 'in het wilde weg' gaan toepassen van Data Mining heeft geen enkele zin. Het is belangrijk is een duidelijke lijn uit te zetten en die te volgen. Gedurende het gehele proces kan ook steeds teruggekoppeld worden naar de probleemdefinitie. Op deze manier wordt het gevaar bestreden dat aan de doelstelling voorbij geschoten wordt. Aangezien het vooraf nooit helemaal duidelijk is wat daadwerkelijk uit de data te halen is, moet men wel uitkijken dat men zich niet aan de definitie vastpint. Tijdens het proces kan de definitie onhaalbaar blijken te zijn en zal dan aangepast moeten worden.

4.3.2 Data collectie

Na de probleemdefinitie zullen de relevante gegevens uit de organisatie moeten worden verzameld. Hiervoor is nodig om duidelijk te weten wat men precies hoopt te ontdekken. Echter ook gegevens waarvan niet direct de relevantie wordt gezien zullen bekeken moeten worden, zodat eventuele onbekende verbanden niet over het hoofd worden gezien. Hierbij dient opgemerkt te worden dat het verzamelen van relevante gegevens zeker geen eenvoudige zaak is. De gegevens zijn vaak verspreid over verschillende (grote) databases en staan vaak in verschillende formaten en het komt regelmatig voor dat niet eens duidelijk is wat de gegevens precies voor betekenis hebben.

4.3.3 Data preprocessing

De volgende stap is het gereedmaken van de data voor gebruik.
De gegevens zoals ze in databases zijn opgeslagen zijn vaak 'vuil'. Dit heeft over het algemeen niets te maken met nalatigheid of luiheid van degenen die de data invoeren. Er kunnen problemen zijn met ontbrekende of verkeerd ingevulde records en met de actualiteit van de gegevens. Aangezien vervuiling zeer negatieve invloed kan hebben op de uiteindelijke resultaten van het Data Mining proces, zal dus eerst afgerekend moeten worden met deze vervuiling, de data dient opgeschoond te worden.
Een andere subfase van het preprocessen van de data is het coderen. Met het coderen wordt eigenlijk een soort contextinformatie in de gegevens geplaatst. Met kan namelijk niet verwachten dat de te gebruiken Data Mining tool alle mogelijke interessante informatie vindt; er zal een bepaalde zoekrichting bepaald moeten worden.
Stel dat men geïnteresseerd is in het vinden van bepaalde trends uit de data. Om deze trends te kunnen vinden zullen de gegevens als tijdsreeksen gecodeerd moeten worden. Een andere vraag zou kunnen zijn of bepaalde seizoensinvloeden aanwezig zijn. In dit geval zou men de gegevens direct naar seizoenen kunnen coderen. Op deze manier wordt de nodige achtergrondkennis ingebracht.

Nu is de data in principe geschikt voor het eigenlijke onderzoek. Over deze preprocessing fase moet zeker niet te licht gedacht worden, het goed uitvoeren van deze fase kost vaak enorm veel tijd.

4.3.4 Knowledge Discovery

In deze fase vindt eigenlijk het echte onderzoek van de gegevens plaats. Allereerst zal een geschikte tool moeten worden gezocht om de Data Mining taak te gaan uitvoeren. Hierbij is het vooral van belang welke techniek of technieken gebruikt zullen worden om de patronen in de gegevens te ontdekken. Afhankelijk van de output (het doel) en de input (de invoergegevens) zal een keuze gemaakt moeten worden. Voorbeelden van bekende technieken zijn: regel inductie, statistische methoden (zoals regressie analyse), neurale netwerken, beslissingsbomen en genetische algoritmen. Een aantal technieken zal verder in dit verslag nog besproken worden. Nadat de keuze met betrekking tot de te gebruiken tool gemaakt is, kan de tool op de data worden losgelaten en kunnen de resultaten worden verzameld.

4.3.5 Toetsing resultaten

Zodra de resultaten van de vorige stap verzameld zijn, kunnen ze getoetst worden. Dat wil zeggen dat er wordt gekeken of de uitkomsten wel aan bepaalde normen voldoen. Hierbij kan gedacht worden aan voorwaarden met betrekking tot nauwkeurigheid, betrouwbaarheid en leesbaarheid van de gevonden informatie. Wanneer de resultaten onder de norm blijken te zijn, moet weer worden terug gegrepen naar de fase van de datacollectie; er is sprake van een duidelijke terugkoppeling om tot optimale resultaten te komen.
Pas wanneer de resultaten aan de gestelde normen voldoen, kan verder worden gegaan met de volgende stap in het Data Mining proces.

4.3.6 Rapportage

Nadat de resultaten bevredigend zijn bevonden, zal hierover gerapporteerd moeten worden. Hierbij is het van groot belang dat goed duidelijk is wat er precies uit de rapportage naar voren dient te komen. Het is namelijk goed mogelijk dat er voor het ene project regels of procedures gevonden en beschreven dienen te worden, terwijl het bij het andere project de bedoeling is om meer begrip over of inzicht in het onderzochte gebied te krijgen. Verder kan men kiezen voor een gedetailleerde beschrijving van de resultaten of een minder formele beschrijving waarbij een eenvoudige presentatie van de resultaten wordt gegeven. Het voordeel van de tweede aanpak is dat het voor een breder publiek te begrijpen is. Er moet hierbij wel op gelet worden dat de vertaalslag van 'ruwe' resultaten naar 'leesbare' resultaten wel met de nodige zorgvuldigheid gebeurt. Een klein onjuistheid in de rapportage naar het management kan tot grote gevolgen leiden.

4.3.7 Beslissingen en acties

De laatste fase is het nemen van beslissingen en ondernemen van acties aan de hand van de in de rapportage beschreven resultaten met betrekking tot de probleemdefinitie. Hierbij kan gedacht worden aan het bijstellen van prognoses, aanpassen van productie of het doen van investeringen in een bepaald gebied. Hierbij moet men niet blind varen op de resultaten van het Data Mining onderzoek en gewoon het gezonde verstand blijven gebruiken. Het is beter om de uitkomsten van het onderzoek als ondersteunend hulpmiddel te zien.

4.4 De Data Mining taken

Het Data Mining proces kan verschillende doelen hebben. De meest voorkomende doelen, ook wel taken genoemd, zullen in deze paragraaf besproken worden. Deze taken kunnen worden opgedeeld in twee hoofdgroepen, namelijk *verification-driven* Data Mining taken en *discovery-driven* Data Mining taken.

Bij de verification-driven taken probeert de gebruiker een door hem opgestelde hypothese te toetsen door bepaalde informatie uit de gegevens te halen. Afhankelijk van het resultaat hiervan zal de gebruiker de hypothese accepteren, waarna het proces stopt, of zal er opnieuw informatie uit de gegevens opgevraagd worden om zo toch tot verificatie van de gestelde hypothese te komen. Aan het eind van dit proces zal dat, ofwel de hypothese geaccepteerd zijn, ofwel de gebruiker tot de conclusie komen dat de hypothese niet opgaat voor de betreffende dataset.

Bij de discovery-driven taken zijn er eigenlijk vooraf geen hypothesen opgesteld, maar wordt het vinden daarvan overgelaten aan een Data Mining tool. Een Data Mining techniek wordt toegepast op de gegevens en de uitvoer die daar resultaat van is, wordt dan nader bekeken om de gewenste informatie daar uit te selecteren. Bij deze taak speelt de gebruiker duidelijk een minder actieve rol dan bij de verification-driven taak.

4.4.1 Verification-driven taken

Binnen de verification-driven zijn de drie belangrijkste taken: Query en rapportage, multidimensionale analyse en statistische analyse. Hieronder volgt een korte beschrijving van deze taken.

Query en rapportage
Query en rapportage is de meest simpele vorm van Data Mining. Het doel van deze taak is het valideren van een door de gebruiker opgestelde hypothese. Een voorbeeld van zo'n hypothese is: "in de winter worden meer handschoenen verkocht dan in de zomer." Het valideren van deze hypothese met behulp van query en rapportage gebeurt door het creëren van een query of verzameling queries, die het best de gestelde hypothese vertolkt, de query stellen van de database en het analyseren van de uitkomsten om vast te stellen of de hypothese wordt ondersteund of niet. Elke interpretatie van de uitkomsten van een query kan weer leiden tot additionele queries. Door de queries vast te leggen in rapporten, kunnen analyses automatisch herhaald worden op vastgestelde tijden.

Multidimensionale analyse
Query en rapportage is toereikend voor een groot deel van de verification-driven taken. In sommige domeinen vereist effectieve Data Mining echter het creëren van erg complexe queries, waarbij vaak de dimensie tijd om de hoek komt kijken. Een voorbeeld: de regionale manager van een warenhuisketen kan vragen om de verkoopcijfers per week van het eerste kwartaal van 2001 van zijn winkels in het westen van het land. Multidimensionale databases organiseren de gegevens volgens vastgestelde dimensies, zoals tijd, gebied, etc., en zijn voorzien van de mogelijkheid om queries 'langs' dimensies uit te voeren. Ook laten deze databases het toe de gegevens hiërarchisch naar dimensie te organiseren, met samenvattingen

op de hogere niveaus en de eigenlijke data op de lagere niveaus. Met behulp van deze multidimensionale databases is het dus mogelijk multidimensionale analyse uit te voeren en de complexere queries te beantwoorden.

Statistische analyse
Bij de beide bovenstaande taken wordt gebruik gemaakt van eenvoudige statistische operaties. Wanneer de hypothesen complexer worden, vereist de verificatie van deze hypothesen meer geavanceerde statistische methoden. Om effectief te zijn, dient de statistische analyse gebaseerd te zijn op een bepaalde methodologie. Deze vorm van Data Mining wordt vanwege de gegronde statistische kennis die het vereist niet erg vaak toegepast binnen bedrijven, maar meer op het gebied van onderzoek en ontwikkeling.

4.4.2 Discovery-driven taken

Binnen de discovery-driven zijn ook een aantal belangrijke taken te onderscheiden. De vijf belangrijkste zijn: *associatie*, *sequentiële patronen*, *classificatie*, *clustering* en *scoring*. Hieronder weer een korte beschrijving van deze taken.

Associatie
Een associatie functie is een operatie die op zoek gaat naar relaties tussen attributen uit een verzameling records. Deze relaties kunnen uitgedrukt worden in regels, zoals "80 % van alle records die de attributen A en B bevatten, bevat ook het attribuut C." Hierbij wordt het genoemde percentage de zekerheidsfactor van de regel genoemd. Waar het bij associatie dus op neer komt, is dat de aanwezigheid van bepaalde eigenschappen de aanwezigheid van andere eigenschappen direct impliceert.
Associatie wordt veel gebruikt in verkooporganisaties. De analyse wordt gebruikt om te kijken in hoeverre het verkoopbeleid effectief is, en hoe dit aangepast dient te worden voor verbetering. Er kan gedacht worden aan een regel als: "als een klant meer dan drie zakken chips koopt, zal in 80% van de gevallen ook een fles cola worden gekocht." Aan de hand van deze informatie kan dan het verkoop- en marketingbeleid worden af- en bijgesteld.

Sequentiële patronen
In het hierboven genoemde voorbeeld van de verkooporganisatie bevat de database over het algemeen geen informatie over de identiteit van de klant. Als deze informatie wel voorhanden is, zou het mogelijk zijn een analyse te maken van de transactie toegespitst op één bepaalde klant. Er kan dan op de specifieke interesses en het aankoopgedrag van de individuele klant worden ingespeeld. Een functie voor sequentiële patronen zal een verzameling van records die betrekking hebben op één bepaalde klant analyseren, en zal daarin vaak voorkomende patronen van in de loop der tijd gekochte producten ontdekken.

Classificatie
Classificatie kan toegepast worden wanneer een dataset bestaat uit een verzameling records, ieder bestaand uit een aantal attributen, een verzameling klassen (of labels) en een relatie tussen een record en een klasse. Een classificatie functie analyseert de gelabelde records en produceert een beschrijving van de kenmerken van de records die tot de verschillende klassen behoren. Er wordt dus een soort karakterschets gegeven van de gedefinieerde klassen. Deze karakterschets die door de classificatie wordt gegenereerd kan expliciet (bijvoorbeeld een verzameling regels die iedere klasse beschrijven) of impliciet zijn (bijvoorbeeld een wiskundige functie die als invoer een record neemt en als uitvoer de bijbehorende klasse

geeft). De gevormde klassenbeschrijvingen kunnen nu dus gebruikt worden om nieuwe records waar de bijbehorende klasse nog niet van bekend is te classificeren. In de loop van de tijd zijn er vele classificatie modellen ontwikkeld. Een aantal hiervan zal later nog besproken worden.
Classificatie functies zijn in veel applicaties toegepast, zoals kredietrisico analyse of gezondheidsrisico analyse, maar ook op het gebied van bijvoorbeeld beeld- en spraakherkenning.

Clustering
Bij classificatie is de klasse van de records die als invoer worden gebruikt bekend. De invoer bij clustering bestaat uit een collectie van ongelabelde records. Hier zijn dus geen klassen bekend op het moment dat de clustering operator wordt toegepast. Bij clustering worden de klassen geproduceerd aan de hand van een aantal criteria. Hierbij wil men een zo groot mogelijke samenhang tussen de records binnen een bepaalde klasse en zo min mogelijk samenhang tussen de records van verschillende klassen. De klassenbepaling hangt af van de keuze van de criteria.
Als voorbeeld van het gebruik van clustering zou gedacht kunnen worden aan het indelen van het klantenbestand van een verzekeringsmaatschappij. De maatschappij zou dan op basis van deze indeling kunnen kijken naar de specifieke kenmerken van elk segment en aan de hand daarvan een speciale polis voor iedere groep op kunnen stellen.

Scoring
Bij scoring worden aan alle records in een database een score of waarde toegekend, door aan alle attributen een gewicht te geven en deze te wegen. Wanneer een record een hogere waarde heeft dan een andere, kan dit betekenen dat dit record beter aan een bepaalde voorwaarde voldoet dan de andere. Een voorbeeld hiervan is dat een verzekeringsmaatschappij bij een hoge score wel een bepaalde auto zal verzekeren en bij een lagere score. Voor scoring bestaan dus veel praktische toepassingen.

5. RISICO-ANALYSE KLIC MELDINGEN BINNEN KPN

Aan de hand van de stappen uit het Data Mining proces die besproken zijn in het vorige hoofdstuk zal in dit hoofdstuk de stage-opdracht besproken worden.

5.1 Probleem definitie

De probleem definitie is het eerste en zeker niet onbelangrijkste deel van de opdracht. Bij de definitie is een doelstelling gedefinieerd die het richtsnoer is voor de hele opdracht. Bij elke stap of beslissing die genomen wordt, wordt gekeken of deze stap of beslissing ertoe bijdraagt dat de gestelde doelstelling behaald wordt. In hoofdstuk 2 is de doelstelling als volgt opgesteld:

Het maken van een risicoprofiel van KLIC meldingen zodat de Toezicht Werken Derden toezicht kunnen houden bij grondroeringen (m.b.v. prioriteitsstelling) zodat de totale schadepost voor KPN geminimaliseerd wordt.

5.2 Data collectie

Nadat het probleem goed in kaart is gezet, begint het verzamelen van bruikbare data. Voor de stage-opdracht is gebruik gemaakt van data afkomstig uit verschillende informatiesystemen van verschillende afdelingen.

KLIC meldingen
Elk jaar vinden circa 200.000 grondroering plaats. Bij 120.000 van deze grondroeringen, wordt vooraf een melding gedaan bij KLIC. KLIC geeft deze melding door aan KPN. Voor de verwerking van deze KLIC meldingen wordt gebruik gemaakt van een speciaal computersysteem: Geoversum.

Geoversum wordt zowel door KPN als het landelijke KLIC gebruikt. KPN heeft een speciale KLIC afdeling waar alle meldingen binnenkomen.
Wanneer een grondroerder melding doet bij het landelijke KLIC wordt deze melding ingevoerd in Geoversum.
Vanuit KLIC worden alle meldingen waarbij KPN betrokken is, dat wil zeggen, de meldingen waarvan de grondroering plaatsvindt op een locatie waar KPN kabels heeft liggen, digitaal doorgestuurd naar de KLIC afdeling bij KPN.

Deze meldingen worden door KPN op de volgende manier verwerkt:
KPN heeft (digitale) tekeningen van al haar kabels in Nederland. Wanneer melding gedaan wordt van een grondroering, stuurt KPN tekeningen van de kabels die op de graaflocatie liggen naar de grondroerder (meestal betreft dit een aannemer).
Alle informatie over deze grondroering is opgenomen in de KLIC melding. De informatie van alle KLIC meldingen wordt opgeslagen in Geoversum. Deze informatie kan als tekstbestand verkregen worden. In bijlage 2 is een overzicht te vinden van de informatie per KLIC melding.

Met de data uit Geoversum is het één en ander aan de hand. KPN werkt nu een aantal jaar met dit systeem. In mei 2002 zijn echter een aantal dingen veranderd. Voor die tijd werden de meldingen op verschillende (decentrale) computers behandeld en opgeslagen. Er was geen centrale database waar alle informatie werd opgeslagen.
Na mei 2002 zijn alle computers die gebruik maken van het systeem aangesloten op een centrale database via het netwerk. Er is toen nog wel geprobeerd alle meldingen van vóór mei 2002, die op verschillende computers waren opgeslagen, te verzamelen. Van een aantal van deze computers is de data echter verloren gegaan of de data is opgeslagen op cd's die onleesbaar geworden zijn.
Het gevolg hiervan is dat de data van vóór mei 2002 onvolledig is.
Van rayon Zuid zijn vrijwel alle KLIC meldingen verloren gegaan. Rayon West heeft circa 40% van de meldingen teruggevonden. In rayon Oost is dit percentage circa 65%. Rayon Noord is het enige rayon waar bijna alle KLIC meldingen van voor mei 2002 beschikbaar zijn.
Tegenwoordig worden alle meldingen centraal in Amersfoort opgeslagen.
De meldingen van ná mei 2002 zijn daarom wel allemaal beschikbaar.

Schademeldingen
Bij de 200.000 grondroeringen die jaarlijks plaatsvinden, gebeurt het circa 12.000 keer dat er schade wordt gemaakt aan één of meerdere kabels van KPN.
Wanneer een kabel beschadigd wordt, wordt een schadeformulier ingevuld door de schadeveroorzaker of monteur. Dit formulier wordt later door de schadebehandelaars van KPN ingevoerd in het computersysteem: Figtree

Voor de schadebehandeling van kabels is Nederland opgedeeld in drie rayons: Noord Oost, Zuid en West. De vestigingen van de verschillende rayons zijn gevestigd in Zwolle (NO), Den Bosch (Z) en Rotterdam (W). De kabelschades worden decentraal behandeld in verschillende rayonvestigingen.
Schades met een schadebedrag boven de € 5000 en schades die juridisch gezien erg gecompliceerd blijken te zijn worden doorgestuurd naar de afdeling Juridische Zaken in Amersfoort. Zowel de rayonvestigingen als de afdeling Juridische Zaken werken met Figtree. Ook Figtree heeft de mogelijkheid om gegevens af te drukken als tekstbestand. In bijlage 2 is ook een overzicht van alle informatie die beschikbaar is wat betreft de kabelschades.

Figtree wordt gebruikt vanaf 1999. Er is dus informatie beschikbaar van alle kabelschades die gemaakt zijn in 1999 en daarna.

5.3 Data preprocessing

Op het moment dat een KLIC melding binnenkomt, is het (nog) onbekend of deze melding betrekking heeft op een grondroering waar kabelschade zal worden gemaakt of dat het een grondroering betreft die zonder kleerscheuren zal verlopen. Via een model dat gebaseerd is op KLIC meldingen uit het verleden hoopt men inzicht te krijgen in de kans dat er bij een nieuwe KLIC melding kabelschade wordt gemaakt.

Enige informatie is hiervoor wel noodzakelijk. De KLIC meldingen uit het verleden op zichzelf zeggen niet zoveel. Deze data wordt interessant wanneer van deze meldingen bekend is of het gaat om een grondroering mét of zonder kabelschade.

Om deze informatie boven water te krijgen, moet er een koppeling gemaakt worden tussen de data met alle KLIC meldingen en de data met alle kabelschades. De systemen Geoversum en Figtree moeten aan 'aan elkaar gekoppeld' te worden. Bij elke kabelschade zal geprobeerd worden de KLIC melding te vinden van de grondroering waar de kabelschade werd gemaakt.

Zoals al eerder is opgemerkt vinden jaarlijks 200.000 grondroeringen plaats, waarbij 120.000 keer 'geKLICt' wordt. Er zijn elk jaar dus circa 80.000 grondroeringen waarvan geen melding wordt gedaan bij KLIC.
Van een deel van de kabelschades zal dus nooit een KLIC melding teruggevonden worden, omdat deze simpelweg nooit is gedaan.
Het aantal kabelschades waar KPN mee te maken heeft, bedraagt circa 12.000 per jaar. Er geldt dus dat (gelukkig) slechts een klein deel van de grondroeringen te maken heeft met kabelschade.

In de overzichten uit bijlage 2 zien we dat zowel de KLIC melding als het schaderecord een veld bevat waar het KLIC nummer vermeld staat. Dit veld zou dus makkelijk gebruikt kunnen worden om de KLIC melding van de grondroering waar de schade werd gemaakt, terug te vinden. Helaas is dit KLIC nummer slechts terug te vinden op een klein deel (circa 5%) van de schadeformulieren.
Er is dus geen sleutelveld waarmee de beide systemen makkelijk te koppelen zijn. Er zal een andere manier gevonden moeten worden om de KLIC melding terug te vinden bij een kabelschade. De methode die gebruikt is voor de koppeling van Geoversum en Figtree is beschreven in hoofdstuk 7.

5.4 Knowledge discovery

Met de informatie die nu voorhanden is, kan gezocht worden naar een geschikte Data Mining techniek. Voordat een techniek gekozen wordt, is het goed om goed voor ogen te hebben wat voor informatie we precies uit de data willen halen.
Het is belangrijk om duidelijk te stellen wat we precies willen hebben, hoe willen we dat de uitvoer van het traject eruit zal zien?

De nieuwe KLIC meldingen willen we als een gesorteerde lijst meldingen kunnen weergeven, met bovenaan de lijst de meldingen die een hoog risico hebben als het gaat om kabelschade en onderaan de meldingen die minder risicovol zijn.

Het probleem waar we mee te maken kan op meerder manieren bekeken worden. Het kan zowel als een classificatieprobleem als een scoringsprobleem gezien worden.

Als we het probleem behandelen als een classificatieprobleem, wil dat zeggen dat we de meldingen die binnenkomen willen classificeren naar verschillende klassen. De klassen die we bij deze opdracht hanteren zijn voor de hand liggend, namelijk: **wél schade** en **géén schade**. Meldingen met een kleine kans op kabelschade zullen geclassificeerd worden in de klasse 'geen schade'. De overige meldingen zullen het label 'wel schade' krijgen.

De andere mogelijkheid is om het probleem als scoringsprobleem te zien. In dit geval willen we aan elke melding een score of waarde toekennen, door alle attributen een gewicht te geven en deze te wegen. Meldingen met een hoge waarde staan voor risicovolle meldingen, terwijl lage waarden juist een lage kans op kabelschade hebben.

Het is ook nog mogelijk om het probleem te zien als een combinatie van de twee verschillende soorten problemen: elke melding krijgt een score of waarde toegekend. Meldingen met een waarde boven een bepaalde grens zijn de risicovolle meldingen en worden geclassificeerd in de klasse 'wél schade'. De meldingen die een score krijgen onder deze grens worden geclassificeerd in de klasse 'géén schade'.

Voor het analyseren van een aantal verschillende technieken die mogelijk geschikt zijn voor de opdracht, is gebruik gemaakt van het software pakket WEKA.
WEKA is een collectie van algoritmes die gebruikt kunnen worden voor het oplossen van data mining problemen. WEKA heeft als groot voordeel dat met weinig voorwerk verschillende technieken toegepast kunnen worden. Een ander voordeel is de prijs van het pakket, het is namelijk gratis verkrijgbaar.
Het programma heeft ook een aantal nadeel. Eén hiervan is de snelheid. Het programma is geschreven in de programmeertaal Java, wat bepaald niet één van de snelste talen is. Doordat het aantal berekeningen dat bepaalde technieken met grote datasets met zich meebrengen soms wel in de miljarden kan lopen, heeft dit als gevolg dat afhankelijk van de processorsnelheid van de computer de runtimes enorm kunnen oplopen.
Een ander belangrijk nadeel is de bewerkbaarheid en inzichtelijkheid van de methodes. Het is lastig om duidelijk te krijgen wat er precies gebeurd bij sommige methodes, ook is het nauwelijks mogelijk zelf dingen aan te passen.

Voor het toetsen van de verschillende technieken is gebruik gemaakt van twee datasets: de *trainingdata* en de *testdata*.
De trainingset bevat alle beschikbare KLIC meldingen van 2001. Met behulp van de in paragraaf 5.3 beschreven koppeling is voor deze KLIC meldingen bepaald of dit meldingen mét of zónder kabelschade waren. Van deze meldingen is dus bekend tot welke klasse ('wel schade' of 'geen schade') ze behoren.
De testdata bevat alle KLIC meldingen van januari, februari en maart van 2002. Ook voor deze meldingen is met behulp van de koppeling bepaald of dit meldingen mét of zónder kabelschade waren.
Op basis van de trainingdata is met behulp van Data Mining technieken een classificatie of score berekend voor alle KLIC meldingen uit de testdata. Omdat de klasse van de testdata dankzij de koppeling al bekend is, kan nu gekeken worden hoe goed de Data Mining techniek de meldingen uit de testdata classificeert.

Uit verscheidene tests bleek dat de Kernel Dichtheid Classificatie bij dit probleem goede resultaten gaf. Deze methode wordt verder besproken in bijlage 3.5. De Kernel Dichtheid Classificatie berekent steeds een 'afstand' tussen een KLIC melding uit de testdata en een KLIC melding uit de trainingdata. Deze afstand is altijd een waarde tussen 0 en 1. KLIC meldingen uit de trainingset met een kleine afstand tot de melding uit de testset hebben grote invloed op de uiteindelijke classificatie van de melding uit de testset.
Wanneer de twee KLIC meldingen erg op elkaar lijken, levert dit een kleine afstand op. Het kan hier zijn dat de meldingen in een zelfde stad of plaats zijn gedaan of dat het bij beide meldingen om dezelfde soort werkzaamheden gaat.

De Kernel Dichtheid Classificatie doet vooral goede zaken wanneer attributen numeriek zijn, dat wil zeggen dat het attribuut een getal is. Tussen twee getallen is makkelijk een afstand te berekenen. Je kunt zeggen dat de afstand tussen de getallen 4 en 5 kleiner is dan de afstand tussen de getallen 3 en 7. Bij niet-numerieke attributen is dit lastiger. Stel je hebt een attribuut

dat de naam van een meubelstuk bevat. Het is moeilijk om een afstand tussen een stoel en een tafel te bepalen. Is deze groter of kleiner dan de afstand tussen een schemerlamp en een boekenkast?
De Kernel Dichtheid Classificatie behandelt niet-numerieke attributen als volgt: wanneer de attributen gelijk zijn, is de afstand 0, bij ongelijkheid is de afstand 1.

Een KLIC melding bestaat uit meerdere velden of attributen.
Bij records met meerdere attributen gaat de Kernel Density Schatter op de volgende manier te werk: wanneer alle attributen van twee records allemaal aan elkaar gelijk zijn, is de afstand 0, wanneer er ook maar één attribuut ongelijk is, wordt direct de maximale afstand van 1 toegekend.
Alleen KLIC meldingen uit de trainingset die volledig gelijk zijn aan de KLIC melding uit de testset krijgen dus een afstand van 0. Dit betekent dat slechts de KLIC meldingen die volledig gelijk zijn invloed uitoefenen op de classificatie van de melding uit de testset.
Dit betekent voor ons probleem dat bij het voorspellen van de schade voor een bepaalde KLIC melding alleen naar de meldingen uit het verleden wordt gekeken die identiek zijn aan de nieuwe melding. Omdat we ook de invloed willen weten van meldingen die niet helemaal hetzelfde zijn, maar waar bijvoorbeeld de aannemer of plaatsnaam wel gelijk is, is er een methode ontwikkeld die wel het principe van de Kernel Dichtheid Classificatie hanteert, maar die de afstand op een andere manier berekent. Mee hierover is te lezen in hoofdstuk 6.

5.5 Toetsing resultaten

Resultaten van zowel de koppeling tussen Geoversum en Figtree als het uiteindelijke model zijn beschreven in hoofdstuk 7.

Het is moeilijk om te zeggen wanneer de resultaten voldoende bevredigend zijn.
Allereerst is er geen vergelijkend materiaal waarmee de resultaten vergeleken kunnen worden om zo tot een kwaliteitsoordeel te komen, omdat er in het verleden nog niet eerder een risico-analyse is uitgevoerd.
Verder is er weinig informatie beschikbaar over de grondroeringen die de TWD's in het verleden bezocht hebben. Er kan dus geen vergelijking gemaakt worden tussen de selectie dat voorkomt uit het model en de selectie die gemaakt worden door de TWD (zonder wiskundige en Data Mining hulpmiddelen). De TWD bezoekt de grondroeringen die hij risicovol acht. Hij maakt eigenlijk gebruik van een model dat gebaseerd is op zijn ervaringen en gezond verstand.

Doordat er weinig informatie beschikbaar is over de grondroeringen die de TWD's in het verleden bezocht hebben, is er ook weinig bekend over de invloed van een bezoek van een TWD op de schadekans van een grondroering.

Om toch een kwaliteitsoordeel te kunnen geven aan de resultaten, introduceren we het begrip 'verdikkingsfactor'. Deze factor geeft aan in hoeverre het model een betere voorspelling geeft dan wanneer we een willekeurige set meldingen zouden nemen.

Als voorbeeld nemen we een set van 10.000 meldingen waarvan er 1000 (= 10%) zijn waar schade blijkt te zijn.
We kiezen een selectie van de 500 risicovolste meldingen hebben.

In het geval van een willekeurige selectie is het verwachte aantal meldingen met schade binnen deze selectie 50 (10% * 500).
Wanneer we gebruik maken van het model, zullen de 500 meldingen met de hoogste risico-factor geselecteerd worden. Als van deze 500 meldingen 200 meldingen schade blijken te hebben, betekent dit dat het model 40% van de voorspelde schades goed zat, en levert dit een verdikkingsfactor op van 40/10 = 4.

5.6 Rapportage

Een duidelijke rapportage van de resultaten is erg belangrijk. Leesbaarheid en begrijpelijkheid zijn hierbij sleutelwoorden. De rapportage van de resultaten komt uitgebreid aan de orde in hoofdstuk 7.

5.7 Beslissingen en acties

Naar aanleiding van de resultaten zijn een aantal beslissingen genomen over te nemen acties. Zo worden de KLIC meldingen voortaan tweemaal per week naar de rayons doorgestuurd in plaats van éénmaal.

Deze acties zijn samen met de conclusies van het onderzoek beschreven in hoofdstuk 8.

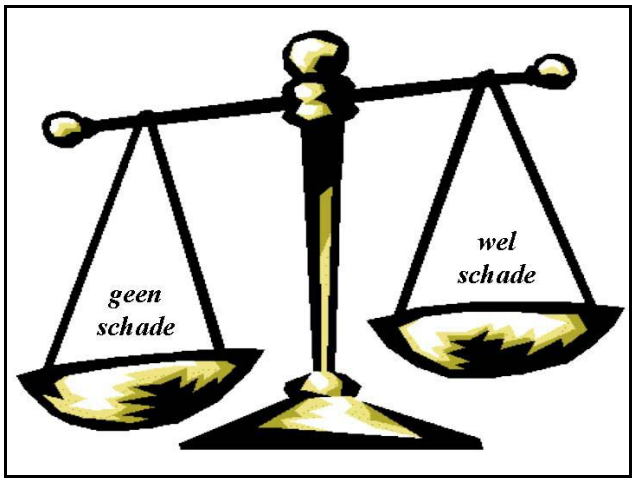
6. HET MODEL

De data is verzameld, opgeschoond en gebruiksklaar gemaakt voor toepassing van een geschikte Data Mining techniek.

Analyse van een aantal Data Mining technieken heeft uitgewezen dat de Kernel Dichtheid Schatter, beschreven in bijlage 3.5, goede resultaten gaf. Uiteindelijk is een nieuwe techniek ontworpen die gebruik maakt van het basisidee van de Kernel Dichtheid Schatter. Een nieuwe KLIC melding krijgt een risicofactor toebedeeld op basis van historische meldingen die op de nieuwe melding lijken.

6.1 Het basisprincipe: de weegschaal

Het model kan het best worden vergeleken met een weegschaal.
Een nieuwe KLIC melding wordt vergeleken met meldingen die in het verleden gedaan zijn. Van de meldingen uit het verleden is dankzij de koppeling bekend of het een grondroering betreft waar schade werd gemaakt of niet. Alle meldingen die op één of andere manier overeenkomstighe(i)d(en) hebben met de nieuwe melding brengen een gewicht op de weegschaal.



Figuur 6.1 Het principe van een weegschaal

Meldingen waar géén schade was brengen hun gewicht op één van de twee schalen (geen-schade-schaal), terwijl de andere schaal (wel-schade-schaal) gevuld wordt met de gewichten van de meldingen waar wél schade was. Meldingen die een grote overeenkomst met de nieuwe melding hebben, brengen een groot gewicht op de schaal, terwijl meldingen met kleine overeenkomsten minder invloed hebben op het totaalgewicht.

Bij het vergelijken van twee KLIC meldingen wordt naar drie volgende velden gekeken:
- de plaatsnaam,
- de aard van de werkzaamheden,
- de aannemer

Wanneer alle drie de velden van twee KLIC meldingen dezelfde informatie blijken te bevatten, hebben we te maken met twee meldingen die erg op elkaar lijken en zal een groot gewicht worden toegekend.

Meldingen kunnen ook overeenkomstigheden vertonen zonder dat alle drie de velden gelijk zijn. Er is dan sprake van een KLIC melding waarbij één of twee van de drie velden gelijk zijn.

In het geval van één gelijk veld zijn er drie mogelijkheden:
- de meldingen zijn gedaan door dezelfde aannemer,
- beide meldingen betreffen dezelfde soort werkzaamheden of
- de meldingen zijn van grondroeringen in dezelfde stad of dorp.

Ook wanneer er twee velden gelijk zijn, zijn er drie mogelijkheden:
- De aannemer en de aard werkzaamheden zijn gelijk
- De aannemer en de plaatsnaam zijn gelijk
- De plaatsnaam en de aard werkzaamheden zijn gelijk.

In het laatste en verreweg meest voorkomende geval zijn er geen overeenkomstigheden tussen de twee KLIC meldingen: alle velden zijn verschillend.

In totaal zijn er dus acht mogelijkheden. In het laatste genoemde geval zijn er geen overeenkomsten en zal er geen gewicht worden toegekend. Elk van de overige zeven mogelijkheden krijgt zijn eigen gewicht. De grootte van elk van deze gewichten worden bepaald aan de hand van een optimalisatiemethode, die later in dit hoofdstuk aan de orde zal komen.

Nadat alle KLIC meldingen uit een bepaalde periode uit het verleden zijn vergeleken met een nieuwe KLIC melding, zijn er twee totaalgewichten op de weegschaal. Het totaalgewicht op de wel-schade-schaal (in verhouding tot de som van beide totaalgewichten) bepaalt de risicofactor. Deze factor zal altijd tussen de 0 en de 1 liggen.

6.2 Praktijkvoorbeeld

De werking van het model kan het beste duidelijk worden gemaakt aan de hand van een voorbeeld. In het voorbeeld wordt gebruik gemaakt van fictieve meldingen en gewichten.

Er wordt in het voorbeeld gebruik gemaakt van de volgende gewichten:

α_{pwa} = 100
α_{pw} = 13,67
α_{pa} = 40,5
α_{wa} = 34,9
α_{p} = 0,45
α_{w} = 0,38
α_{a} = 0,12

waarbij geldt:

α_{paw} is het gewicht dat wordt toegekend aan een melding waarvan zowel de p (=plaatsnaam) als de a (=aannemer) als de w (=werkzaamheden) gelijk zijn.
α_{pw} is het gewicht voor een melding met gelijke p en w, maar een ongelijke a,
α_{pa} is het gewicht voor een melding met gelijke p en a, maar een ongelijke w, etc…

Verdere gegevens zijn:

NieuweKLIC melding:

Plaats	***Aannemer***	***Werkzaamheden***
Zwolle	*Siers*	*kabels leggen*

KLIC meldingen uit het verleden:

Plaats	***Aannemer***	***Werkzaamheden***	***Schade***
Zwolle	*Heijmans*	*kabels leggen*	*Ja*
Amsterdam	*Fugro*	*bodemonderzoek*	*Nee*
Rotterdam	*Meijssen*	*kabels leggen*	*Nee*
Nunspeet	*Siers*	*Grondboring*	*Ja*
Nijkerk	*Nieboer*	*tuin aanleggen*	*Nee*
Zwolle	*Siers*	*wegwerkzaamheden*	*Nee*

De nieuwe melding wordt met elk van de meldingen uit het verleden vergeleken.

Vergelijking met de eerste melding levert het volgende op:
- De plaatsnaam is in beide gevallen Zwolle.
- De aannemers zijn verschillend, Heijmans ≠ Siers.
- De werkzaamheden zijn gelijk, nl. kabels leggen.

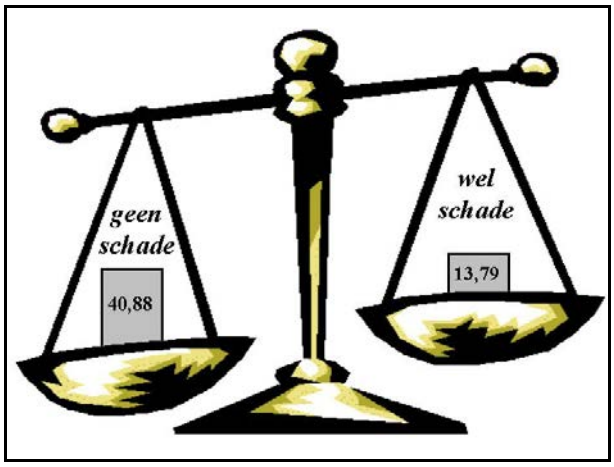
Het gewicht voor deze melding is dus: $\alpha_{pw} = 13{,}67$.
Bij de grondroering werd schade gemaakt, het gewicht wordt dus op de wel-schade-schaal gelegd.

Vergelijking met de tweede melding:
- Plaatsnaam is niet gelijk
- Aannemer is niet gelijk
- Werkzaamheden zijn niet gelijk

Geen gelijke velden → geen gewicht op de weegschaal.

Wanneer deze vergelijking wordt gedaan voor alle (zes) historische meldingen, levert dit de volgende totaalgewichten op:

GEEN SCHADE	***WEL SCHADE***
0,38 + 40,5 = **40,88**	13,67 + 0,12 = **13,79**



Figuur 6.2 voorbeeld gewichten

De risicofactor van de nieuwe melding wordt als volgt berekend: $\frac{13,79}{13,79+40,88} \approx 0,252$

6.3 Optimalisatie van de gewichten

In het model wordt gebruik gemaakt van zeven gewichten. Voor deze gewichten worden niet zomaar een aantal waarden gekozen. Met behulp van een genetisch algoritme, beschreven in bijlage 3.3, is een optimale set gewichten gegenereerd.

De optimale set is gedefinieerd als de set gewichten die bij een dataset met KLIC meldingen de risicofactors zó berekent dat de hoogste risicofactors (dit zijn de KLIC meldingen die op basis van het model door de TWD bezocht zullen worden) zoveel mogelijk terechtkomen bij meldingen waar uiteindelijk ook echt schade wordt gemaakt.
Momenteel worden 8.000 van de in totaal 120.000 KLIC meldingen bezocht door een TWD. Circa 6,7% van alle KLIC meldingen worden dus bezocht.
Er is daarom gekozen om de gewichten zó te kiezen zodat de selectie KLIC meldingen met de hoogste risicofactors (6,5% van alle meldingen) een zo groot mogelijk percentage 'schademeldingen' bevat.

Het algoritme maakt gebruik van:

- populatie omvang: N = 100,
- crossover operator: 1-punts-crossover,
- stop functie: 500 iteraties,
- fitness functie f() berekent het aantal ‘schademeldingen’ binnen 6,5% hoogste risicofactoren,
- mutatie operator: een vector $x = [x_1, ..., x_7]$ wordt vervangen door $[x'_1, ..., x'_7]$,
  - $x'_i = x_i * \sigma$.
  - σ random gekozen waarde uit het interval [0,8 ; 1,25]

6.3.1 Het algoritme

Het volgende genetische algoritme wordt toegepast bij berekening van de optimale gewichten.

1. Een selectie van 100 x 7 verschillende willekeurige gewichten wordt gekozen.

2. De 100 sets worden geevalueerd aan de hand van de fitness functie f().De kwaliteit wordt afgemeten aan het percentage schadegevallen binnen de 6,5% KLIC meldingen met de hoogste risicofactors.

3. Wanneer nog niet aan de stop criterium functie is voldaan, dit is wanneer alle stappen 500 keer zijn uitgevoerd, ga naar stap 4, anders, stop.

4. Selecteer aan de hand van de bij 2. uitgevoerde evaluatie op probabilistische wijze individuen om een mating pool van omvang 100 te genereren.

5. Kies 50 keer twee willekeurige individuen en voer een 1-punts-crossover hierop uit met kans 0,5.

6. Muteer willekeurig de inviduen uit de mating pool met kans 0,5 door middel van de gegeven mutatie operator en ga naar stap 2.

7. RESULTATEN

In dit hoofdstuk zullen de uitkomsten aan de orde komen van de verschillende trajecten van het onderzoek.

Eerst worden de resultaten besproken van de koppeling. Hier zijn de kabelschades gekoppeld aan de bijbehorende KLIC melding.
Vervolgens zullen de resultaten worden besproken van het model aangevuld met een overzicht van de optimale gewichten en de invloeden van deze gewichten.

7.1 Resultaten koppeling schade - KLIC melding

Het doel van de koppeling tussen de KLIC meldingen en de kabelschades is: duidelijk krijgen of historische KLIC meldingen horen bij grondroeringen mét of zónder kabelschade.
Met de koppeling wordt geprobeerd om bij elke KLIC melding die gedaan is in het verleden de eventuele bijbehorende kabelschade terug te vinden.

Er is bekend dat er jaarlijks zo'n 15.000 meldingen van kabelschade binnenkomen bij KPN. Van deze schades zijn er circa 6.000 (40 %) waarbij geen sprake is van een bijbehorende KLIC melding. De grondroerder heeft in dit geval geen KLIC melding gedaan door nalatigheid, onwetendheid of tijdgebrek. Soms is er helemaal geen sprake van een grondroerder, doordat de schade gevolg is van een (verkeers)ongeluk of natuurgeweld. Deze schades zijn dus ook niet aan een KLIC melding te koppelen.

Bij de overige 9.000 schades is er wel sprake van een grondroering waar een KLIC melding is gedaan. We kunnen dus zeggen dat bij zo'n 7,5% van de 120.000 KLIC meldingen sprake is van een grondroering waarbij kabelschade werd gemaakt.

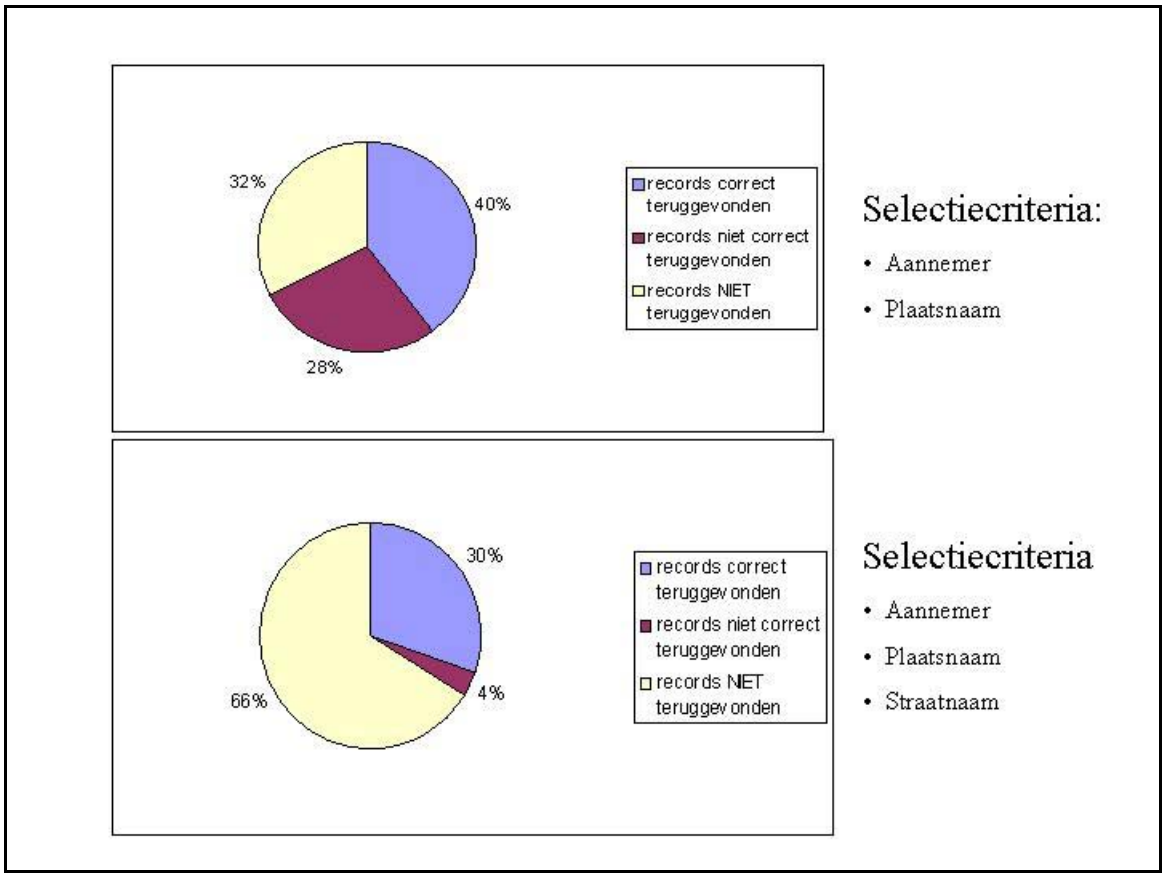
Omdat de KLIC melding en de schaderecord geen gezamenlijk sleutelveld bevatten waarmee direct gekoppeld kan worden, moest een andere manier gevonden worden om de 'juiste KLIC melding' te vinden bij een kabelschade.

Er is gekozen voor een methode die een kabelschade vergelijkt met alle KLIC meldingen uit een zelfde periode. Verschillende velden zoals adres, datum en aannemer worden vergeleken. De KLIC melding waarvan deze velden overeenkomen met die van de kabelschade wordt geselecteerd.

Nu zijn er veel mogelijkheden om te koppelen. Welke velden (en hoeveel velden) moeten gelijk zijn voordat een KLIC melding als dé juiste KLIC melding wordt benoemd?
Er kan bijvoorbeeld gekozen worden om te koppelen op basis van slechts één gelijk veld, zeg de plaatsnaam.
De resultaten zullen in dit geval zeer onbetrouwbaar zijn, omdat de kans erg groot is dat een verkeerde KLIC melding gekozen wordt. In de meeste plaatsen worden namelijk wel meer dan één KLIC melding gedaan.
Wanneer gekozen wordt om te koppelen op basis van gelijkheid van erg veel velden, zeg plaatsnaam, straatnaam, huisnummer, aannemer en opdrachtgever, zal de betrouwbaarheid erg groot zijn. Wanneer al deze velden overeenkomen is de kans groot dat je ook wel de juiste KLIC melding te pakken hebt. Het aantal teruggevonden KLIC meldingen zal echter erg klein

zijn, omdat in een groot aantal gevallen wel één van de velden zal verschillen. Er zijn vaak meerder aannemers aan het werk bij een grondroering. De aannemer die geKLICt heeft, hoeft lang niet altijd dezelfde te zijn als de aannemer die de schade veroorzaakt heeft. Ook beslaat een grondroering vaak meerdere straten (en zeker meerdere huisnummers). Het adres van de KLIC melding is dus vaak anders dan het adres waar de kabel is beschadigd.

Het spanningsveld tussen aan de ene kant de betrouwbaarheid van de resultaten en aan de andere kant het aantal teruggevonden KLIC meldingen wordt geïllustreerd in figuur 7.1.



Figuur 7.1 Resultaten bij verschillende selectiecriteria

Het bovenste cirkeldiagram geeft aan wat de resultaten zijn wanneer geselecteerd wordt op basis van gelijkheid van twee velden: Aannemer en plaatsnaam. Wanneer meerder KLIC meldingen voldoen aan deze criteria wordt de KLIC melding gekozen die qua datum het dichtst bij de schadedatum ligt.
Het onderste diagram geeft de resultaten bij een selectie op gelijkheid van drie velden: Aannemer, plaatsnaam en straatnaam. Ook hier wordt de KLIC melding gekozen die qua datum het dichtst bij de schadedatum ligt.

Voor het verkrijgen van de bovenstaande resultaten is gebruik gemaakt van 244 kabelschades waarvan de KLIC melding met zekerheid bekend was omdat het KLIC nummer ingevuld was op het schadeformulier. Hier kon dus gecontroleerd worden of de koppeling op basis van de selectiecriteria de juiste KLIC melding terugvond die op het schadeformulier stond vermeld.

We zien dat in het onderste geval (selectie op drie velden) de betrouwbaarheid van de teruggevonden KLIC meldingen vrij groot is: 30 van de 34 teruggevonden KLIC meldingen zijn juist. Het percentage niet-teruggevonden KLIC meldingen is echter ook erg groot, namelijk 66%.

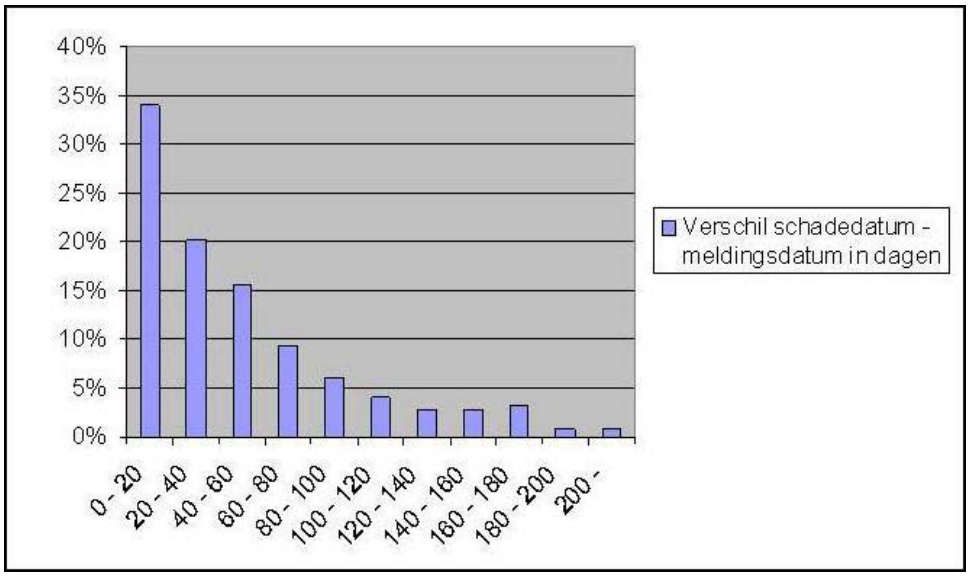
Bij de selectie op twee velden is het percentage niet-teruggevonden meldingen lager, namelijk 32%, maar de betrouwbaarheid is daarentegen ook lager: 40 van de 68 (40+28) teruggevonden KLIC meldingen zijn juist.
In KLIC melding te gaan met het KLIC nummer dat vermeld stond op het schadeformulier. Een andere KLIC melding werd teruggevonden, waarvan de plaatsnaam en aannemer ook overeen kwamen met die van het schadeformulier. Deze KLIC melding lag qua datum dichter bij de schadedatum dan de KLIC melding die op het schadeformulier werd vermeld.
Hier werd dus een KLIC melding geselecteerd en aan de schade gekoppeld die in principe schadevrij was. De KLIC melding van de grondroering waar de schade in werkelijkheid werd gemaakt wordt hierdoor als schadevrij aangezien.

Uiteindelijk is gekozen voor de bovenste optie waar geselecteerd wordt op twee selectiecriteria: aannemer en plaatsnaam. Reden hiervan is dat het aantal niet-teruggevonden KLIC meldingen bij de andere optie erg hoog is (66%). Verder geldt dat de KLIC meldingen die onterecht werden teruggevonden (28%) hetzelfde karakter hebben als de KLIC meldingen die op het schadeformulier stonden vermeld en dus eigenlijk teruggevonden hadden moeten. De aannemer én plaatsnaam van deze KLIC meldingen waren namelijk hetzelfde en ook de soort werkzaamheden bleken na verder onderzoek in bijna alle gevallen gelijk te zijn.
Het percentage niet-teruggevonden KLIC meldingen is weliswaar ook bij deze optie vrij groot, namelijk 32%, maar bij een keuze voor een selectie op slechts één overeenkomend veld zouden de resultaten te onbetrouwbaar maken.

Het feit dat een KLIC melding helemaal niet wordt teruggevonden kan twee redenen hebben:

- De plaatsnaam in de KLIC melding is anders dan de plaatsnaam in de schaderecord. Dit gebeurt heel af en toe. Het betreft dan een grondroering in de buurt van een grensgebied tussen twee plaatsen.
- De veroorzaker van de kabelschade is een andere aannemer dan de aannemer die de KLIC melding heeft gedaan. Bij veel grondroeringen zijn meerdere aannemers betrokken. De aannemer die de melding heeft gedaan is niet altijd degene die verantwoordelijk is voor de schade.

Uit dit onderzoek kwam ook naar voren dat na het doen van een KLIC melding er nog al eens wat tijd overheen ging voordat de schade werd gemaakt of opgemerkt. Dit verklaarde ook waarom het percentage KLIC meldingen dat onterecht werd teruggevonden zo groot was. In figuur 7.2 is dit met behulp van een histogram weergegeven.

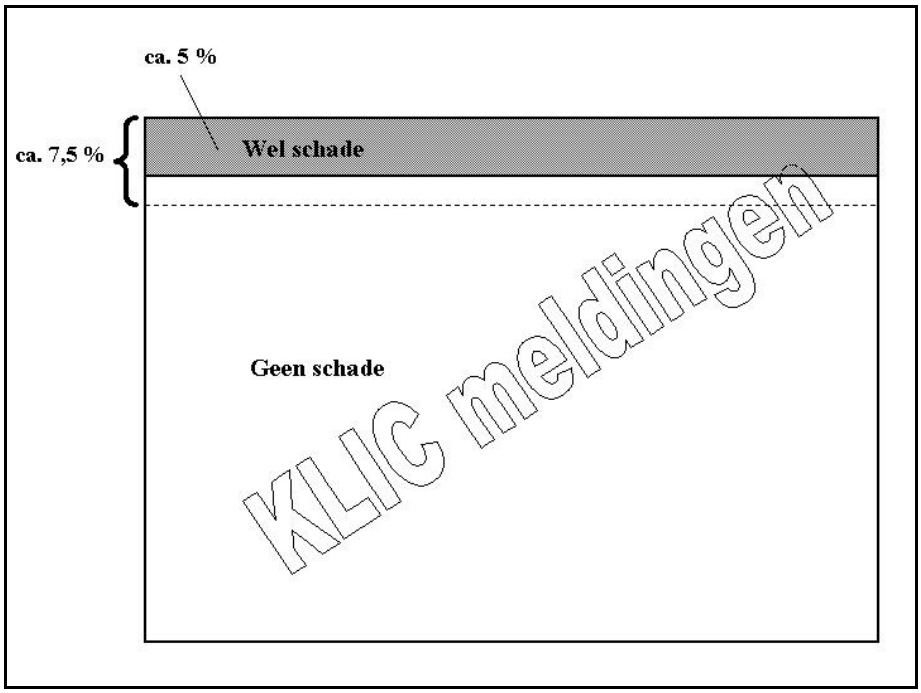


Figuur 7.2 Verschillen tussen KLIC datum en schadedatum

De schade blijkt gemiddeld circa 49 dagen later pas plaats te vinden. In één van de vijf gevallen is het verschil zelfs groter dan drie maanden.

Van alle jaarlijkse KLIC meldingen 120.000 zijn circa 9.000 KLIC meldingen die betrekking hebben op grondroeringen waar kabelschade wordt gemaakt. Deze 9.000 KLIC meldingen bedragen 7,5% van alle KLIC meldingen. Van deze 7,5% van alle KLIC meldingen worden circa 5% (68%*7,5%) gekoppeld aan de bijbehorende kabelschade.

Het resultaat van de koppeling is weergegeven in figuur 7.3. Het gebied boven de stippellijn staan voor alle KLIC meldingen van grondroeringen waar kabelschade werd gemaakt. Het grijze gebied hierbinnnen zijn de KLIC meldingen die de koppeling terug wist te vinden.



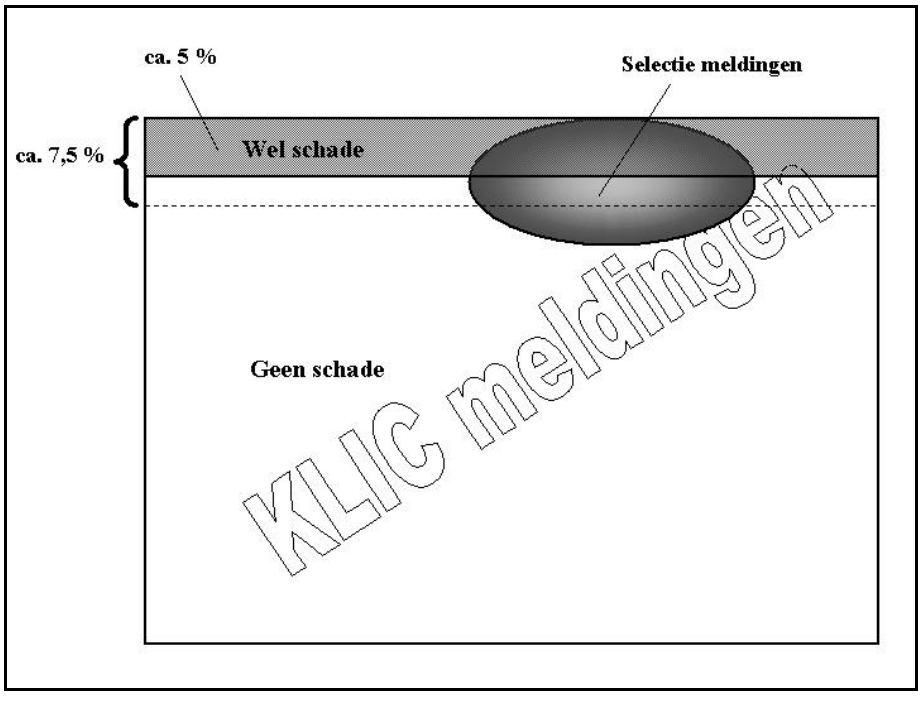
Figuur 7.3 Resultaten na de koppeling

7.2 Resultaten model

Zoals al eerder beschreven maakt het model gebruik van een dataset met KLIC meldingen uit het verleden. Van deze meldingen is door middel van de koppeling bekend geworden of het Voor nieuwe KLIC meldingen wordt op basis van deze historische meldingen een risicofactor berekend.
Om een betrouwbare risicofactor te verkrijgen is het belangrijk dat de historische meldingen regelmatig 'ververst' worden. Door de koppelingsmethode regelmatig te draaien (circa twee keer per jaar), blijft de data up-to-date.

Het model geeft iedere nieuwe KLIC melding een risicofactor. Deze factor is afhankelijk van de historische meldingen en de gewichten die gebruikt worden. De meldingen met de hoogste gewichten zijn de meest risicovolle meldingen. Deze meldingen komen boven aan de lijst te staan.
Om de kwaliteit van het model te toetsen, zal eerst een toetscriterium moeten worden gedefinieerd. Hiervoor dient de volgende vraag beantwoord te worden: Wanneer is er sprake van een kwalitatief 'goed' model?
Een goed model zal een lijst creëren waarbij de KLIC meldingen die bovenaan komen te staan (de risicovolle meldingen) ook daadwerkelijke KLIC meldingen zijn van grondroeringen waarbij schade wordt gemaakt.

Wanneer we dit vertalen naar het figuur waar de resultaten van de koppeling werden gegeven, is het dus de wens om via het model de lijst KLIC meldingen zó gesorteerd weer te geven, zodat wanneer we een selectie van KLIC meldingen met de hoogste risicofactors maken, deze selectie zo veel mogelijk binnen de ‘KLIC meldingen met schade’ zal liggen. De selectie weergegeven in figuur 7.4 willen we voor een zo groot mogelijk deel boven de stippellijn laten vallen.



Figuur 7.4 Het proces na de koppeling

De grootte van zo'n selectie kan zelf bepaald worden. De KLIC meldingen uit deze selectie verdienen de voorkeur bij het bezoeken van grondroeringen door de Toezicht Werken Derden. De grootte van de selectie zal dus afhangen van het aantal bezoeken dat de Toezicht Werken Derden af kunnen leggen.

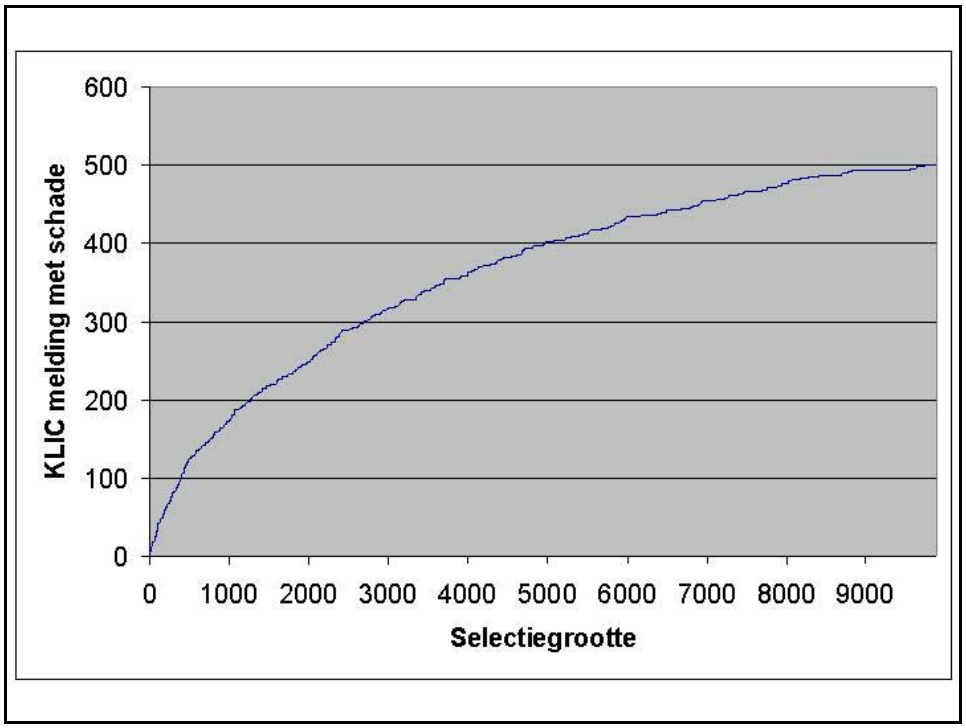
Bij het toetsen van de kwaliteit is gekeken naar de capaciteit van het huidige aantal Toezicht Werken Derden. Jaarlijks leggen de Toezicht Werken Derden circa 8.000 bezoeken af. Aangezien er per jaar circa 120.000 KLIC meldingen binnenkomen, betekent dit dat circa 6,7% van de KLIC meldingen bezocht wordt.
Bij de toetsing is daarom een selectie KLIC meldingen geselecteerd met een selectiegrootte van 6,7% van het totale aantal KLIC meldingen. Het percentage 'schademeldingen' binnen deze selectie is de kwaliteitsmeter.

De kwaliteit van het model wordt voornamelijk bepaald door de keuze van de gewichten (zie paragraaf 6.3).
Elke keer wanneer de set historische KLIC meldingen die voortkomt uit het koppelingstraject wordt ververst, wordt er bij deze KLIC meldingen ook een nieuwe set (optimale) gewichten bepaald.
Een set gewichten waarbij de selectie KLIC meldingen met de hoogste risicofactors (6,7% van alle KLIC meldingen) en zo hoog mogelijk percentage 'schademeldingen' heeft, wordt met behulp van de in paragraaf 6.3 beschreven methode gekozen.

Bij de toetsing van de kwaliteit van het model is gebruik gemaakt van alle beschikbare KLIC meldingen van 2001. Deze dataset vormde de set met historische KLIC meldingen. Met behulp van deze dataset zijn optimale gewichten bepaald.
Hierna is met behulp van deze dataset met bijbehorende gewichten, het model toegepast op alle KLIC meldingen uit de maanden januari, februari en maart van 2002.

Ook van deze meldingen was al van tevoren bekend (weer met behulp van de koppelingsmethode) of er de KLIC melding een schademelding was of niet.
De dataset met meldingen uit de maanden januari, februari en maar van 2002 bevat 9890 KLIC meldingen. Van deze 9890 KLIC meldingen waren volgens de koppelingsmethode 500 (5,1%) schademeldingen. Hierbij dient opgemerkt te worden dat er in werkelijkheid meer dan deze 500 KLIC meldingen zijn waarbij kabelschade werd gemaakt. Dit heeft te maken met het feit dat bij slechts 68% van de kabelschades (waarbij wél een KLIC melding werd gedaan) deze bijbehorende KLIC melding ook daadwerkelijk werd teruggevonden door de koppelingsmethode.
Naar schatting zullen van deze 9890 KLIC meldingen dus circa 750 (ipv 500) schademeldingen zijn.
Het is goed om dit het achterhoofd te houden bij het lezen van de rest van dit hoofdstuk.

Voor al de meldingen uit de maanden januari, februari en maart van 2002 is een risicofactor berekend. Vervolgens zijn deze KLIC meldingen (gesorteerd op risicofactor) in een lijst gezet.



Grafiek 7.1 Resultaten bij verschillende selectiegroottes

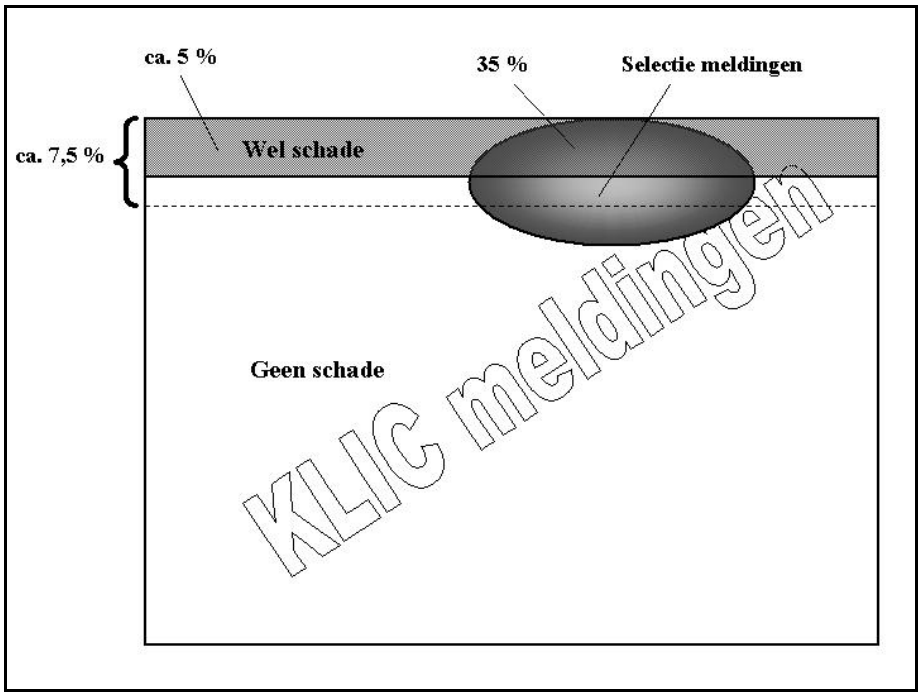
Grafiek 7.1 en tabel 7.1 geven aan hoeveel 'schademeldingen' er zich binnen een willekeurige selectie met hoogste risicofactoren bevonden: wanneer dus gekozen wordt voor een selectie van 1.000 meldingen, zullen hier (minimaal) 173 meldingen tussen zitten die een kabelschade als gevolg hadden. 173 van de 500 schademeldingen worden dus geselecteerd. Omdat in werkelijk circa 750 schademeldingen zijn binnen de 9890 KLIC meldingen, zal het aantal schademeldingen binnen deze selectie groter zijn dan 173.

Het is duidelijk uit de tabel en de grafiek af te lezen dat een kleine selectie KLIC meldingen procentueel meer schademeldingen bevat dan een grote selectie.

Selectiegrootte	Aantal schademeldingen'	Percentage
50	17	34,0%
100	35	35,0%
150	48	32,0%
200	59	29,5%
300	80	26,7%
400	101	25,3%
500	124	24,8%
750	147	19,6%
1000	173	17,3%
9890	500	5,1%

Tabel 7.1 Resultaten bij verschillende selectiegroottes

Ook dit kan weer vertaald worden naar het plaatje met de resultaten van de koppeling. Het percentage van selectie dan binnen het grijze gedeelte valt is 35% (bij een selectiegrootte van 100).



Figuur 7.5 Resultaten aan het eind van het proces

Omdat we niet weten wat dit percentage is bij de huidige bezoeken van de TWD's introduceren, is het niet mogelijk om te weten in hoeverre het model een verbetering is van de huidige situatie. Wel kunnen we bepalen in hoeverre de selectie die het model geeft 'beter' is dan een willekeurige selectie KLIC meldingen. Hiervoor gebruiken we de in paragraaf 5.5 geintroduceerde verdikkingsfactor.

Bij een willekeurige selectie is het verwachte percentage schademeldingen 5,1%.
Het model is dit percentage 35%.
De verdikkingsfactor is nu: 35/5,1 = 6,9

7.3 Resultaten gewichten

Met behulp van de parameteroptimalisatie en een set met historische KLIC meldingen wordt een set van optimale gewichten bepaald. De set bevat zeven gewichten naar het aantal mogelijke overeenkomsten tussen twee KLIC meldingen:

α_{paw}, α_{pa}, α_{pw}, α_{aw}, α_{p}, α_{a} en α_{w}

waarbij geldt:
α_{paw} is het gewicht dat wordt toegekend aan een melding waarvan zowel de p (=plaatsnaam) als de a (=aannemer) als de w (=werkzaamheden) gelijk zijn.
α_{pa} is het gewicht voor een melding met gelijke p en a, maar een ongelijke w,
α_{pw} is het gewicht voor een melding met gelijke p en w, maar een ongelijke a, etc…

De dataset van alle KLIC meldingen van 2001 gaf de volgende gewichten:

α_{pwa} = 100
α_{pw} = 0,71196
α_{pa} = 41,7777
α_{wa} = 1,38265
α_{p} = 0,06644
α_{w} = 0,00689
α_{a} = 0,03271

Uit deze gegevens valt af te lezen dat van de zeven gewichten α_w het kleinste gewicht is. Toch heeft dit gewicht meer invloed dan de gewichten α_p en α_a. Dit heeft te maken met het feit dat KLIC meldingen met gelijke soort werkzaamheden (veel) vaker voorkomt dan KLIC meldingen van gelijke aannemers of KLIC meldingen in dezelfde plaats. Het aantal verschillende soort werkzaamheden is namelijk slechts 90, terwijl er circa 12.000 verschillende aannemers zijn die KLIC meldingen doen verdeeld over een paar duizend verschillende plaatsen.

Om een beeld te krijgen wat voor invloed de afzonderlijke gewichten hebben, moeten we deze gewichten bekijken in verhouding tot de frequentie waarmee deze gewichten voorkomen.

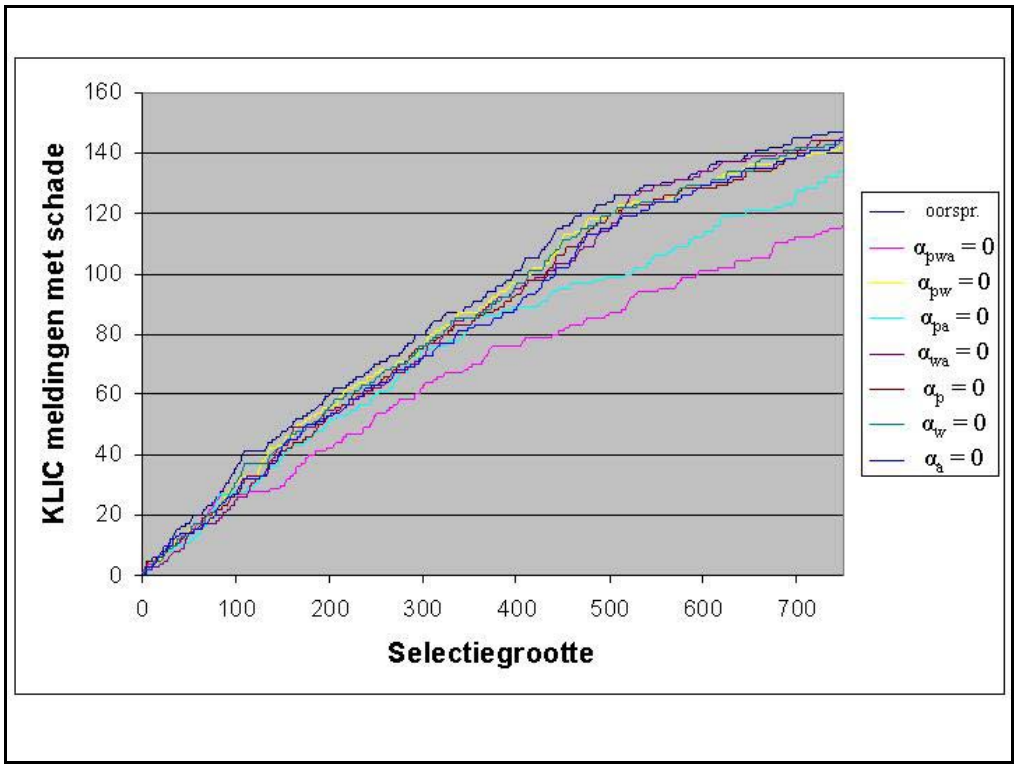
Onderstaande tabel geeft aan hoe vaak de afzonderlijke gewichten gemiddeld voorkomen wanneer een willekeurige KLIC melding vergeleken wordt met 40.000 andere willekeurige meldingen. Door deze frequentie te vermenigvuldigen met de absolute gewichten, wordt een relatief gewicht zichtbaar.

	Absolute Gewicht	***Frequentie per 40.000 vergelijkingen***	***Relatieve gewicht***
α_{pwa}	100	4,73	472,90
α_{pw}	0,712	80,19	57,09
α_{pa}	41,7777	4,53	189,17
α_{wa}	1,3827	134,26	185,64
α_{p}	0,0664	147,86	9,82
α_{w}	0,0069	3874,30	26,69
α_{a}	0,0327	163,02	5,33

Tabel 7.2 Relatieve gewichten

Er is verder onderzoek verricht naar de invloed van de afzonderlijke gewichten door te kijken wat er gebeurt wanneer één van de zeven gewichten weggelaten wordt. Dit is heel simpel te verwezenlijken door het desbetreffende gewicht gelijk aan 0 te stellen.

In grafiek 7.2 zien we het effect op de kwaliteit van de selectie wanneer één van de gewichten weggelaten wordt.
De bovenste lijn is de oorspronkelijke lijn geeft de resultaten weer waarbij alle gewichten gebruikt worden. Deze lijn wordt ook weergegeven in grafiek 7.1. De lijnen daaronder geven aan wat de kwaliteit is van het model wanneer één van de gewichten gelijk aan 0 wordt gesteld. De kwaliteit wordt gemeten aan het aantal KLIC meldingen met schade dat binnen de gekozen selectie valt. Des te lager de lijn, des te groter het verschil in kwaliteit, wat betekent dat het weggelaten gewicht grote invloed heeft op het resultaat van het model.



Grafiek 7.2 Invloeden van de individuele gewichten

Te zien valt dat de roze lijn het grootste kwaliteitsverlies oplevert. Deze lijn geeft de resultaten weer van het model waarbij α_{pwa} is weggelaten. Dit betekent dat dit gewicht (dat wordt toegekend aan meldingen waarvan zowel de plaatsnaam als de aannemer als de werkzaamheden gelijk zijn) grote invloed heeft op de kwaliteit van het model. Dit was ook wel te verwachten, want het gaat hier om (historische) meldingen die identiek zijn aan de nieuwe KLIC melding.
Na α_{pwa} blijkt α_{pa} ook met name van betekenis te zijn voor de kwaliteit.

8. CONCLUSIES, AANBEVELINGEN EN MOGELIJKE VERVOLGSTUDIES

KPN wil meer inzicht krijgen in de risico's van de KLIC meldingen. Dankzij de stage-opdracht is er meer inzicht gekomen in deze risico's.

Dit inzicht is niet zozeer uit de drukken in lijsten met risicovolle werkzaamheden of risicovolle aannemers. Dankzij Data Mining technieken is er dieper in de data gedoken.
Er is ook gekeken naar de risico's die verschillende combinaties van werkzaamheden, aannemer en/of plaats mogelijk met zich meebrengen.
Zo zou het misschien kunnen zijn dat een bepaalde soort werkzaamheden, zeg baggerwerkzaamheden, erg risicovol blijken te zijn, behalve wanneer deze werkzaamheden uitgevoerd worden door een bepaalde aannemer, omdat deze aannemer altijd speciale voorzorgsmaatregelen treft bij baggerwerkzaamheden.
Het model kijkt ook naar deze deze combinaties bij het beoordelen van een KLIC melding.

Om de historische KLIC meldingen up-to-date te houden, zal tweemaal per jaar een koppeling worden uitgevoerd tussen de KLIC meldingen en de schademeldingen.

Voortaan zullen de KLIC meldingen niet meer eens per week, maar tweemaal per week naar de rayons doorgestuurd worden. Omdat minimaal drie dagen voordat de graafwerkzaamheden plaatsvinden, een KLIC melding moet worden gedaan, is eenmaal per week te weinig. De werkzaamheden zijn al begonnen, voordat men in de rayons op de hoogte is van de grondroering.

De KLIC meldingen worden nu inclusief risicofactor doorgestuurd. Hiervoor zal het programma dat gebruikt wordt door de TWD's aangepast worden. Dit programma leest de KLIC meldingen in, de TWD kan alle informatie per KLIC melding inzichtelijk maken. Ook kan hij met dit programma bijhouden bij welke KLIC meldingen hij toezicht houdt.

Het is niet bekend wat precies de financiele baten zijn van mijn stage-opdracht. Hierover is te weinig bekend over:

- de invloed van een bezoek van een TWD bij een grondroering
- de kwaliteitsverbetering van de selectie die het model maakt tov de selectie die de TWD in zijn huidige situatie maakt

Er is weinig data beschikbaar over de grondroeringen die de TWD in het verleden heeft bezocht. Hierdoor is ook niet bekend wat de invloed van zo'n bezoek is. Hopenlijk zal de aannemer bij grondroeringen waar een TWD toezicht houdt voorzichtiger te werk gaan en minder schade maken. De grondroeringen zijn echter vaak langdurige projecten; de TWD kan niet voortdurend aanwezig zijn. Wanneer meer informatie beschikbaar, zou een vervolgstudie naar deze invloed zeker waardevol zijn.

Doordat niet of weinig bekend is over de historische bezoeken van de TWD, is het ook niet mogelijk een vergelijking te maken tussen de kwaliteit van het model en de kwaliteit van het 'model van de TWD'. De TWD zal bij het selecteren van risicovolle KLIC meldingen afgaan op zijn ervaring en gevoel. Ook beschikt de TWD over kennis die het model niet heeft.

Vooral wanneer de TWD al lang een bepaald gebied onder zijn hoede heeft, weet hij precies wat risicovolle plaatsen zijn, bijvoorbeeld in de buurt van belangrijke kabelkasten.
Hierbij kan een studie gedaan worden naar de beste aansturing van de TWD's.
De volgende mogelijkheden kunnen worden onderzocht:

- De TWD bezoekt de KLIC meldingen met de hoogste risicofactors.
- De TWD bezoekt KLIC meldingen op eigen inzicht.
- De TWD bezoekt KLIC meldingen op verzoek van de aannemer of adviseur.

Natuurlijk zijn combinaties van bovenstaande mogelijkheden ook mogelijk.

Wanneer de invloed van de bezoeken van de TWD bekend is, is het een optie een vervolgstudie te doen naar de optimale bezetting van TWD's.

Bijlage 1: Overzicht van de velden uit figtree en geoversum

FIGTREE

1 CTR
2 Oud claim
3 Dossier nr
4 Schadedtm
5 Datum Aans
6 Boeking FA
7 Factuur da
8 Schadelocatie
9 Correspondent
10 Herst,kstn:Eur
11 Openstaand:Eur
12 Rappeldat1
13 Rappeldat2
14 Rappeldatm
15 Opmerkingen
16 DatumOverd
17 Plaats do
18 Gemeld RA
19 Eerste Ra
20 2e rappel
21 Created
22 ASN region
23 Owner
24 Besch. Ob
25 Betaaldatu
26 bet.dat.ER
27 Bevestigin
28 Code KPN
29 Cor betdat
30 Correspondent
31 Datum Afge
32 Datum brie
33 Kosten EP
34 MM07 nummer
35 Materiaalkosten
36 Ontvanst F
37 Ontvangst
38 Oorzaak S
39 Kosten AP
40 rap1jz
41 Rayon:
42 Rente Betaald
43 Ontvangen:
44 Veroorzaker
45 Storingdtm
46 OverdrJZ
47 KO/WO nummer
48 KLIC
49 KLIC nummer

GEOVERSUM

1 TWD Naam
2 TWD Plaats
3 TWD Mobiel
4 Behandeld door
5 KLIC kantoor
6 Startdatum
7 KLIC verz.
8 KLIC ontv.
9 KLIC nummer
10 Graaflocatie
11 Plaats
12 Aard werkzaamheden
13 Uitvoerend bedrijf
14 Contactpersoon
15 Tel. Nr.
16 Handwerkprocedure
17 UTN_NR
18 OVERLEG_NR
19 DATUM_BEH
20 DATUM_ONTV
21 DATUM_RET
22 SCHADE_NOT
23 Opdrachtgever
24 Contactpersoon
25 Tel. Nr.
26 KLIC Notitie / Locatie
27 KENL_SUB

Uitleg van de velden die van belang zijn voor het onderzoek

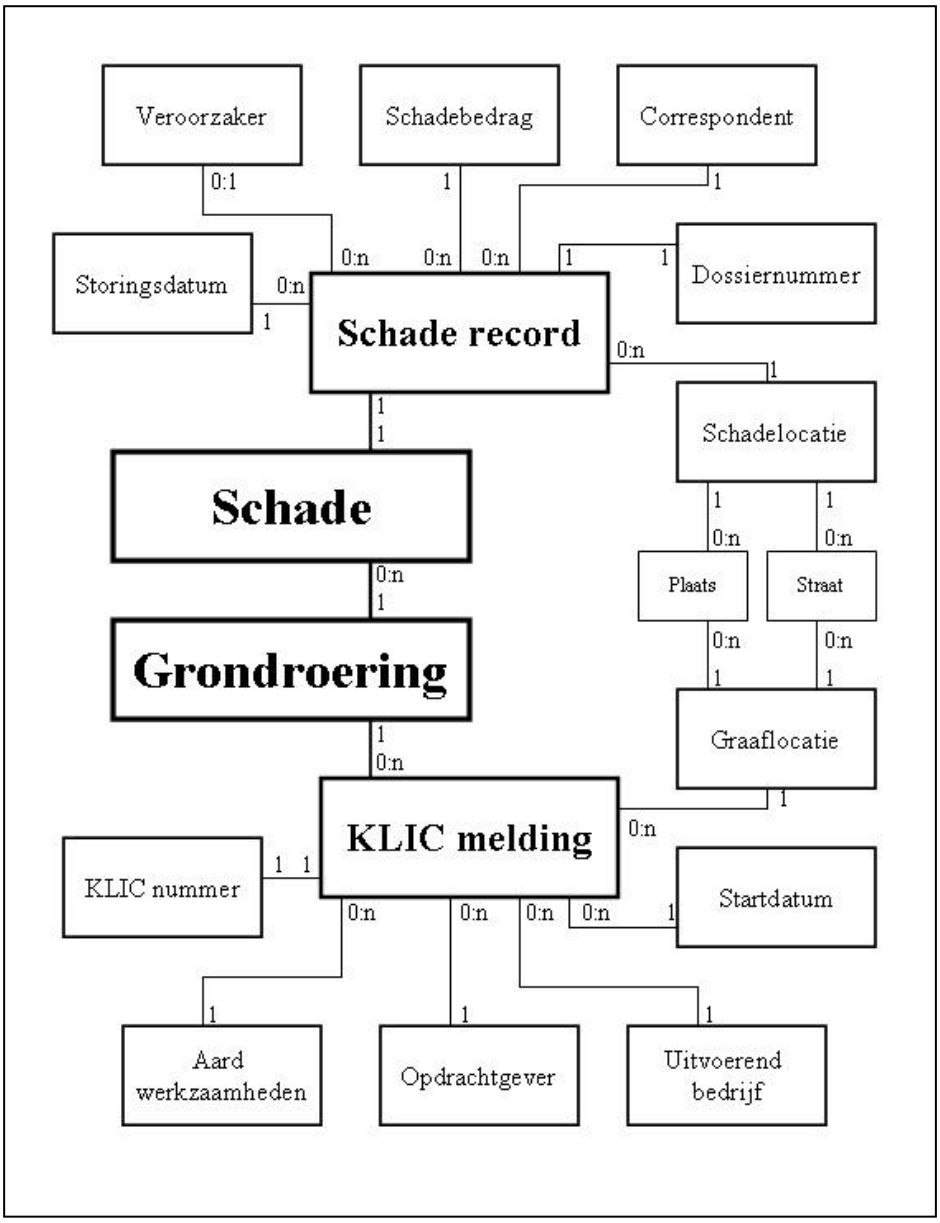
FIGTREE

1	CTR	Rayon waar de schade behandeld is (NO, W, Z)
3	Dossier nr	Code bestaand uit 7 cijfers dat dient als kenmerk voor de schade
4	Schadedtm	Datum waarop de schade is gemaakt
8	Schadelocatie	Plaatsnaam, straatnaam, huisnummer
9	Correspondent	Corresponderende partij
10	Herst,kstn:Eur	Herstelkosten in Euro's
30	Correspondent	Code van de corresponderende partij
44	Veroorzaker	Partij die de schade veroorzaakt heeft
45	Storingdtm	Datum waarop de schade tot een storing leidde
48	KLIC	KLIC-vermelding (Ja, Nee, Onbekend, Niet van toepassing)
49	KLIC nummer	KLIC-nummer dat betrekking heeft op de grondroering waar de schade werd gemaakt

GEOVERSUM

6	Startdatum	Startdatum van de grondroering
9	KLIC nummer	Code dat dient als kenmerk voor de melding (bestaand uit jaartal/regio/nummer)
10	Graaflocatie	Huisadres waar de grondroering plaatsvindt
11	Plaats	Plaatsnaam
12	Aard werkzaamheden	Soort werkzaamheden dat uitgevoerd gaat worden
13	Uitvoerend bedrijf	Naam van het bedrijf dat de werkzaamheden verzorgt
14	Contactpersoon	Naam van de contactpersoon
23	Opdrachtgever	Naam van de opdrachtgever

Bijlage 2: Entiteit relatie diagram KLIC - Schade



Bijlage 3: Enkele Data Mining Technieken

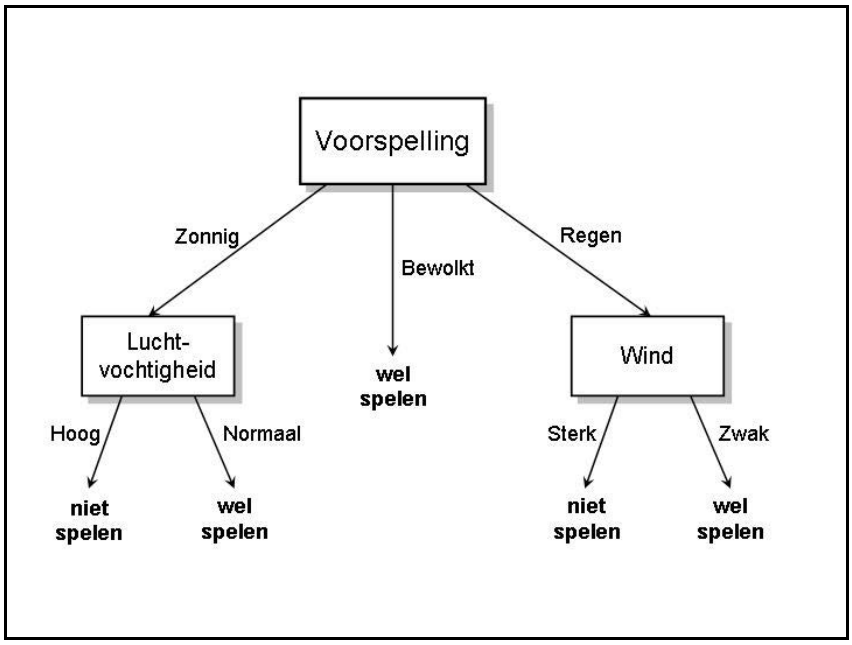
Er zijn veel verschillende Data Mining technieken aanwezig. In deze paragraaf zal een aantal bekende technieken besproken worden die gebruikt voor het uitvoeren van de discovery fase uit het Data Mining proces. Bij de verification-driven Data Mining wordt over het algemeen gebruik gemaakt van statistische methoden.
De technieken zullen kort worden toegelicht.

3.1 Beslisbomen

3.1.1 Inleiding

Beslisbomen is een vrij eenvoudige techniek die een dataset partitioneert in steeds kleiner wordende subverzamelingen, terwijl de omvang van de beslisboom steeds meer toeneemt. Er wordt gebruik gemaakt van een zogenaamde top-down strategie.

Een beslisboom bestaat uit een startpunt (ook wel wortel of root genaamd), tussenliggende punten of knopen en bladeren. De wortel is het beginpunt van waaruit de beslisboom kan worden uitgebreid. Tussenliggende knopen zijn punten die verder kunnen worden uitgebreid, terwijl bladeren juist niet verder worden uitgebreid.



Figuur 3.1 Beslisboom wel/niet tennissen

Ter illustratie wordt in figuur 3.1 een beslisboom getoond die gebruikt kan worden om tot een keuze te komen bij de vraag of er wel of niet getennisd zal worden bij gegeven weersomstandigheden.
De wortel is hier ‘voorspelling’; ‘niet spelen’ is één van de bladeren. Daar een blad niet verder uitgebreid kan worden, is dit een uiteindelijke beslissing van de boom. Bij een

classificatieprobleem zou dus aan elk blad een klasse worden toegewezen. In het geval van het voorbeeld zijn er twee klassen, namelijk ‘niet spelen’ en ‘wel spelen’.
Bij elke tussenliggende knoop (en bij de wortel) moet er een test worden uitgevoerd om de opvolgende knoop of blad te bepalen. Na elke beslissing wordt er als het ware voor elke mogelijke uitkomst een nieuwe beslisboom gecreerd. Deze beslisboom kan weer een knoop, een nieuwe test, zijn of een blad, een uiteindelijke beslissing.

3.1.2 Het genereren van een beslisboom

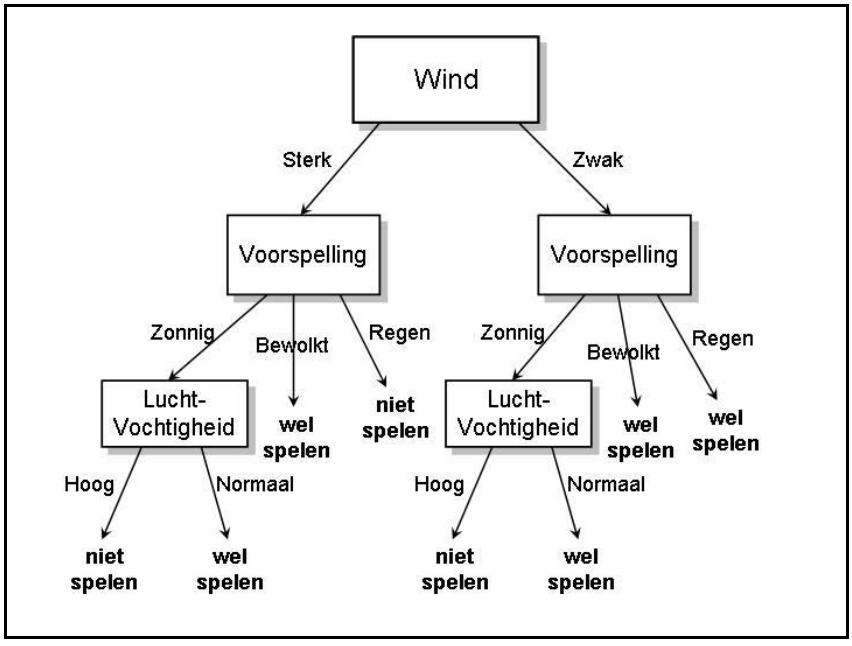
De beslisboom in figuur 3.1 geeft één specifieke manier om de gegevens zoals die in tabel 3.1. staat te classificeren. Dit betekent niet dat dit de enige mogelijkheid is om de gegevens met behulp van een beslisboom te representeren.

Voorspelling	*Luchtvochtigheid*	*Wind*	*Spelen?*
Zonnig	*Hoog*	*Sterk*	*Nee*
Bewolkt	*Hoog*	*Zwak*	*Ja*
Regen	*Hoog*	*Zwak*	*Ja*
Regen	*Normaal*	*Sterk*	*Nee*
Bewolkt	*Normaal*	*Sterk*	*Ja*
Zonnig	*Hoog*	*Zwak*	*Nee*
Zonnig	*Normaal*	*Zwak*	*Ja*

Tabel 3.1 Tennisdata

Figuur 3.2 is een beslisboom die dezelfde gegevens weergeeft. Hoewel beide beslisbomen de gegevens goed classificeren, zal de eerste intuïtief een stuk duidelijker zijn.

Duidelijk is dat de selectie van het attribuut voor een test van grote invloed is op de uiteindelijke vorm van de boom. De volgorde waarin de attributen die in de testen gebruikt worden, is namelijk cruciaal voor het produceren van een begrijpbare en efficiënte beslisboom.



Figuur 3.2 Uitgebreidere beslisboom

Om de kwaliteit van de beslisboom te kunnen bepalen, zullen er bepaalde eisen aan de beslisboom worden gesteld, die invloed hebben op de vorm van de resulterende boom. Bij het begrip kwaliteit kan worden gedacht aan bijvoorbeeld:

- *Omvang*. Om de overzichtelijkheid niet het gedrang te laten komen, zullen er bepaalde beperkingen aan de omvang van de boom worden gesteld.

- *Diepte*. Er zal een grens worden getrokken met betrekking tot het aantal 'lagen' dat de boom diep kan worden. Dit in verband met de mogelijkheid dat de boom zo ver naar beneden uitgroeit, dat de knopen die zich daar bevinden statistisch nauwelijks nog relevant zijn en ook hier de overzichtelijkheid van de boom in gevaar komt.

- *Accuraatheid*. Het is natuurlijk de bedoeling dat de boom zoveel mogelijk voorbeelden juist classificeert. Er kunnen verschillende algoritmen, die in de volgende paragraaf aan de orde komen, geprobeerd worden om dit te bewerkstelligen, waarbij door de verschillende werkwijzen van de algoritmen er ook verschillende bomen zullen ontstaan.

3.1.3 Algoritmen voor beslisbomen

Er zijn verschillende manieren om een beslisboom aan de hand van een verzameling voorbeelden te genereren. Enkele algoritmen waarmee een boom kan worden geinduceerd, zullen aan de orde komen. Het algemene doel dat door deze algoritmen wordt nagestreefd is om een 'zo goed mogelijke' boom te genereren. Dat er zoveel verschillende algoritmen zijn, is het gevolg van de vage definitie van het begrip 'goed'. Elk algoritme heeft wel een eigen definitie van dit begrip. Door de complexiteit van het geheel is het onmogelijk een algoritme te vinden waarmee de optimale beslisboom kan worden gegenereerd. Er zullen drie verschillende algoritme worden besproken: De Concept Learning System, de ID3 en de Classification And Regression Trees.

Concept Learning System
Allereerst zal het Concept Learning System (CLS) worden behandeld. CLS construeert een beslisboom waarbij er wordt geprobeerd de kosten die verbonden zijn aan het classificeren van een dataset te minimaliseren.
Deze kosten bestaat uit het incorrect classificeren van

CLS gebruikt een vooruitziende strategie.
Om een precieze beschrijving te geven van het CLS algoritme: laat een dataset S bestaan uit verschillende voorbeelden, elk behorende tot één van de klassen C_1, C_2, ..., C_k. Het algoritme om beslisboom te genereren kan nu worden gegeven door de volgende stappen:

1. Als alle voorbeelden in verzameling S tot dezelfde klasse C behoren, dan bestaat de beslisboom uit een met deze klasse geassocieerd blad,
2. Anders, laat T een test met mogelijke uitkomsten O_1, O_2, ..., O_n zijn. Elk voorbeeld uit S heeft één uitkomst voor T, dus de test partitioneert S in de subsets S_1, S_2, ...,S_n, waarbij ieder voorbeeld uit S_i uitkomst O_i heeft voor T. T wordt de wortel van de beslisboom en, voor iedere uitkomst O_i, wordt een onderliggende beslisboom geconstrueerd door het recursief uitvoeren van het algoritme met de verzameling S_i.

Een belangrijke vraag is nu of de attributen wel voldoende informatie bevatten om een beslisboom te kunnen genereren. De attributen zijn niet geschikt wanneer er twee voorbeelden in de dataset zijn die dezelfde waarde voor ieder attribuut hebben, maar tot verschillende klasse behoren.
Wanneer de volgende twee voorbeelden in een dataset met sportartikelen aanwezig zouden zijn, zou het onmogelijk zijn om verschil te kunnen zien tussen een pingpongbal en een golfbal:

Grootte	Vorm	Kleur	Classificate
Klein	Rond	Wit	Pingpongbal
Klein	Rond	Wit	Golfbal

Als de attributen geschikt zijn, kan er altijd een test worden geconstrueerd die een niet-triviale partitie van de training set produceert. De basisalgoritme houdt echter geen rekening met complicaties die voor kunnen komen bij toepassing van inductie in de praktijk; de attributen zijn gewoon niet altijd geschikt.

ID3

ID3 is in 1986 door Quinlan ontwikkeld naar aanleiding van CLS. Het vervangt de vooruitziende op kosten gebaseerde functie van CLS door een evaluatie functie gebaseerd op informatie. Het zogenaamde *gain criterium* selecteert die test die het meeste aan informatie winst oplevert. Het gain criterium heeft echter een voorkeur voor attributen met meer waarden. Om dit te ondervangen kan gebruik gemaakt worden van een subset criterium die de waarden in subsets verdeelt zodat alle testen slechts twee uitkomsten hebben. Nadeel hiervan is dat de resulterende beslisbomen nauwelijks te begrijpen meer zijn en veel meer rekentijd kosten om te construeren.
Het idee achter ID3 is als volgt. Bij een gegeven dataset wordt getracht een test voor de wortel te vinden die het meeste informatie zal gaan opleveren. Dat wil zeggen een test die een zo ‘zuiver’ mogelijke splitsing in de verzameling oplevert. Hiermee wordt bedoeld dat de klassen zo goed mogelijk door de splitsing worden gescheiden, en dat er dus zo min mogelijk voorbeelden met verschillende classificaties door de test op één hoop worden gegooid. Zodra het attribuut dat het meeste informatie oplevert is gevonden, wordt het attribuut geselecteerd, vindt de splitsing plaats en herhaalt het proces zich voor de knopen die dan ontstaan zijn.

Nadelen van ID3:

- Dit type algoritme heeft aanzienlijke problemen om gegevens waar veer ruis in zit goed te kunnen verwerken. Met ruis wordt hier bedoeld: fouten in de attributen of fouten in de informatie over de klassen. Ruis zorgt er bij gebruik van dit algoritme voor dat de resterende bomen vaak groot en erg complex zijn.

- ID3 blijft de subsets opdelen tot er geen enkele uitzondering meer aanwezig is. Dit streven om 100% correctheid te bereiken kan er voor zorgen dat de ID3 bijzonder uitgebreide bomen genereert met knopen die zeer weinig voorbeelden bevatten.

- Omdat beslisbomen een representatie voor geleerde concepten zijn, is het checken op equivalentie niet eenvoudig. Het is gewoon moeilijk om de concepten te begrijpen wanner het uitgedrukt is in de vorm van een grote beslisboom. Verder kan de manier

waarop een beslissing wordt genomen op basis van een beslisboom geheel anders zijn dan de manier waarop een menselijke expert dit zou doen. Het gevaar bestaat door de bomen het bos niet meer te zien.

Classification And Regression Trees

Het algoritme Classification And Regression Trees (CART) is eveneens ontwikkeld naar aanleiding van CLS. CART, ontwikkeld nog voor het ID3 algoritme, construeert een beslisboom door het herhaaldelijk opslitsen van de dataset in kleinere subsets. Hierbij wordt getracht een systematische manier te vinden waarmee de klasse waartoe een object behoort kan worden voorspeld. Daarnaast wordt ook geprobeerd de voorspellende structuur van het probleem te ontrafelen.

Het construeren van een boom wordt bepaald door:

- de selectie van de splits,
- de beslissing om een knoop verder op te splitsen of niet, en
- het toewijzen van een klasse aan een niet-gesplitste knoop.

Een split deelt een dataset in twee subsets. In principe zijn de splits van CART dus gelijk aan de testen die gebruikt worden in ID3. Een belangrijk verschil is dat CART alleen binaire splits kent. Bij het selecteren van een split wordt getracht de homogeniteit van de klassen van de resterende subsets te optimaliseren. Met andere woorden: CART probeert een split zo te kiezen dat de resterende subsets voorbeelden van één klasse hebben. De bruikbaarheid van een split kan dus worden gedefinieerd als een soort mate van zuiverheid of onzuiverheid, waarbij met het splitsen van de dataset wordt geprobeerd de beste scheiding van de klassen te bereiken volgens een bepaald criterium voor de zuiverheid daarvan.
Een boom wordt op de volgende manier gecreeerd. Om te beginnen worden alle mogelijke kandidaat-splits bekeken om die split te vinden die de grootste afname bewerkstelligt met betrekking tot één of andere maatstaf voor de onzuiverheid. Het uitbreiden (opsplitsen) van een knoop stopt zodra er geen significante afname van deze maatstaf meer is. De klasse die dan aan deze knoop wordt toegewezen is de klasse die het meest voorkomt tussen de voorbeelden uit de subset. Het CART algoritme bevat een standaard verzameling splits of testen.

3.2 Regel inductie

3.2.1 Inleiding

Met regel inductie wordt het extraheren van voor de mens (en computer) duidelijke en begrijpelijke kennis uit gegevens bedoeld. Deze kennis wordt dan gerepresenteerd in de vorm van regels met daarbij relevante gegevens met betrekking tot die regels. Regel inductie op zich is eigenlijk geen techniek, maar een vorm van classificatie die kan worden bewerkstelligd door gebruik te maken van andere technieken, zoals beslisbomen. In deze paragraaf zal het begrip ‘regel’ formeel worden beschreven en zal er een voorbeeld van een regel inductie worden gegeven om het gebruik ervan te verduidelijken.

3.2.2 Regels

De regels die getracht worden te extraheren uit de gegevens zijn eigenlijk een soort beschrijvingen van de te classificeren klassen in een voor de mens eenvoudig te begrijpen vorm. Deze beschrijvingen refereren dan aan de voorspellende attributen van de dataset, dat wil zeggen de attributen die niet de informatie bevatten waar aangegeven wordt tot welke klasse een voorbeeld hoort. In het ideale geval zullen alle voorbeelden uit de dataset die aan de voorwaarden van een regel voldoen correct door die regel worden geclassificeerd, maar over het algemeen zal dit niet het geval zijn en geeft de bij de regel behorende accuraatheid aan in hoeveel procent van de gevallen dat wel zo is.
Daar er alleen attributen van de voorbeelden uit de dataset bekend zijn, en geen relaties tussen die voorbeelden, kunnen de beschrijvingen slechts bestaan uit condities op deze attributen. De beschrijvingen die over het algemeen gebruikt worden bij Data Mining tools zijn een subset van de selectie condities uit de relationele algebra.

Een regel bestaat uit één of meerdere condities A en een klasse B_i.
In gewone taal komt dit neer op: Als A, dan B_i.
Hierbij dient opgemerkt te worden dat een regel voor een klasse niet uniek is. Verschillende beschrijvingen (één of meerder condities) kunnen geconstrueerd worden die correct zijn met betrekking tot een dataset. Het probleem is echter dat niet al deze beschrijvingen op correcte wijze nog niet bekende voorbeelden zullen classificeren. Dit is een direct gevolg van het feit dat niet alle kennis die is afgeleid uit observaties per definitie juist is.
Verder zijn er nog een drietal belangrijke begrippen die betrekking hebben op de afzonderlijke gevonden regels.

Accuraatheid

Met de accuraatheid van een regel wordt de verhouding aangegeven, waarmee de regel de objecten die aan de beschrijving van de regel voldoen juist classificeert. Dit gebeurt meestal in de vorm van een percentage, met behulp van de volgende formule:

$$\frac{\text{aantal juist geclassificeerde objecten}}{\text{aantal te classificeren objecten}} \text{ x } 100\%$$

Coverage of dekking

Met de coverage of overdekking van een regel wordt het aantal objecten bedoeld, die aan de beschrijving voldoen. Dit is een zeer interessante statistiek, daar het duidelijk is dat in de praktijk aan regels met een lage coverage weinig of geen waarde kan worden gehecht. Er wordt daarom hier een bepaalde grenswaarde gesteld aan het minimale aantal objecten dat aan de beschrijving van de regels moet voldoen.

Betrouwbaarheid

De betrouwbaarheid van een regel is een bepaald significantieniveau, dat aangeeft in welke mate de gevonden regel in de praktijk (buiten de gebruikte dataset) nog steeds correct zal zijn. De betrouwbaarheid kan met behulp van statistische technieken worden berekend en is nauw verbonden met de accuraatheid en coverage van een regel.

3.2.4 Een voorbeeld

Ter illustratie nemen we de dataset die we ook gebruikt hebben bij de beslisbomen gegeven in tabel 3.1. De bedoeling is dat er regels worden gevonden voor het eenvoudig kunnen bepalen of er wel of niet getennist zal worden aan de hand van de weersomstandigheden die in de tabel staan.

Wanneer er bijvoorbeeld gekozen is om beslisbomen als techniek voor het vinden van regels te gebruiken, zou dit dus kunnen resulteren in een boom zoals in figuur 3.1 is gegeven. De regels voor het classificeren van tennisomstandigheden zijn nu eenvoudig uit deze boom te halen:

1. Als (Voorspelling = Zonnig) en (Luchtvochtigheid = Hoog) → Niet spelen
2. Als (Voorspelling = Zonnig) en (Luchtvochtigheid = Laag) → Wel spelen
3. Als (Voorspelling = Bewolkt) → Wel spelen
4. Als (Voorspelling = Regen) en (Wind = Sterk) → Niet spelen
5. Als (Voorspelling = Regen) en (Wind = Zwak) → Wel spelen

Bij deze regels worden alle voorbeelden uit de dataset correct geclassificeerd. De regels hebben dus een accuraatheid van 100%. Er zijn nog wel meer regels te bedenken zoals: *Als (Voorspelling = Regen) en (Luchtvochtigheid = Hoog) → Wel spelen, m*aar de bovenstaande vijf regels zijn voor classificatie al voldoende.

Wanneer de dataset uitgebreider wordt, zullen de regels niet zo makkelijk uit de dataset te halen zijn. Daarom zal nu naar een iets uitgebreidere dataset gekeken worden zoals die in tabel 3.2 staat.

Voorspelling	*Luchtvochtigheid*	*Temperatuur*	*Wind*	*Spelen?*
Zonnig	*Hoog*	*Heet*	*Sterk*	*Nee*
Bewolkt	*Hoog*	*Heet*	*Zwak*	*Ja*
Regen	*Hoog*	*Normaal*	*Zwak*	*Ja*
Regen	*Normaal*	*Koud*	*Sterk*	*Nee*
Bewolkt	*Normaal*	*Heet*	*Sterk*	*Ja*
Zonnig	*Hoog*	*Heet*	*Zwak*	*Nee*
Zonnig	*Normaal*	*Normaal*	*Zwak*	*Ja*
Bewolkt	*Hoog*	*Normaal*	*Sterk*	*Nee*
Regen	*Normaal*	*Koud*	*Zwak*	*Nee*
Zonnig	*Normaal*	*Normaal*	*Hoog*	*Ja*
Regen	*Hoog*	*Koud*	*Zwak*	*Nee*
Zonnig	*Hoog*	*Normaal*	*Zwak*	*Ja*

Tabel 3.2 Uitgebreidere tennisdataset

De regels zijn nu lastiger uit de dataset te halen, omdat de dataset meer voorbeelden heeft gekregen. Als we kijken naar de vijf opgestelde regels, zien we dat de accuraatheid van enkele regels geen 100% meer is.

- Als (Voorsp. = Zonnig) en (Luchtv. = Hoog) → Niet spelen (met acc. = 67%)
- Als (Voorsp. = Zonnig) en (Luchtv. = Normaal) → Wel spelen (met acc. = 100%)
- Als (Voorsp. = Bewolkt) → Wel spelen (met acc. = 67%)
- Als (Voorsp. = Regen) en (Wind = Sterk) → Niet spelen (met acc. = 100%)

- Als (Voorsp. = Regen) en (Wind = Zwak) → Wel spelen (met acc. = 33%)

Er kan nu gekozen worden voor óf meer en complexere regels óf een lagere accuraatheid.

3.2.5 Voordelen en Nadelen

Voordelen van regel inductie zijn:

- Regels kunnen simpel worden geïnterpreteerd door menselijke experts. Een regel kan dus worden begrepen zonder referentie naar andere regels.
- Regels zijn over het algemeen vrij eenvoudig te verifiëren.

Daarnaast zijn er ook enkele problemen met het gebruik van regels:

- Er zal een afweging moeten worden gemaakt tussen het stellen van hoge en lage eisen aan de accuraatheid en aan de coverage van de regels. Stelt met aan beide begrippen te hoge eisen, dan zullen er nauwelijks nog regels gevonden worden die hieraan voldoen. Worden de eisen echter naar beneden bijgesteld, dan zullen de gevonden regels bijna niet betrouwbaar genoeg meer zijn.
- Regels zijn moeilijk te onderhouden en niet altijd geschikt om alle typen kennis te representeren.
- Met de toename van het aantal objecten en attributen van een dataset, zoal in het voorbeeld, vindt er een explosie plaats met betrekking tot het aantal regels dat gevonden kan worden. Er zal dus een soort filtering van de regels moeten plaatsvinden, daar de meeste regels niet relevant zullen zijn.
- De uiteindelijke verzameling regels zal niet te veel overlap moeten vertonen, zodat geen onnodig groot aantal regels gebruikt en verwerkt zal moeten worden.

3.3 Genetische algoritmen

3.3.1 Inleiding

Genetische algoritmen (GA) is een benaming voor een adaptieve heuristische zoektechniek die gebaseerd is op de evolutionaire ideeën omtrent natuurlijke selectie en genetica.
Een GA werkt met verzamelingen van potentiele oplossingen, de zogenaamde *populaties*. De populaties evolueren aan de hand van bepaalde genetische regels. In het bijzonder wordt gebruik gemaakt van de principes van 'survival of the fittest', zoals die door Darwin zijn opgesteld. Het resultaat hiervan is dat de 'kwaliteit' van de populaties steeds beter wordt en zo een optimale oplossing wordt gevonden.

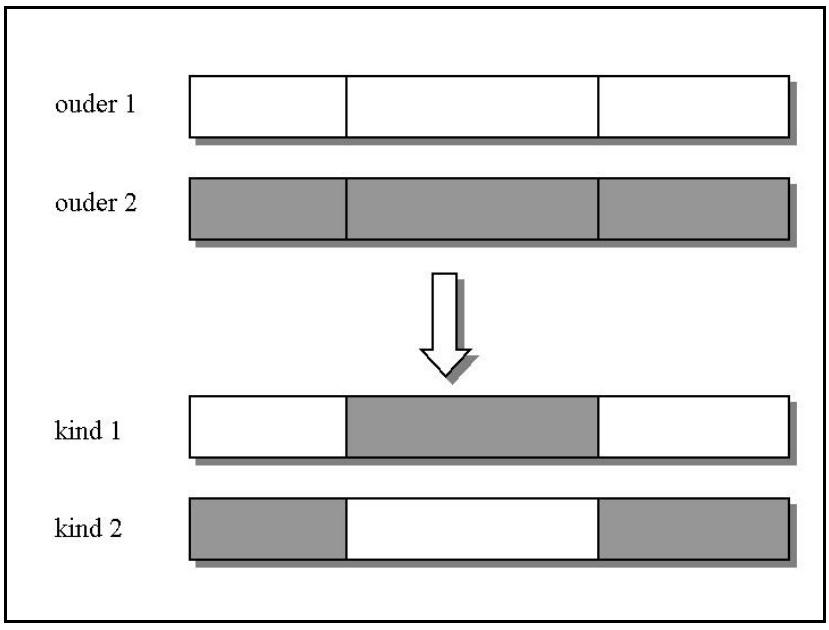
3.3.2 De werking van genetische algoritmen

Het fundamentele onderliggende principe dat werkt aan de hand van een populatie van individuen maakt gebruik van drie operaties:

1. Evaluatie van de individuele geschiktheid (of kwaliteit) van de individuen, aan de hand van een fitness-functie
2. Formatie van een 'mating pool' en
3. Hercombinatie en mutatie

De individuen die het gevolg zijn van het uitvoeren van deze drie operaties vormen de populatie van de volgende generatie. Het proces wordt herhaal tot er geen significante verbetering meer plaatsvindt, dat wil zeggen de 'sterkste' individuen hebben het selectieproces overleefd. De omvang van de populatie in iedere generatie is constant. Over het algemeen wordt ieder individu uit de populatie gerepresenteerd door een vector van bepaalde lengte met binaire of numerieke waarden.
Na de evaluatie van de individuele geschiktheid worden de individuen afhankelijk van hun relatieve fitness aan de 'mating pool' toegevoegd. Dit houdt in dat individuen die het goed doen, meerdere keren zullen terugkomen terwijl de zwakkere exemplaren minder of geen bijdrage leveren aan de volgende generatie.

Hercombinatie gebeurt met behulp van een n-point crossover operator. Een operator wordt weergegeven in figuur 3.3. De operator selecteert willekeurig twee ouders uit mating pool, samen met een n crossover posities binnen de vector. De ouders ruilen dan van 'staart' (het gedeelte rechts van de crossover posities), wat resulteert in twee nakomelingen. Deze kinderen zullen deel uitmaken van de volgende generatie. Soms is het wenselijk slechts een deel van de volgende generatie door crossover te genereren.



Figuur 3.3 cross-over

Daarnaast is er nog de mutatie operator, waarbij er (met kleine kans) een willekeurige waarde in de vector van een individu wordt veranderd. Op deze manier kan er door puur toeval een sterkere generatie van indviduen ontstaan.

3.3.3 Het algoritme

Het formele genetische algoritme is als volgt gedefinieerd:

1. Selecteer een eindig en begrensd domein voor de zoektocht en representatie die de zoekruimte discretiseert

2. Kies een geschikte populatie omvang N en initialiseer de start populatie.

3. Evalueer de individuen aan de hand van de fitness functie f().

4. Als aan de criteria functie om te stoppen is voldaan, stop. Zo niet, selecteer aan de hand van de bij 3. uitgevoerde evaluatie op probabilistische wijze individuen om een mating pool van omvang N te genereren.

5. Zolang de omvang van de tijdelijke populatie TP minder dan N bedraagt, kies een willekeurige ouder uit de mating pool. Voeg de ouder aan de tijdelijke TP toe met kans p_1. Met kans $(1-p_1)$, kies een tweede ouder uit de mating pool, laat de crossover operator op beide ouders los en voeg het resultaat aan de TP toe.

6. Muteer willekeurig gekozen individuen uit TP met kleine kans p_2. Laat P = TP en ga terug naar stap 3.

De volgende opmerkingen dienen te worden gemaakt:

- Het is mogelijk om meerdere kopieën van een individu in een populatie en/of de mating pool te hebben.
- Er zijn enorm veel keuze mogelijkheden; denk aan verschillende fitness-functies, variaties van de genetische operatoren en selectieschema's en andere representaties van de individuen.
- Mutatie is een hulpmiddel om diversiteit in de populatie te garanderen. Het is niet de primaire zoekoperator
- De criteria om de stoppen worden vaak gegeven in termen van een grenswaarde met betrekking tot de toename in verbetering of het totale aantal generaties

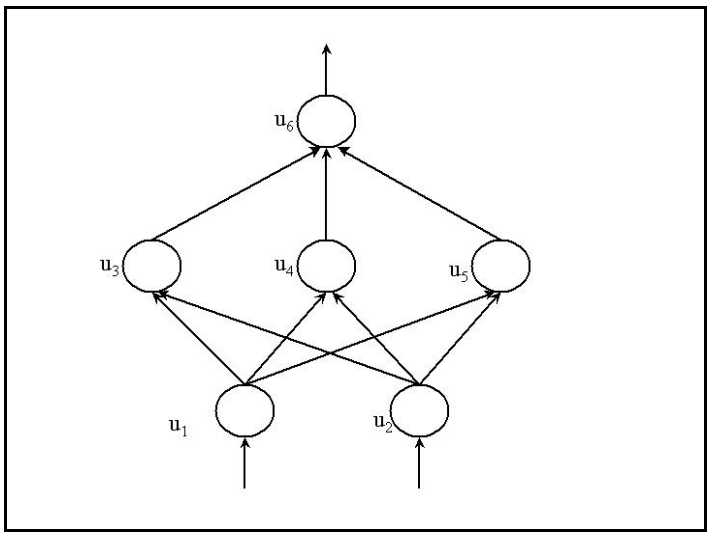
3.4 Neurale netwerken

3.4.1 Inleiding

De motivatie voor het ontstaan van neurale netwerken komt uit de neurologie. Hierbij werd getracht met een netwerk het menselijk brein na te bootsen. De eerste neurale netwerken zijn afkomstig uit de jaren vijftig.
Neurale netwerken genieten op het moment een zeer grote populariteit en worden bij een groot aantal verschillende toepassingen gebruikt. Classificatie, scoring en clustering zijn de belangrijkste Data Mining taken waar deze netwerken voor gebruikt worden. In deze paragraaf zal de structuur van een neuraal netwerk worden beschreven.

3.4.2 Enkele begrippen

Een neuraal netwerk kan worden beschreven aan de hand van de netwerk, cel en dynamische eigenschappen, met daarbij de leeralgoritmen die gebruikt worden om het netwerk te trainen. Elk van deze begrippen zal hieronder nader worden toegelicht.



Figuur 3.4 Neuraal netwerk

Netwerk eigenschappen

Figuur 3.4 geeft een voorbeeld van een neuraal netwerk, waarin is te zien dat het netwerk bestaat uit zelfstandige verwerkingseenheden, cellen genaamd, die verbonden worden door pijlen. De cellen en de pijlen van een netwerk tezamen vormen de *netwerk topologie*. Elke pijl (connectie) heeft een numerieke waarde wij, hetgeen het gewicht genoemd wordt. Dit gewicht geeft de mate van invloed van cel u_j op cel u_i aan. Positieve gewichten geven bekrachtiging aan, terwijl negatieve gewichten duiden op terughouding. De gewichten bepalen het gedrag van het netwerk.
In sommige modellen wordt een subset van de cellen als *input cellen* beschouwd, dat wil zeggen dat ze van buitenaf gestuurd worden. In het voorbeeld zijn dit de cellen u_1 en u_2.
De *output cellen* zijn de cellen waarvan de uitvoer wordt gezien als totale uitvoer van het neurale netwerk (cel u_6). Cellen die noch input cel noch output cel zijn, worden *intermediate cellen* genoemd.

Cel eigenschappen

Iedere cel ui berekent een enkele numerieke cel uitvoer of *activatie*. In figuur 3.4 wordt de activatie van bijvoorbeeld u_2 gebruikt om de activatie van de cellen u_3, u_4 en u_5 te berekenen. De invoer en de activaties van de cellen kunnen zowel discreet, met waarden uit {0,1} of {-1,0,1}, als continu zijn, met waarden uit het interval [0,1] of [-1,1]. Met ui wordt, afhankelijk van de context, zowel de cel als de activatie bedoelt.
Normaal gesproken gebruikt iedere cel hetzelfde algoritme om de activatie te berekenen. Die activatie wordt hierbij berekend met behulp van de activaties van de verbonden cellen en de gewichten van die connecties. Iedere cel u_i (behalve de input cellen) berekent die nieuwe activiteit u_i als functie van de gewogen som van de verbonden activaties en gewichten.

De werking van het netwerk

Voor ieder neuraal netwerk moeten de tijdstippen worden aangegeven waarop het berekenen van de nieuwe activaties van de cellen en het veranderen van de uitvoer van de cellen precies plaatsvindt.
Hiervoor zijn verschillende mogelijkheden.
De cellen kunnen in een bepaalde volgorde afgelopen worden, waarbij elke cel wordt geëvalueerd en de nieuwe activatie wordt berekend voor de volgende cel wordt aangedaan.
Een andere mogelijkheid is om herhaaldelijk door het netwerk te lopen. Het netwerk kan dan in een stabiele toestand belanden ofwel zichzelf gaan herhalen.
Een derde mogelijkheid is om voor alle cellen de activaties tegelijk te berekenen en daarna tegelijk de uitvoer van de cellen te veranderen.
Tenslotte is er de mogelijkheid om een willekeurige cel te kiezen, de nieuwe activatie te berekenen en daarna de uitvoer aan te passen alvorens een nieuwe cel te kiezen.

Leeralgoritmen

Ieder neuraal netwerk is geassocieerd met één of meer leeralgoritmen, waarmee wordt getracht het netwerk 'zich zo goed mogelijk te laten gedragen'. Dit komt neer op het vinden van optimale gewichten voor het netwerk, door deze gewichten steeds aan te passen en de uitvoer van het netwerk te evalueren.

Het belangrijkste aspect van neurale netwerken is dat ze kennis opdoen door getraind te worden aan de hand van voorbeelden. In plaats van geprogrammeerd te worden, leren ze eigenlijk door ervaring op te doen.

3.5 Kernel Dichtheid Classificatie

Bij de Kernel Dichtheid Classificatie wordt een voor een ongelabelde record geclassificeerd aan de hand van (voor elke klasse berekende) verwachtingen.
De record wordt geclassificeerd naar de klasse die de hoogste verwachting geeft.

De ongelabelde record x_0 wordt vergeleken met alle gelabelde invoerrecords x_i.
Met behulp van een Kernel functie $K(x_0,x_i)$ wordt een waarde of gewicht berekend.
De Kernel functie maakt gebruik van het verschil of de afstand tussen x_0 en x_i: $|x_i - x_0|$. Deze afstand is altijd een waarde tussen 0 en 1. De functie heeft de volgende eigenschappen:

- Als $|x_i - x_0| < |x_j - x_0|$ → $K(x_0,x_i) > K(x_0,x_j)$
- Als $|x_i - x_0| = 1$ → $K(x_0,x_i) = 0$

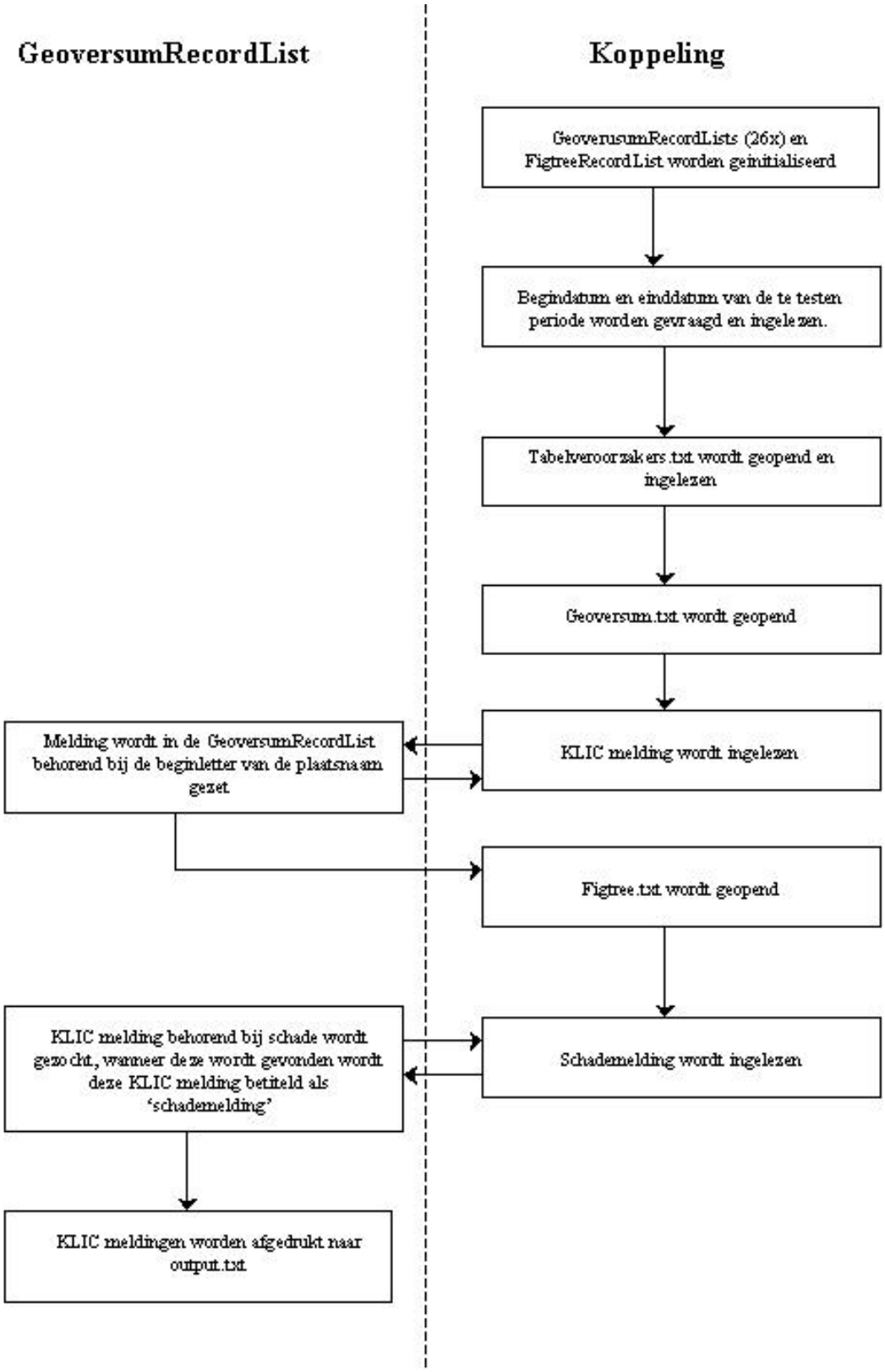
De eerst eigenschap houdt in dat een gelabelde record x_i die een kleine afstand heeft naar de ongelabelde record x_0, een grote waarde of gewicht krijgt.
Gelabelde records met een kleine afstand zijn records die weinig of geen gelijke velden hebben met de ongelabelde record.
Wanneer de afstand tussen x_0 en x_1 is, wordt een waarde van 0 toegewezen. Het gaat hier om records die totaal geen gelijke velden hebben met de ongelabelde te classificeren record.

Voor elke klasse J kan een verwachting P_J worden berekend door de som van alle waarden of gewichten van de met klasse J gelabelde records op te tellen en te delen door de som van de waarden van alle gelabelde records:

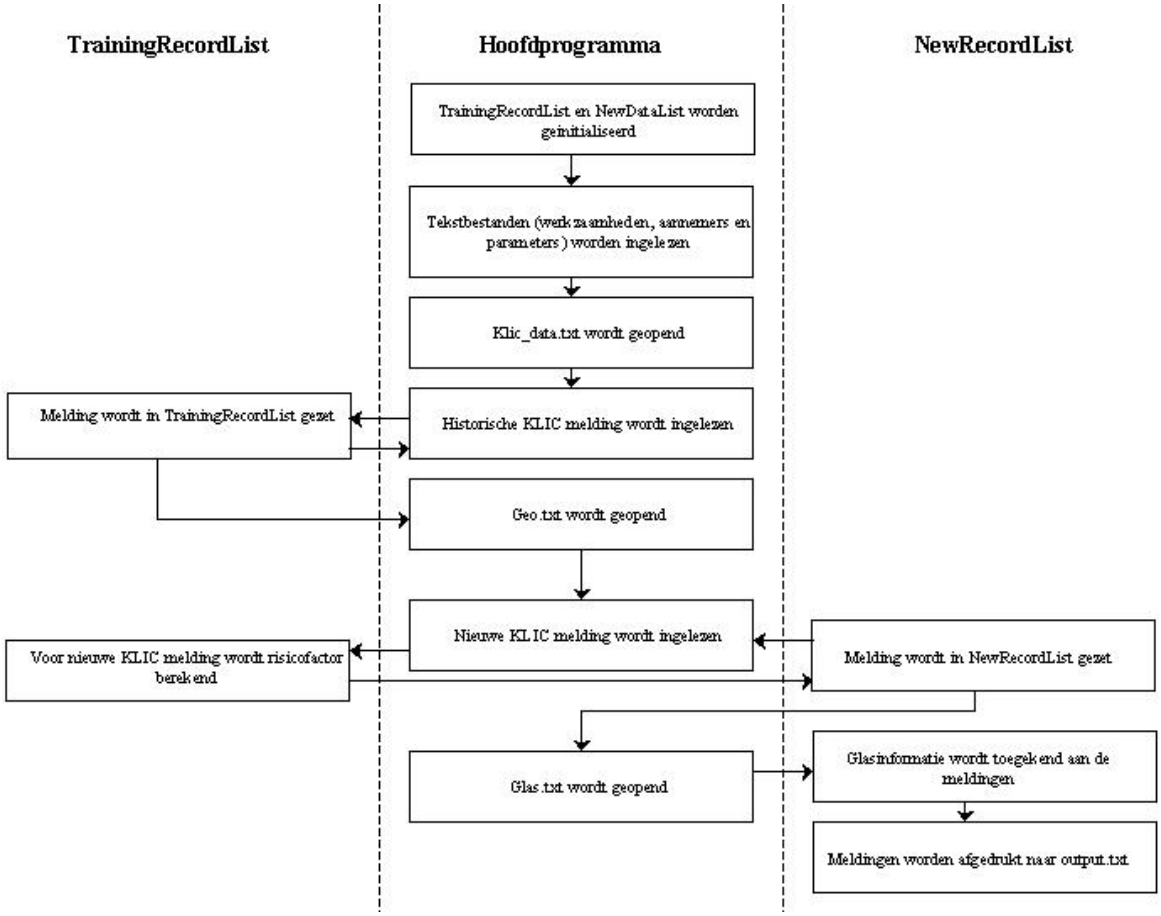
$$P_J = \frac{\sum_{i \in J} K(x_0, x_i)}{\sum_i K(x_o, x_i)}$$

De record x_0 wordt geclassificeerd naar de klasse J met de hoogste verwachting P_J.

Bijlage 4: Activity Diagram Koppeling



Bijlage 5: Activity Diagram Model



Bijlage 6: Handleiding Koppeling Figtree – Geoversum

De koppeling van gegevens uit Figtree en Geoversum zal twee maal per jaar gedraaid worden op de volgende data:

- ± 1 januari
- ± 1 juli

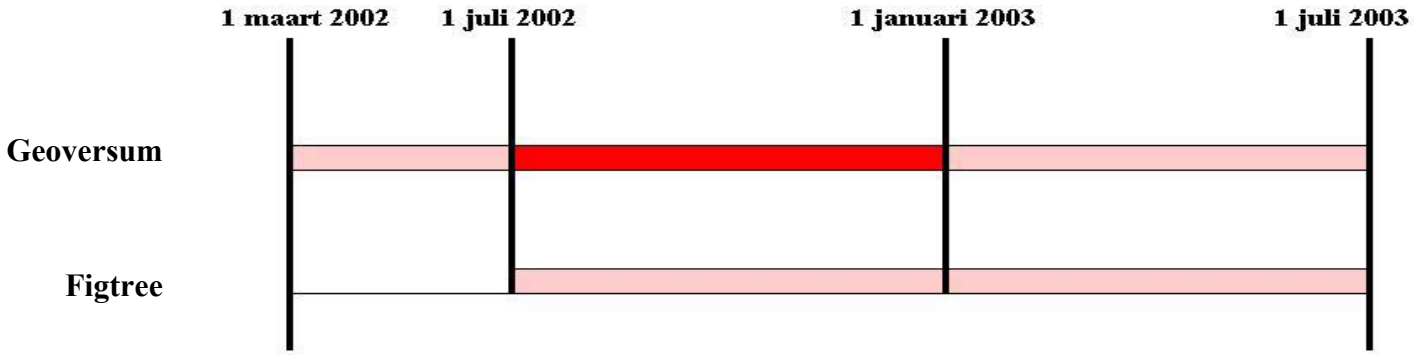
De koppeling dient ter verversing van de input voor het wekelijks te draaien model.

Het meeste werk zit in het aanmaken en verzamelen van benodigde gegevens.

6.1 Voorbereiding: benodigde gegevens

Voor de koppeling is zowel data uit geoversum (KLIC meldingen) als data uit Figtree (kabelschades) nodig.

Het onderstaande plaatje geeft aan welke data nodig is voor de eerstvolgende koppeling rond 1 juli 2003. Uiteraard worden voor een volgende koppeling deze tijdstippen allemaal een (half) jaar doorgeschoven.



Dit geeft dus de volgende benodigde gegevens:

- Alle KLIC meldingen uit Geoversum van 1 maart 2002 t/m 1 juli 2003.
- Alle schades uit Figtree van 1 juli 2001 t/m 1 juli 2003.

- Verder moet een tekstbestand aanwezig zijn met correspondentcodes die gebruikt worden in Figtree.

Geoversum gegevens

Geoversum data zal in samenwerking met Evert Scharrenburg uit Geoversum gehaald moeten worden.
Het bestand moet onder de naam geoversum.txt in de directory koppeling/invoer staan.

De geoversumdata zal als excelbestand door een macro gehaald moeten worden welke Evert ook gebruikt voor de 2-wekelijkse verspreiding van KLIC-meldingen (dit staat beschreven in de handleiding van het model). De macro dient te worden gedraaid op de PC van Evert i.v.m. directorie-verwijzingen. Let hierbij op dat excel niet meer dan 65.536 velden kan bevatten, dus niet in één keer alle meldingen van de vijf kwartalen uit draaien. Via macro per kwartaal opslaan. (Tekstbestanden kun/moet je wel weer aan elkaar plakken tot één bestand).

Figtree gegevens

Figtree heeft als de functionaliteit om data als tekstbestanden uit te draaien.

Met behulp van deze functionaliteit kunnen zowel de schadegegevens als de correspondentcodes verkregen worden.

Schademeldingen:

- Vanuit startscherm Figtree 'Ad hoc' aanklikken.
- Hoofdtabel: Records Management
- Selecteer 'Historie tbv Risico-Analyse KLIC'
- Wijzig de start- en einddatum
- Start
- 'Maak ASCII-bestand' aanvinken
- Browse: lokatie van het tekstbestand bepalen
- OK
- Afdrukken
- Het tekstbestanden openen vanuit excel (gescheiden door komma)
- Alles selecteren
- Bewerken → vervangen
- Zoeken naar: **;** Vervangen door: **veld leeglaten**
- Bestand → opslaan als
- Bestandsnaam: figtree
- Opslaan als: CSV (gescheiden door lijstscheidingsteken) (*.csv)
- Excel afsluiten
- Het bestand figtree.csv wat nu gemaakt is renamen naar figtree.txt
- Dit bestand plaatsen in de directory *koppeling/invoer*

Correspondenten:

- Vanuit startscherm Figtree 'Ad hoc' aanklikken
- Hoofdtabel: **Adres**
- Selecteer rapportage 'adressen'
- 'Maak ASCII-bestand' aanvinken

- Het tekstbestand openen vanuit excel (gescheiden door tab)
- Kolom A t/m C selecteren
- Rechtermuisknop → verwijderen
- Kolom B t/m K selecteren
- Rechtermuisknop → verwijderen
- Kolom C selecteren
- Rechtermuisknop → verwijderen
- Kolom D t/m P selecteren
- Rechtermuisknop → verwijderen
- *Als alles goed is gegaan bevat kolom A nu de code en kolom B de naam van het bedrijf en kolom C het adres*
- Alles selecteren
- Data → Filter → Autofilter
- Pijltje bij kolom B aanklikken → (Lege cellen) *helemaal onderaan*
- De zichtbare rijen selecteren
- Rechtermuisknop → Rijen verwijderen
- Data → Filter → Autofilter
- Alles selecteren
- Bewerken → vervangen
- Zoeken naar: **;** Vervangen door: **veld leeglaten**
- Bestand → opslaan als
- Bestandsnaam: correspondenten
- Opslaan als: CSV (gescheiden door lijstscheidingsteken) (*.csv)
- Excel afsluiten
- Het bestand correspondenten.csv renamen naar correspondenten.txt
- Dit bestand plaatsen in de directory *koppeling/invoer*

6.2 Het model draaien

Wanneer de drie bestanden:
- correspondenten.txt
- figtree.txt
- geoversum.txt

in de directory koppeling/invoer staan, kan de koppeling gedraaid worden door koppeling.exe te starten.

Na een aantal uur (afhankelijk van de processorsnelheid) is de koppeling klaar en is het bestand **klic_historisch.txt** aangemaakt.

6.3 Uitvoer bewerken:

Het bestand klic_historsich.txt is de output van het programma.

Dit bestand bevat alle KLIC meldingen die als input gebruikt zijn nu ook met schadeveld. Omdat alleen het donker gearceerde gedeelte uit het figuur op pagina 2 nuttige informatie bevat, zal deze uit het bestand gehaald moeten worden mbv excel.
- Excel starten
- Tekstbestand openen (gescheiden door puntkomma)

- Sorteren op datum (kolom A)
- Alle rijen die datums bevatten die in het donker gearceerde gebied vallen selecteren en kopieren in een nieuw excelbestand
- Bestand → opslaan als
- Bestandsnaam: klic_historisch_xxx (xxx = periode donker gearceerde gebied; zie benamingen in directory "data_klic_historisch")
- Opslaan als: CSV (gescheiden door lijstscheidingsteken) (*.csv)
- Excel afsluiten (wijzigingen niet opslaan)
- klic_historisch_xxx.csv renamen naar klic_historisch_xxx.txt

Er is nu weer een half jaar historische KLIC meldingen gegenereerd.
Zet dit bestand in de directory *data_klic_historisch*

Het bestand klic_historisch.txt dat als invoer voor het model wordt gebruikt, kan nu gemaakt worden. Dit is een samenvoeging van de drie meest recente bestanden uit de directory *data_klic_historisch*. (Samenvoeging van KLIC-bestanden : openen laatste 3 half-jaren en kopiëren / plakken steeds op nieuwe regel).

6.4 Parameteroptimalisatie:

Bij het nieuwe bestand klic_historisch.txt kan een nieuwe set optimale parameters berekend worden:

- zet de nieuwe klic_historisch.txt in de directory parameteroptimalisatie/invoer
- start het bestand parameteroptimalisatie.exe

Dit proces duurt ook weer een aantal uur (afhankelijk van de processorsnelheid) uitvoerbestand is het bestand. *parameters txt*

De bestanden klic_historsich.txt (met toevoeging schade ja/nee) en parameters.txt zijn de bestanden die als invoer dienen voor het wekelijks te draaien model (toevoegen in directory t.b.v. model)

Bijlage 7: Handleiding Model

Deze handleiding bevat een simpele gebruiksaanwijzing van het programma.

Het excel-bestand dat gewoonlijk opgestuurd wordt naar de rayons, wordt hierdoor vervangen door een excel-bestand met twee extra velden (één met de risico-factor, één met glasinformatie)

Wekelijkse (of twee keer per week) moet het volgende gebeuren:

- Sla het .xls-bestand met KLIC meldingen op in de directory *risico calculator* als 'geo.xls'.
- Open macro.xls en druk op de knop 'Opschonen geoversumdata (excel-bestand)'.
- Wanneer er al een tekstbestand geo.txt in de directory staat zal de volgende vraag gesteld worden:

  'Er bestaat in deze locatie al een bestand met de naam geo.txt. Wilt u dit bestand vervangen?

  Klik dan op 'Ja'.
- Vervolgens wordt de vraag gesteld:

  'Wilt u de wijzigingen in geo.txt opslaan?'

  Klik dan op 'Nee'

- Controleer of de volgende bestanden in de directory *invoer* staan:
  - werkzaamhedentabel.txt
  - parameters.txt
  - glas.txt
  - aannemertabel.txt
  - klic_historisch.txt

- Start het model door op het bestand calculator.exe te klikken.

- Wanneer het model uitgedraaid is, bevat het tekstbestand output.txt de KLIC meldingen inclusief de risicofactor. Dit bestand kan in worden gelezen in excel:
  - Scheidingsteken: puntkomma
  - Tekstindicator: "

- Dit excelbestand wordt het bestand dat opgestuurd wordt naar de rayons.