Onderzoeksrapport

Mogelijkheden predictive maintenance binnen Rijkswaterstaat

Verkenning van SCADA gegevens binnen 21 bruggen

afstudeerscriptie van Roel van Endhoven

studentnummer 00048986

in het kader van de opleiding HBO-ICT aan de

HZ University of Applied Sciences, Vlissingen

Onderzoeksrapport

Mogelijkheden predictive maintenance binnen Rijkswaterstaat

Verkenning van SCADA gegevens binnen 21 bruggen

Versiebeheer

Versie	Auteur	Omschrijving
0.1	R.v.E.	Onderzoeksvoorstel, eigen concepten, inleiding en probleemstelling
1.0	R.v.E.	Onderzoeksvoorstel, eerste formele concept
1.5	R.v.E	Onderzoeksvoorstel, eigen concepten Spelfouten verwijderd, doelstelling beschrijft geen onderzoeksmethode meer, hoofdvraag aangepast op voorstel van Stagedocent.
2.0	R.v.E	Onderzoeksvoorstel, tweede formele concept
2.1	R.v.E	Onderzoeksrapport, eerste concept
2.2	R.v.E	Onderzoeksrapport, verbeterde methode en algehele spellingscontrole. Terminologie geüniformeerd.
2.3	R.v.E.	Resultaten deelvraag 1 t/m 3 toegevoegd.
2.4	R.v.E.	Resultaten deelvraag 4 en totaalbeeld toegevoegd.
2.5	R.v.E.	Discussie en Conclusies toegevoegd
2.6	R.v.E.	Zinsbouw en spelfouten verwijderd.
3.0	R.v.E.	Feedback Jeroen van den Broeke verwerkt.
4.0	R.v.E.	Feedback Daan op laatste concept verwerkt.

Samenvatting

Rijkswaterstaat is de uitvoeringsorganisatie van het Ministerie van Infrastructuur en Waterstaat. Zij beheren binnen Nederland de rijkswegen, -vaarwegen en –wateren en ontwikkelen deze ook. Dit maakt dat Rijkswaterstaat in het bezit is van een breed scala aan infrastructurele assets, zoals bruggen en sluizen. Het onderhoud van deze assets is van groot belang voor de integriteit en het functioneren van de infrastructuur binnen Nederland. Dit feit is leidend bij de uitvoering van verschillende pilot projecten van Rijkswaterstaat binnen Nederland. Eén van deze pilots is Tilburg 3, deel van het Vitale Assets, voorspelbaar onderhoud project. Het doel van dit project is om het onderhoud aan objecten in het bezit van Rijkswaterstaat beter te kunnen timen door te sturen naar datagedreven predictive maintenance.

Het doel van dit onderzoek is de inzetbaarheid van de beschikbare SCADA-gegevens van eenentwintig bruggen in regio Zuid Nederland in kaart brengen voor het doeleind predictive maintenance. Om deze doelstelling te bereiken is de volgende onderzoeksvraag opgesteld: *“In hoeverre is het mogelijk om met de beschikbare SCADA gegevens predictive maintenance toe te passen?”*.

Het onderzoek is vervolgens in vier stappen doorlopen. Eerst is de doelstelling van Rijkswaterstaat uitgediept door overleg te plegen met verschillende stakeholders binnen het vitale assets project. Hieruit is een lijst met eisen en user stories opgesteld welke de richting bepalen voor de rest van het onderzoek. Deze richting is het zoeken naar verbanden in het storingsgedrag van de bruggen. De stap die daarop volgde was het in kaart brengen van de kwaliteit van de beschikbare gegevens om een beeld te creëren over de mate waarin deze ingezet kan worden voor predictive maintenance. Hieruit bleek dat dertien van de eenentwintig bruggen niet consistent hebben gelogd, dit beslaat 12.5% van de gegevens. Eén brug (Brug Son) mist zesentwintig dagen aan loggegevens en scoort 89.03% op de mate van compleetheid. Vervolgens is gezocht naar literatuur over de toepassingen van predictive maintenance op SCADA loggegevens. Vanuit deze literatuur is een lijst van modellen opgesteld die in SCADA gegevens verbanden zoeken. Daarnaast zijn meetinstrumenten vastgesteld om het toegepaste model aan te toetsen. Tot slot is een model uitgewerkt tot een proof of concept en is hiermee een experiment gedaan op de beschikbare SCADA gegevens. De uitkomsten van het experiment geven geen verbanden aan tussen SCADA alarmen die gebruikt kunnen worden voor het toepassen van predictive maintenance. Het onderzoek geeft geen uitsluitsel dat met een grotere dataset ook geen verbanden kunnen worden gevonden. Vanuit deze conclusie wordt aanbevolen de dataset te vergroten en andere datasets te combineren om zo een bredere set aan gegevens te creëren.

Summary

Rijkswaterstaat is the executive organisation that is part of the Dutch Ministry of Infrastructure and Water Management. They are responsible for the design, construction, management and maintenance of the main infrastructure facilities in the Netherlands. Therefore, Rijkswaterstaat is in possession of a wide range of infrastructural assets like bridges and sluices. Maintenance of these assets is of critical importance for the integrity and functioning of the Dutch national infrastructure. This is the leading motivator for the deployment of pilot projects managed by Rijkswaterstaat. One of these pilot projects is Tilburg 3, part of the Vital Assets, predictive maintenance project. The goal of this project is improving the timing of maintenance by steering towards a data driven predictive maintenance strategy.

The goal of this research report is to explore the employability of SCADA data gathered from twenty one bridges in the South Netherlands region (Zuid Nederland) for the purpose of predictive maintenance. To reach this goal, the following research question was formulated: *"To what extent can predictive maintenance be employed given the available SCADA data"*.

The research was conducted and four steps were taken. The first being the exploration and mapping of the business goals Rijkswaterstaat has with regards to predictive maintenance and SCADA data. This was done by consulting different stakeholders within the vital assets project. The goals were mapped to requirements and were recorded in a list. The main requirement this research focusses on is finding patterns and relations in the failure behavior recorded in the SCADA logs. The next step in the research was recording the quality of the available data. This way a statement can be made about the extent predictive maintenance can be employed on the data. The results of this step in the research project showed gaps in the consistency of logged data. In thirteen of the twenty one bridges there was a gap of 12.5% in the data logs. One bridge (Brug Son) produces a completeness score of 89.03% due to missing data. The following step in the research was conducting a literature study towards finding applications of predictive maintenance on SCADA systems. Models which accommodate such applications were found and were recorded in a list. Measures of interestingness were also recorded so that the model could be tested after implementation. The last step of the research was implementing the model into a proof of concept. This proof of concept was used to conduct an experiment by searching for patterns in the SCADA data for all twenty one bridges. The results of this experiment show no patterns that can be used for a predictive maintenance strategy. However, the results give no clear answer that the model will not produce interesting rules given a larger dataset. It is therefore recommended that the dataset be expanded and that other datasets are combined to create a larger set of data.

Voorwoord

Deze scriptie beschrijft mijn onderzoek naar de toepassing van data mining en predictive maintenance op SCADA gegevens. Dit onderzoek heb ik in opdracht van Rijkswaterstaat uitgevoerd in de periode september tot en met december van 2018.

Het uitvoeren van het onderzoek was voor mij een zeer interessante verdieping in data mining, met in het specifiek association rule mining. Daarbinnen heb ik mijn kennis van set theorie flink moeten vergroten. Het onderzoekstraject was een zeer leuke en leerzame ervaring.

Graag neem ik hier de tijd om mijn begeleiding te bedanken. Dit zijn vanuit Rijkswaterstaat: Gilbert Westdorp en Jeroen van den Broeke. Zij hebben mij de weg gewezen binnen Rijkswaterstaat en waren altijd beschikbaar voor vragen of feedback op mijn werk. Vanuit school zijn dit: Daan de Waard en Jolène Cijsouw. Ik wil hen bedanken voor het leveren van feedback op mijn werk en het helpen met het invullen en structureren van mijn onderzoek.

Roel van Endhoven,

Breskens, 4 december 2018

Inhoud

1. Inleiding

Dit document beschrijft de methode en uitvoering van het onderzoek naar de mogelijkheid voor het toepassen van predictive maintenance methodieken op SCADA gegevens van eenentwintig bruggen in Zuid Nederland. Gebaseerd op de resultaten van het onderzoek wordt vervolgens een conclusie opgesteld en worden aanbevelingen gedaan voor vervolgonderzoek. Eerst volgt een beschrijving van de organisatie voor welke het onderzoek wordt verricht. Daarnaast wordt de aanleiding van het onderzoek beschreven. Daarop volgt een beschrijving van de doelstelling. De aanleiding en doelstelling vormen de probleemstelling waarop het onderzoek berust. Uit deze probleemstelling is een onderzoeksvraag opgesteld welke verder uiteenvalt in deelvragen. De beantwoording van de deelvragen vormen daarmee het antwoord op de onderzoeksvraag.

Het onderzoek wordt verricht in opdracht van Rijkswaterstaat, regio Zee en Delta. Rijkswaterstaat is de uitvoeringsorganisatie van het Ministerie van Infrastructuur en Waterstaat. Zij beheren binnen Nederland de rijks- vaarwegen en wateren en ontwikkelen deze ook. Dit maakt dat Rijkswaterstaat in het bezit is van een breed scala aan infrastructurele assets, zoals bruggen en sluizen. Het onderhoud van deze assets is van groot belang voor de integriteit en het functioneren van de infrastructuur binnen Nederland.

1.1 Aanleiding

Rijkswaterstaat heeft de visie om meer te doen met de gegevens en informatie die zij in bezit hebben. In de Informatievisie (I-visie) van Rijkswaterstaat wordt benadrukt dat er veel waarde wordt gehecht aan het inzetten van Big Data binnen de bedrijfsvoering van Rijkswaterstaat (Rijkswaterstaat, 2017). Informatie is leidend bij de uitvoering van verschillende pilot projecten van Rijkswaterstaat binnen Nederland. Eén van deze pilots is Tilburg 3. Deze pilot maakt deel uit van het project Vitale Assets, Voorspelbaar Onderhoud. Het doel van dit project is om het onderhoud aan objecten in het bezit van Rijkswaterstaat beter te kunnen timen door te sturen naar datagedreven predictive maintenance.

De pilot Tilburg 3 bestaat uit de gegevens die worden gegenereerd door 21 bruggen in regio Zuid Nederland, specifiek Brabant en Limburg. Deze bruggen zijn voorzien van een PLC besturingssysteem welke data verzamelen over de staat van functioneren. Dit gaat over duizenden records aan data die per minuut gegenereerd worden. Sinds 1 februari 2018 worden deze gegevens maandelijks ontsloten op locatie. Dit gaat nu over een totaal van ongeveer tien miljoen regels data. Deze gegevens worden verzameld en opgeslagen, maar verder wordt er nog niets met deze gegevens gedaan. Er is daarmee nog geen duidelijkheid in hoeverre deze gegevens bruikbaar zijn voor het toepassen van predictive maintenance. Predictive maintenance is een onderhoudsstrategie die zich richt op het voorspellen van defecten in apparatuur door voorspellende analyses (Technopedia, z.d.).

1.2 Probleemstelling

Rijkswaterstaat verzamelt in regio Zuid Nederland met behulp van sensoren gegevens vanuit de Supervisory Control And Data Acquisition (SCADA) systemen waarmee de eenentwintig bruggen bediend worden. Dit gaat om gegevenssets met miljoenen regels data. Ook worden daarnaast energieverbruiksgegevens op de verschillende objecten (bruggen) geregistreerd. Momenteel wordt er met de verzamelde gegevens niets gedaan. Er zijn vanuit Rijkswaterstaat nooit eisen gesteld aan welke data de sensoren en het SCADA systeem precies rapporteren en welk formaat deze gegevens

hebben. Hierdoor is er onzekerheid over de operationaliseerbaarheid van de data en de inzetbaarheid van het huidige systeem als instrument voor predictive maintenance.

1.2.1 Doelstelling

Het doel van dit onderzoek is de inzetbaarheid van de beschikbare SCADA-gegevens in kaart brengen voor het doeleind predictive maintenance. Op deze manier kan een uitspraak worden gevormd over de mate waarin de SCADA-gegevens zich lenen voor predictive maintenance. Deze inzichten kunnen vervolgens gebruikt worden in het vervolg van het Tilburg 3 project om een systeem in te richten dat predictive maintenance beter faciliteert.

1.2.2 Relevantie

1.2.2.1 Bedrijfsrelevantie

Het werken volgens predictive maintenance stelt Rijkswaterstaat in staat om tijdiger te kunnen handelen wanneer gezien wordt dat een object dreigt defect te gaan. Tevens kan periodiek onderhoud beter worden ingepland. Zo wordt voorkomen dat onderhoud onnodig gebeurt of dat het onderhoud te laat komt. Dit bespaart kosten en tijd. Daarnaast levert het inzicht in de staat van objecten een manier om de aannemers die onderhoud uitvoeren te controleren op accuraatheid.

1.2.2.2 Maatschappelijke relevantie

Naast Rijkswaterstaat zijn er vele andere eigenaren van infrastructurele objecten. Gemeenten, waterschappen en provincies bezitten eigen wegen en bruggen. Inzetten van predictive maintenance kan bij hen ook een besparing van kosten betekenen. Daarbij vormt het reduceren van down time door het voorspellen van defecten ook voordeel voor logistieke organisaties. Bruggen bieden betere doorstroming voor vervoer en vaarwegcorridors waardoor bestemmingsverkeer tijdiger op plaats van bestemming arriveert.

1.2.2.3 Kennisgebied en theoretische relevantie

De uitkomst van het onderzoek biedt een raamwerk voor de benodigdheden om predictive maintenance in te zetten binnen infrastructurele objecten, specifiek bruggen. Deze kennis kan vervolgens worden toegepast of uitgebreid met vervolgonderzoek over de toepassing op andere objecten of assets.

1.2.3 Vraagstelling

1.2.3.1 Centrale vraag

De vraag die dit onderzoek probeert te beantwoorden luidt:

“In hoeverre is het mogelijk om met de beschikbare SCADA gegevens predictive maintenance toe te passen?”

1.2.3.2 Deelvragen

De hiervoor genoemde vraag valt uiteen in de volgende deelvragen:

1. *Wat is de doelstelling van het toepassen van predictive maintenance op het SCADA systeem?*
2. *Wat is de kwaliteit van de beschikbare gegevens?*
3. *Welke data mining technieken zijn beschikbaar om deze doelstellingen te behalen?*
4. *Hoe kunnen deze data mining technieken worden ingezet op de SCADA gegevens?*

De opbouw van de deelvragen is zodanig dat zij een pad vormen voor het analyseren van de gegevens en de eisen aan het predictive maintenance systeem te kunnen realiseren. De eerste deelvraag zoekt naar een antwoord op de doelstelling van predictive maintenance binnen Rijkswaterstaat en de toepassing daarvan op de SCADA gegevens. Het antwoord op deze deelvraag geeft richting aan welke technieken worden toegepast om de doelstelling te bereiken. De tweede deelvraag gebruikt deze bevindingen om verkennend onderzoek te doen naar de gegevens en om de kwaliteit van de beschikbare gegevens in kaart te brengen. Dit brengt inzicht in de inzetbaarheid van de gegevens voor de toepassing van predictive maintenance. De derde deelvraag beoogt antwoord te geven op de mogelijkheid om eisen te stellen aan data mining methodes en deze toe te passen op de gegevens, alsmede het in kaart brengen van verschillende data mining technieken om toe te passen op de huidige gegevens. De vierde deelvraag is opgesteld om de gegevens om te vormen zodat deze gebruik makend van de gevonden data mining technieken kunnen worden ingezet om antwoord te geven op de hoofdvraag.

1.3 Scope

Gezien de omvang van het onderzoek en de brede vraagstelling wordt de scope van het onderzoek vastgelegd. Dit wordt gedaan in overeenstemming met Rijkswaterstaat. Hieruit zijn de volgende voorwaarden gekomen.

1.3.1 CRISP-DM

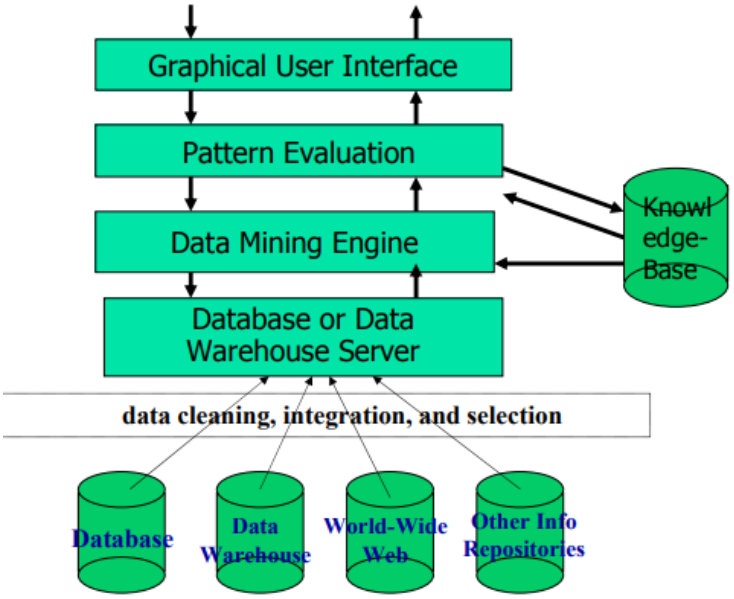
Het onderzoek zal conform de methodieken opgesteld in CRISP-DM worden gepleegd. CRISP-DM, ofwel Cross Industry Standard Process for data mining (Chapman et al., 2000), is een framework en handleiding die handvatten biedt voor het doorlopen van een data mining proces. Dit framework wordt gedurende het onderzoek aangehouden om vorm te geven aan het proces.

2. Theoretisch Kader

Dit hoofdstuk geeft een overzicht van de beschikbare theorie in relatie tot het onderzoek. Daarnaast geeft het een overzicht van de belangrijkste concepten die tijdens het onderzoek aan de orde komen. De samenhang tussen deze onderdelen wordt per onderwerp toegelicht.

2.1 Data mining

Data mining is het extraheren en ontdekken van bruikbare informatie uit een dataset (Khan, Mohamudally, & Babajee, 2013). Dit proces leidt tot kennis welke helpt bij besluitvoering binnen organisaties. Eén van de taken die hierbij hoort is het zoeken naar patronen en overeenkomsten vanuit data opgeslagen in een database (Yao, 2001). De grote hoeveelheden data die tegenwoordig worden opgeslagen zijn moeilijk om op zichzelf te interpreteren. Vanuit de patronen die gevonden worden in deze 'low level' data kunnen abstracties worden gemaakt zoals rapporten, maar ook voorspellende analyses om te benaderen hoe de gegevens zich in de toekomst gaan ontwikkelen (Fayyad, Piatetsky-Shapiro, & Smyth, 1996a). Een typische architectuur voor een dataminingsysteem is opgenomen in figuur 1.



Figuur 1: Structuur Datamining systeem (Han, Kamber, 2006)

2.2 Predictive maintenance

De toepassing van data mining binnen dit onderzoek wordt toegespitst op de verkenning van mogelijkheden die toewerken naar predictive maintenance. Predictive maintenance is de benaming voor het voorspellen van onderhoud aan objecten. In figuur 2 is afgebeeld waar het principe van predictive maintenance valt binnen het Analytics Maturity Model van Gartner (Gartner, 2013). Dit model geeft de sturingsrichting weer van analyse strategieën.

Visuele inspectie van een object met als doel een defect voor te zijn is de oudste en meest toegepaste vorm van predictive maintenance. Moderne toepassingen van predictive maintenance bestaan uit objecten voorzien van sensoren welke continue of periodieke metingen verrichten op basis van welke wordt besloten om onderhoudsacties uit te voeren (Carnero, 2006). Met de groei van datagedrevenheid en de toename van technologische kennis is het vakgebied uitgegroeid tot het

toepassen van geautomatiseerde oplossingen waaronder patroonherkenning en neutrale netwerken (Hashemian & Bean, 2011).

Figuur 2: Analytics Maturity Model (Gartner, 2013)

De moderne variant van predictive maintenance wordt onderverdeeld in twee sub categorieën:

- **Statistisch gebaseerd predictive maintenance:** Bij deze variant wordt informatie verzameld van alle stremmingen en storingen binnen een systeem om statistische modellen te ontwikkelen (Wang, 2016). Met deze modellen kunnen defecten worden voorspeld om zo een predictive maintenance beleid te creëren (Hashemian & Bean, 2011).
- **Conditie gebaseerd predictive maintenance:** Bij conditie gebaseerd predictive maintenance wordt het slijtageniveau van mechanische componenten gemonitord. De slijtage veroorzaakt veranderingen in het gedrag van de machine (het systeem) die niet direct leiden tot mechanische storing (Hashemian & Bean, 2011).

2.3 Knowledge discovery

Binnen data mining is knowledge discovery het hoofddoel. Knowledge discovery is het extraheren van bruikbare informatie uit grote hoeveelheden data. Kurgan & Musilek (2006) stellen dat voordat een poging gedaan kan worden om bruikbare informatie te extraheren eerst een aanpak moet worden geformuleerd over hoe dit proces wordt ingedeeld. Het opstellen van een knowledge discovery processtructuur kwam vanuit de problemen die ontstonden door het blind toepassen van data mining op gegevenssets. Dit principe heet data dredging en dit kan leiden tot het ontdekken van nutteloze informatie voor de organisatie (Fayyad, Piatetsky-Shapiro, & Smyth, 1996b). Er zijn verschillende van deze methodieken en processen.

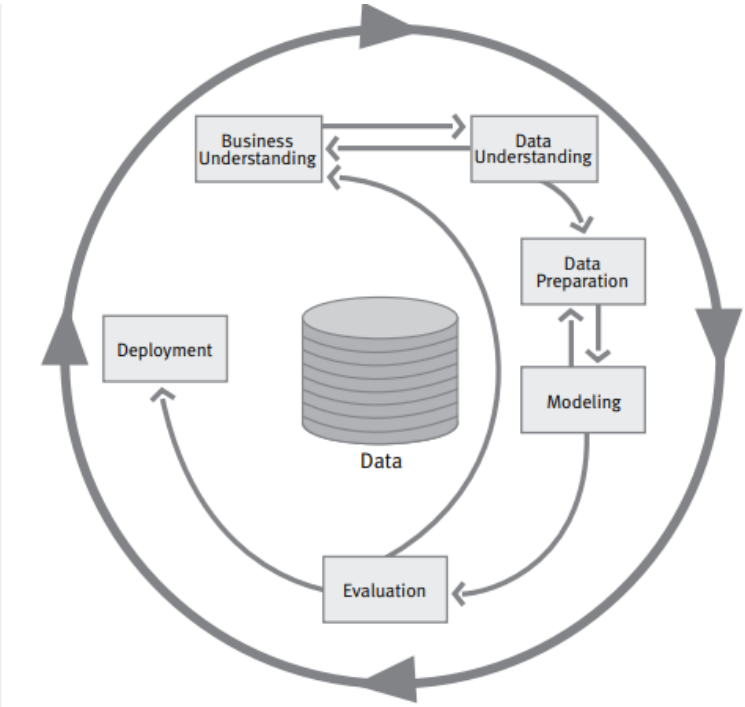
2.4 SCADA

Supervisory Control and Data Acquisition (SCADA) systemen worden ingezet om over grote afstanden objecten te beheren. Daarbij is het verkrijgen en beheren van data een kritiek onderdeel van het functioneren van een SCADA systeem. Daarnaast worden SCADA beheer centra ingezet om vanuit

een centraal punt objecten te beheren en monitoren over lange afstanden (Stouffer, K., & Falco, J. 2006). SCADA systemen communiceren met Programmable Logic Controllers (PLCs) binnen een systeem en kunnen deze aansturen en uitlezen, dit gebeurt in scan cycli. PLCs voeren iedere *n* milliseconden een scan cyclus uit waar de status van sensors wordt uitgelezen of onderdelen worden aangestuurd. Het SCADA systeem voert hiernaast ook scan-cycli uit welke het geheugen van de PLCs weer uitlezen of manipuleren. Daarnaast wordt SCADA gebruikt voor het rapporteren van fouten of het vastleggen van event sequenties (Gaushell & Darlington, 1987).

3. Methode

Binnen het onderzoek wordt de CRISP-DM methodologie voor data mining gehanteerd (zoals beschreven in paragraaf 1.3). CRISP-DM onderscheidt binnen het data mining proces zes verschillende fases die meerdere keren cyclisch kunnen worden doorlopen. Een schematisch overzicht hiervan is te vinden in figuur 2. Daarna volgt een toelichting van iedere fase. Bij iedere deelvraag wordt toegelicht onder welke fase van CRISP-DM deze valt en hoe CRISP-DM wordt ingezet om antwoord te geven op de deelvraag.



Figuur 3: CRISP-DM Cyclus (Chapman et al, 2000)

De stappen van de CRISP-DM cyclus vallen als volgt uiteen:

- **Business understanding:** De eerste fase in het in kaart brengen van de doelstellingen van het project en de eisen aan het systeem vanuit een bedrijfsperspectief. Vanuit deze eisen wordt een data mining plan opgesteld om deze doelstellingen te bereiken.
- **Data understanding:** De data understanding fase begint met het verzamelen en verkennen hiervan. Hier worden de eerste inzichten gecreëerd in de kwaliteit van de gegevens en worden interessante onderdelen van de dataset gevonden om hypotheses te vormen.
- **Data preparation:** Binnen de data preparation stap horen alle activiteiten die nodig zijn om de dataset voor te bereiden op het toepassen van statistische of machine learning modellen. Hierbij hoort ook het opschonen en transformeren van de gegevens.
- **Modeling:** In de modeling fase worden de gegevens verwerkt in statistische of machine learning modellen. Deze stap vereist vaak dat er teruggegaan wordt naar de Data Preparation stap om te voldoen aan specifiekere eisen van sommige modellen.
- **Evaluation:** De evaluation fase bestaat uit het controleren van het model en het vergelijken met de in de business understanding opgestelde doelstellingen en eisen. Binnen dit onderzoek bevat deze stap ook het evalueren van de bruikbaarheid van de gevonden informatie.

- **Deployment:** Binnen de deployment fase vallen alle werkzaamheden om het model en de bevindingen van het dataminingproces presenteerbaar te maken aan de klant. Dit kan verschillen van het opstellen van een rapport tot het implementeren van een real-time dashboard.

3.1 Deelvraag 1

De eerste deelvraag luidt: *"Wat zijn de doelstellingen van het toepassen van predictive maintenance op het SCADA systeem?"*

Om deze vraag te beantwoorden wordt deze geoperationaliseerd. Dit gebeurt aan de hand van het uitzetten van het meetinstrument en de producten die deze stap in het onderzoek oplevert. Daarnaast wordt de relatie beschreven binnen het CRISP-DM proces.

3.1.1 Meetinstrument en operationalisatie

Het in kaart brengen van de doelstellingen valt binnen de business understanding fase van het CRISP-DM proces en vormt daarmee de eerste stap van het onderzoekstraject. Om de doelstelling vast te stellen wordt met verschillende stakeholders binnen het Vitale Assets project gesproken en wordt achterhaald wat de doelstelling is en welke informatie uit de SCADA gegevens gehaald dient te worden. Dit wordt gedaan aan de hand van een verkenning van gebruikerseisen. Besprekingen worden in hoofdlijnen vastgelegd en bevindingen worden genoteerd, waar de focus wordt gelegd op de vraag naar de doelstelling van het project. Op deze manier wordt vastgelegd hoe Rijkswaterstaat aankijkt tegen predictive maintenance en hoe er invulling gegeven kan worden aan de doelstellingen binnen het vitale assets project.

De besprekingen volgen de volgende semi gestructureerde vorm om de richting vast te stellen. De volgende punten komen overeen met deelproducten van de business understanding fase van het CRISP-dm proces, vastgelegd in een verslag:

- Wat is de achtergrond van het project?
- Wat is de doelstelling van het project?
- Welke inzichten wil Rijkswaterstaat verkrijgen vanuit de SCADA gegevens?

De antwoorden op deze vragen worden verwerkt in dit verslag.

3.1.2 Onderbouwing keuze meetinstrumenten

Doordat er veel betrokken stakeholders zijn binnen het vitale assets project, waaronder asset managers, maar ook de afdelingen netwerkontwikkeling en netwerkmonitoring, worden er meerdere besprekingen gehouden met verschillende stakeholders. Zo kan de doelstelling voor meerdere stakeholders in kaart worden gebracht en kan richting gegeven worden aan de informatiebehoefte vanuit de SCADA gegevens. Om de mogelijkheid van de toepasbaarheid van predictive maintenance voor Rijkswaterstaat goed in kaart te kunnen brengen is het van belang dat er eenduidigheid is over hoe predictive maintenance wordt opgevat en welke informatie meerwaarde levert voor het project. Dit kan alleen door met verschillende stakeholders en experts te spreken.

3.1.3 Analysemethodes

De analysemethodes die gebruikt worden om tot antwoord van deze deelvraag te komen zijn:

1. De bevindingen uit de requirements verkenning worden omgevormd tot user stories.
2. Vanuit de user stories worden functionele eisen opgesteld welke worden beperkt tot de toepassing van predictive maintenance.
3. De opgestelde requirements worden geprioriteerd aan de hand van de MOSCOW methode (Waters, 2009).

De doelstelling wordt uitgedrukt in user stories in het formaat:

“Als <rol> wil ik <actie/kenmerk> om <reden>”.

(“Als data-analist wil ik het aantal storingen kunnen zien van brug x op dag x om inzicht te krijgen in het functioneren op dag niveau“).

Op deze manier wordt de doelstelling achter het data mining proces door middel van een standaard methode in kaart gebracht. De user stories worden geprioriteerd volgens de MOSCOW methode, welke onderscheid maakt in vier verschillende niveaus, namelijk Must, Should, Could en Won’t.

Dit geeft aan welke eisen er absoluut zijn aan het systeem, ofwel welke inzichten gecreëerd moeten worden.

Tot slot wordt er van deze lijst user stories een shortlist gemaakt. Dit is een lijst met de vijf hoogst geprioriteerde eisen binnen dit onderzoek. Dit is om de scope van het onderzoek te beperken en om gerichtere analyse te kunnen doen over de belangrijkste gebruikseisen.

3.1.4 Producten

Deze stap produceert een requirementsdocument met daarin verslag van de gehouden besprekingen en de herleiding van user stories en functionele eisen vanuit de gevonden achtergrond en doelstellingen van het project.

3.2 Deelvraag 2

De tweede deelvraag luidt als volgt: “*Wat is de kwaliteit van de beschikbare gegevens?”.*

De deelvraag wordt geoperationaliseerd zoals te vinden in paragraaf 3.2.1. Daar wordt ook de relatie tot het CRISP-DM proces beschreven.

3.2.1 Meetinstrument en operationalisatie

Om antwoord te geven op de tweede deelvraag worden de SCADA gegevens verkend en wordt de kwaliteit van de gegevens in kaart gebracht. Dit wordt gedaan aan de hand van het opstellen van een analyse waarin de aspecten van de gegevens worden vastgelegd. Dit valt binnen de data understanding fase van het CRISP-DM proces. Dit is een vorm van exploratief onderzoek en een combinatie van desk research in de vorm van een data analyse.

Askham et al. (2013) definiëren zes kwaliteitsdimensies voor datakwaliteit. Deze dimensies zijn als volgt:

- Compleetheid
- Consistentie
- Uniekheid

- Validiteit
- Nauwkeurigheid
- Tijdigheid

Deze dimensies worden gebruikt om een percentuele score te geven aan de kwaliteit van de gegevens.

3.2.2 Onderbouwing keuze meetinstrumenten

Het gekozen meetinstrument creëert voor de onderzoeker en voor de organisatie inzicht in welke gegevens er beschikbaar zijn en hoe deze in elkaar zitten. Het verkennen van de gegevens geeft inzichten in welke richting het onderzoek kan gaan en welke modellen of algoritmes toegepast kunnen worden voor het inzetten van predictive maintenance van de assets. Daarnaast geeft het voor de organisatie inzicht in de volledigheid van de data verzamel strategie. De score die uit het model komt geeft per dimensie een herhaalbaar proces om datakwaliteit te meten.

3.2.3 Analysemethodes

De zes kwaliteitsdimensies worden geoperationaliseerd in het datakwaliteitsrapport. De implementatie van de analyse strategie is in bijlage B: kwaliteitsrapport, uitgewerkt per kritische kwaliteitsdimensie.

3.2.4 Producten

De producten die worden opgesteld voor deze deelvraag is een datakwaliteitsrapportage op basis van de zes dimensies van Askham et al. (2013).

3.3 Deelvraag 3

De derde deelvraag is: "*Welke data mining technieken zijn beschikbaar om deze doelstellingen te behalen?".*

Om deze vraag te beantwoorden wordt literatuuronderzoek verricht naar beschikbare data mining oplossingen voor het werken met SCADA event logs. Het in kaart brengen van verschillende technieken die aansluiten op de doelstellingen van Rijkswaterstaat zal vormt het antwoord op de deelvraag.

3.3.1 Meetinstrumenten en operationalisatie

Om deze deelvraag te operationaliseren zal een literatuuronderzoek worden uitgezet. Daarbij worden alle geselecteerde user stories genomen en wordt hier vanuit gezocht naar data mining technieken om de doelstellingen van de user stories te dekken. Er wordt bij deze vraag gezocht naar algoritmes die toepasbaar zijn op de SCADA dataset. Dit leidt tot een lijst met toepasbare modellen of algoritmen om de gegevens mee te kunnen analyseren. Om tot deze lijst te komen wordt een zoekplan opgesteld. Dit zoekplan bevat uit de user stories onttrokken sleutelwoorden. Naar deze woorden wordt gezocht en wordt in eerste instantie gekeken naar de toepassingen op SCADA gegevens, binnen data mining technieken.

Het literatuuronderzoek zal op de volgende manier plaatsvinden:

1. Standaard sleutelwoorden voor literatuur onderzoek data mining:
    a. data mining
    b. SCADA
    c. Event log

d. Logging
e. Log mining

2. Sleutelwoorden onttrokken uit user stories worden hierbij gevoegd.

3.3.2 Onderbouwing keuze meetinstrumenten

Het blindelings toepassen van statistische modellen zonder af te wegen of deze wel bij de bedrijfsdoelstellingen horen leidt tot data dredging (Fayyad et al., 1996b). Om te zorgen dat de gegevens onderzocht worden op een manier die meerwaarde oplevert voor Rijkswaterstaat wordt daarom eerst vanuit de user stories gekeken welke inzichten gewenst zijn.

3.3.3 Analysemethode

De gevonden data mining methoden worden opgenomen in een overzicht. Daarbij worden de gevonden modellen afgespiegeld aan de data mining requirements en wordt op basis van geschiktheid een selectie gemaakt die wordt opgenomen in het overzicht. Hiernaast worden meetinstrumenten in kaart gebracht om de modellen te toetsen.

3.3.4 Producten

Deze stap in het onderzoek levert de volgende producten op:

- Document met lijst van gevonden data mining technieken.
- Omschrijving gevonden data mining methodes en keuze voor een techniek ontleend aan de analyse stap.
- Meetinstrumenten voor het toetsen van het gekozen model.

3.4 Deelvraag 4

De vierde deelvraag is als volgt: *“Hoe kunnen deze data mining technieken worden ingezet op de SCADA gegevens?”*.

Om deze deelvraag te beantwoorden wordt deze geoperationaliseerd zoals te lezen in paragraaf 3.4.1

3.4.1 Meetinstrument en operationalisatie

In deze stap van het onderzoekstraject worden de SCADA gegevens gevormd tot een model dat tracht uitspraak te doen over het functioneren van de bruggen. Deze stap van het onderzoek vormt een experiment over de toepasbaarheid van het model en resulteert in een prototype in de vorm van een implementatie hiervan. Het model dat wordt gebruikt wordt gekozen op basis van het literatuuronderzoek dat vooraf gaat aan deze deelvraag. De toepasbaarheid van het model wordt vastgesteld door dit te toetsen aan de gevonden meetinstrumenten in deelvraag drie.

3.4.2 Onderbouwing keuze meetinstrumenten

Door het realiseren van een prototype van de gevonden modellen wordt de toepasbaarheid hiervan inzichtelijk gemaakt. Op deze manier kan het model direct worden toegepast op de SCADA gegevens en wordt er direct een uitspraak gedaan over de toepasbaarheid aan de hand van de resultaten die de modellen produceren. Door het model vervolgens te toetsen aan de meetinstrumenten die gevonden zijn in deelvraag drie wordt de mate waarin het model correcte en valide resultaten oplevert vastgesteld.

3.4.3 Analysemethodes

De analysemethodes om de deelvraag te beantwoorden zijn als volgt:

- Een implementatie van het model wordt ontwikkeld.
- Parameters voor het model worden vastgesteld.
- Het model wordt toegepast op de SCADA gegevenssets voor iedere brug met de gegeven parameters.
- De resultaten die het model produceert worden vastgelegd.
- De resultaten worden getoetst aan de gevonden meetinstrumenten.
- Waar mogelijk worden de resultaten gevisualiseerd.

3.4.4 Producten

Het product dat resulteert uit deze stap in het onderzoek is een implementatie van de gevonden modellen. Daarnaast worden ook de resultaten van de toepassing van de modellen in kaart gebracht alsmede de waarde van de meetinstrumenten die gebruikt worden om de modellen te toetsen.

4. Resultaten

De resultaten van het onderzoek geven antwoord op de hoofdvraag: *"In hoeverre is het mogelijk om met de beschikbare SCADA gegevens predictive maintenance toe te passen?".* De vraag wordt beantwoord aan de hand van de resultaten van de beantwoording van alle deelvragen.

4.1 Deelvraag 1

Om antwoord te geven op de eerste deelvraag, *"Wat is de doelstelling van het toepassen van predictive maintenance op het SCADA systeem?",* zijn de doelstellingen van het data mining proces volgens CRISP-DM in kaart gebracht. Hiervoor is overleg geweest met twee verschillende stakeholders van het Tilburg 3 project. Deze stakeholders zijn Robert van Lakwijk en Arnoud de Kruijf, beiden gespecialiseerd in asset management en nauw betrokken met de SCADA gegevens en de eenentwintig bruggen. Deze stakeholders zijn benaderd op advies van Gilbert Westdorp en Jeroen van den Broeke van de afdeling netwerkmonitoring Zee en Delta, en informatiemanager Marcel van Houtert. Zij, gebaseerd op een overleg naar welke betrokkenen er zijn binnen het Vitale Assets project, specifiek het Tilburg 3 contract. Robert van Lakwijk en Arnoud de Kruijf hebben veel kennis van het project en kunnen daarom het meest volledige beeld geven van de eisen en wensen aan het project. Op deze manier, naast de free form inrichting van de interviews, is getracht een zo volledig mogelijk overzicht van eisen te creëren aan een systeem dat analyses uitvoert op SCADA logs. Vanuit de overleggen is de achtergrond van het probleem besproken en zijn doelstellingen in kaart gebracht. Deze doelstellingen zijn uitgewerkt naar user stories. Op basis van deze user stories zijn de data mining goals opgesteld als eisen aan het onderzoek. Opvolgend aan het overleg met Robert van Lakwijk is bij het Datalab Delft een plan opgesteld waar points of interest liggen voor het data minen binnen SCADA gegevens. Aan de hand hiervan zijn de data mining goals geprioriteerd.

4.1.1 Resultaten

In bijlage A is het requirementsdocument opgenomen waar de verslagen van het overleg met Robert van Lakwijk en Arnoud de Kruijf is opgenomen. Vanuit het overleg met Robert van Lakwijk is de achtergrond vastgelegd. Deze is als volgt:

"Het doel van het project is, vanuit onderhoudsmanagement oogpunt, hoe onderhoud op een slimmere manier ingericht kan worden."

De doelstelling binnen het onderzoek wordt vervolgens ingekaderd op de volgende punten:

1. Creëer inzicht in wat er gebeurt met een object aan de hand van de SCADA gegevens / Breng het gedrag van de brug in kaart.
2. Achterhaal waarom dingen gebeuren / Kan worden herleid wat de oorzaak (bron) is van een storing?
3. Kunnen de storingsmeldingen specifieker worden gemaakt door events uit de SCADA gegevens te analyseren?
4. Kunnen oneigenlijke bedienhandelingen worden vastgesteld en in verband gebracht worden met storingen?

Deze doelstelling is vertaald naar user stories. De volgende user stories zijn van toepassing op data mining en zijn richten zich op predictive maintenance door het functioneren van de brug te analyseren:

Tabel 1: user stories vanuit doelstelling

#	user story
U02	Als gebruiker wil ik de storingen van een object kunnen zien om inzicht te krijgen in het functioneren van de brug.
U03	Als gebruiker wil ik abnormaliteiten in storingen kunnen inzien om beter te kunnen sturen op onderhoud.
U06	Als gebruiker wil ik kunnen zien welke andere factoren invloed hebben op het storingsgedrag van een object om beter te kunnen sturen op onderhoud.

4.1.2 Analyse

Vanuit de user stories zijn data mining requirements opgesteld. Deze worden binnen de CRISP-DM methode data mining goals / requirements genoemd. Aan de hand van deze requirements worden de derde en vierde deelvraag ingevuld. In tabel 2 is de lijst met data mining requirements opgenomen. In bijlage A is de relatie te zien met de user stories waar zij bij horen. De requirements zijn geprioriteerd volgens de bespreking bij het Datalab Delft, te vinden in Bijlage A. De belangrijkste gevonden requirement is DMR 2. De vraag naar inzicht in het verband van storingen binnen een object is na bespreking het meest relevant gebleken. Het onderzoek richt zich op de vervulling van deze requirement.

Tabel 2: data mining requirements

Requirement	**Omschrijving**	**Prioriteit**
DMR 1	Het systeem legt aanloop naar storing vast. 1.1 Het systeem legt vast welke alarmen leiden tot een storing 1.2 Het dashboard toont de storingen en aanloop in een overzicht	SHOULD
DMR 2	Het systeem toont patronen in Storingsgedrag. 2.1 Het systeem zoekt verbanden tussen voorkomende alarmen en storingen. 2.2 Het systeem toont verbanden in overzicht.	MUST
DMR 3	Systeem brengt oneigenlijke bedienhandelingen in kaart. 3.1 Het systeem zoekt verbanden tussen bedienpatronen en storingen 3.2 Het systeem toont storingen die gevolg (kunnen) zijn van afwijkend bediengedrag.	COULD
DMR 4	Het systeem classificeert storingen. 4.1 Systeem maakt onderscheid in storingen die uitval veroorzaken en storingen die geen uitval tot gevolg hebben.	COULD

DMR 5	Het systeem zoekt verbanden tussen storingen en andere factoren. 5.1 Weersgegevens ten opzichte van storingsgedrag	COULD

4.2 Deelvraag 2

De tweede deelvraag luidt: *"Wat is de kwaliteit van de beschikbare gegevens?"*. Om antwoord te geven op deze vraag zijn alle SCADA gegevens van alle bruggen afzonderlijk geanalyseerd. Hiervan is een datakwaliteitsrapport opgesteld welke te vinden is in bijlage B. De events van iedere brug zijn getoetst op basis van kwaliteitsdimensies Askham et al. (2013). De kwaliteit van de gegevens wordt vervolgens afgeleid van de hoeveelheid meldingen per dag over de periode februari tot september.

4.2.1 Resultaten

Het in kaart brengen van de aantallen events per dag geeft interessante informatie over de compleetheid en kwaliteit van de gegevens die de bruggen genereren. In Tabel 3 is een overzicht opgenomen van opgestelde scenario's aan welke de data is getoetst. De scenario's zijn besproken met informatiemanager Marcel van Houtert die ze logisch en stuurbaar bevond. Er is gekozen voor het toetsen van gegevens op dagniveau omdat dit op hoog niveau de kwaliteit van de logging illustreert. De operationalisatie van de dimensies is gebaseerd op de templates zoals beschreven door Askham et al. (2013).

Tabel 3: Scenario's Compleetheid SCADA gegevens

Scenario ID	**Dimensie**	**Beschrijving**	**Implementatie**
001	Completeness	Compleetheid wordt gedefinieerd door missende gegevens in kaart te brengen. Dit gaat in het geval van de SCADA gegevens om de dagen waarop events niet zijn gelogd of waar data ontbreken. Dit geeft inzicht in de kwaliteit van de logging op dagniveau	Compleetheid = totaal aantal dagen zonder gegevenslogging / aantal dagen van eerste record naar laatste record
002	Completeness	Een ander geval van missende gegevens is waanneer het geheugen van een brug vol loopt. Op dit moment stopt de brug met gegevens loggen tot er weer geheugen beschikbaar is.	Compleetheid = timedelta(geheugen-vol-melding, eerstvolgende-melding) voor iedere geheugen-vol-melding in log-events
003	Completeness	Inzicht in compleetheid kan ook verkregen worden door simpelweg te tellen hoe vaak het geheugen vol loopt.	Compleetheid = COUNT(geheugen-vol-melding)

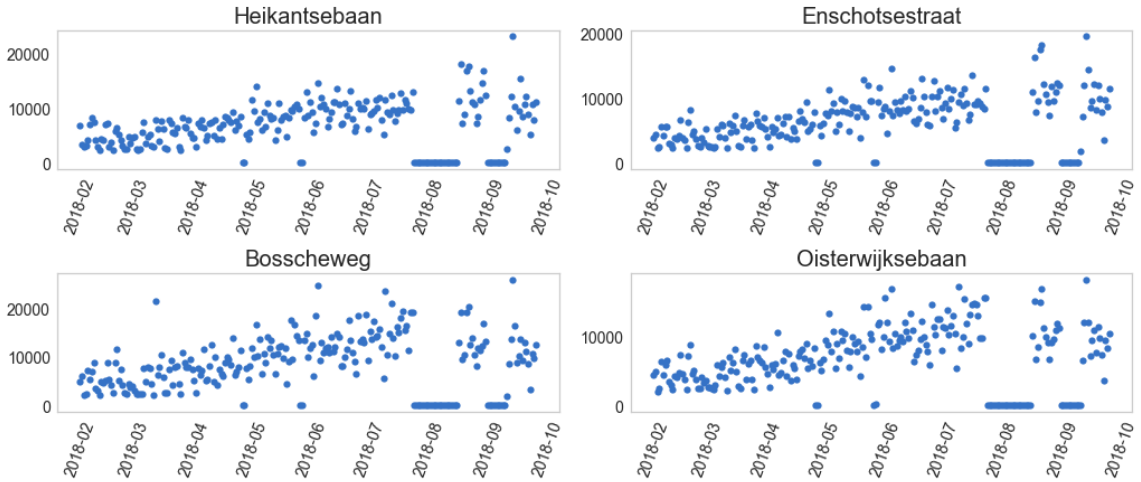
004	Consistency	De mate van consistentie wordt gemeten door te kijken naar of de data voldoet aan de verwachte trends of patronen.	Consistency = Geregistreerde logging is conform het patroon van voorgaande gelogde gegevens

In Tabel 4 is opgenomen hoe de verschillende bruggen scoren op compleetheid. Interessant hieraan is dat de Sonse brug de meeste gaten heeft in logging data en tevens de enige brug is die meldingen heeft over een vol geheugen. Tien procent van het totaal verwacht aantal dagen logging gegevens (237) ontbreekt.

Tabel 4: Score compleetheid gegevens

Brug	**Missende dagen**	**Score**	**Geheugen Vol Meldingen**
Sonse brug	26	89.03%	3
Dr. Deelenlaan	3	98.73%	0
Stadsbrug Weert	3	98.73%	0
Amertakbrug	2	99.16%	0

Om de consistentie (Scenario 004) en validiteit van de SCADA logs in kaart te brengen zijn de gegevens per brug geplot waarin het aantal records per dag is genomen als meetwaarde. In de gegevens is een duidelijk seizoenspatroon te zien bij alle bruggen welke meer gegevens loggen in de zomermaanden. Daarnaast is bij een groot aantal bruggen te zien dat er gaten in de logging lijken te zijn. In de periode half juli tot half augustus worden heel weinig gegevens gelogd. Dit is niet consistent met de rest van de gegevens welke een groeiend aantal events voorspelt. In Figuur 4 is een deel van deze resultaten opgenomen ter illustratie (Dit zijn nieuwere gegevens dan opgenomen in het datakwaliteitsrapport, waardoor ook een gat in september zichtbaar is).



Figuur 4: Missende gegevens logging bruggen.

4.2.2 Analyse

Het in kaart brengen van missende gegevens en gaten in de consistentie van de gelogde gegevens geeft een antwoord op de vraag: “Wat is de kwaliteit van de beschikbare gegevens”. Er is een substantieel aantal bruggen dat apart gedrag vertoont bij het loggen van gegevens. De Sonse brug mist zesentwintig aaneengesloten dagen aan data. Daarnaast zijn er bij verschillende bruggen gaten gevonden in de maanden juli en augustus waarop slechts enkele records worden gelogd en geen storingen of openingen worden geregistreerd. Het gat in deze logging loopt op tot bijna een maand. Welke op de beperkte dataset van februari tot en met september een substantieel gat in de data veroorzaakt, namelijk 12,5%. Het is belangrijk dat de data consistent wordt gelogd wanneer er voorspellingen gemaakt worden over het functioneren van assets.

4.3 Deelvraag 3

Om antwoord te geven op de derde deelvraag is een literatuuronderzoek gedaan naar de mogelijkheden van data mining op SCADA gegevens. Hierbij is eerst in kaart gebracht welk vooronderzoek er is gedaan van het toepassen van data mining op SCADA event logs. Vanuit deze bevindingen zijn modellen gezocht om dit proces te operationaliseren. Deze gevonden modellen zijn opgenomen in een lijst. Het zoekplan dat is gehanteerd voor het literatuuronderzoek is te vinden in bijlage C: “Zoekplan Bronnenonderzoek: Toepassing Data Mining op SCADA gegevens en beschikbare modellen”. De gekozen oplossing uit het bronnenonderzoek wordt verder uiteengezet in bijlage D: “Overzicht Pattern mining algoritmes”. Met de opgedane kennis kan een proof of concept worden gerealiseerd om Data Mining Requirement 2 (DMR 2) te vervullen.

4.3.1 Resultaten

Er zijn verschillende onderzoeken uitgevoerd over het toepassen van data mining technieken binnen SCADA systemen om inzicht te verkrijgen in het functioneren van civieltechnische bouwwerken. Verma (2012) beschrijft een onderzoek naar predictive maintenance in windmolens. Binnen dit onderzoek worden statuspatronen van wind turbines geïdentificeerd door middel van een frequentieanalyse van foutcodes in het SCADA systeem. Hieruit wordt door gebruik te maken van Association Rule Mining (ARM) gezocht naar interessante relaties, gebruikmakend van het Apriori algoritme voor het vinden van frequente statuspatronen. Czora et al. (2017) verkennen de mogelijkheden van rare itemset mining en ARM op systeemniveau. Zij stellen dat functioneren van verschillende deelsystemen invloed op elkaar kunnen hebben en dat daarom onderzoek naar SCADA logs op systeemniveau nuttig zijn.

Daarnaast is er ook onderzoek verricht op het voorspellen van falen van machines aan de hand van log mining. Wang et al. (2017) beschrijven een techniek om log mining te gebruiken om falen in pinautomaten te voorspellen. Binnen dit onderzoek wordt gebruikt gemaakt van verschillende classificatiemodellen zoals RandomForest, AdaBoost, en Support Vector Machines. Sipos et al. (2014) gebruiken multi-instance learning om voorspellingen te doen over het faalgedrag van medische apparatuur. Zij maken gebruik van sequentiële patronen en classificatie om voorspellingen te doen. Log mining wordt ook toegepast voor het analyseren van web logs (Pei, 2000). Dit onderzoek past sequential pattern mining toe op log gegevens om patronen in gebruikersgedrag te vinden.

Verdere methodieken van pattern mining die zijn gevonden zijn constraint based pattern mining (Bayardo et al., 1999) en multilevel mining (Han & Fu, 1999). Beide van deze methoden breiden association rule mining uit door het toepassen van aggregatieniveaus op Item databases of door het toevoegen van restricties aan association rule mining voor het verbeteren van gevonden verbanden.

4.3.2 Association rule mining

De meeste conditiemonitoring oplossingen richten zich op het functioneren van één bepaalde soort apparatuur binnen een systeem (Eickmeyer et al., 2015). Omdat de SCADA gegevens zich niet richten tot het falen van één bepaald apparaat of onderdeel van de bruggen is gekozen voor de oplossing van Czora et al (2017) dat functioneren op systeemniveau in kaart brengt. De SCADA loggegevens verzamelen alarmen en storingen van alle deelsystemen van de bruggen. Deze loggegevens kunnen aan de hand van association rule mining worden onderzocht op mogelijke relaties tussen de meldingen en het gedrag van de deelsystemen.

Frequent itemset mining is een data analyse methode die oorspronkelijk werd gebruikt voor market basket analyse. Het werd ingezet om sets van producten te vinden die frequent samen gekocht werden (Borgelt, 2012).

Frequent itemset mining is een proces dat begint met een set van transacties $T = \{t_1 \dots t_n\}$ waar iedere transactie een set is van items in een database $I = \{i_1 \dots i_m\}$, volgens de regel $t \subseteq I$. Een voorbeeld hiervan zijn winkelmanden met producten in een supermarkt. Een frequente itemset wordt gedefinieerd als een itemset welke in meer transacties voorkomt dan een vooraf gespecificeerde support:

$$supp(X) = \frac{|t \in T; X \subseteq t|}{|T|}$$

De support is het aantal keer dat een itemset X als subset voor komt in T. Wanneer de support hoger is dan gespecificeerd door de gebruiker wordt het gezien als frequent *(Leskovec, Rajaraman, & Ullman, 2014).* Rare itemset mining is een variant van itemset mining die zich richt op het vinden van zeldzame verbanden die een zeer lage support hebben (Koh & Rountree, 2005).

Association rule mining bestaat uit het vinden van verbanden in frequente itemsets. Het is één van de meest gebruikte vormen van data mining (Koh & Rountree, 2005). Deze verbanden, genaamd association rules, zijn regels in de vorm van $A \Rightarrow B$ waar A het antecedent is en B de consequent waartoe het antecedent leidt. De significantie van deze regels kan worden beschreven door de support en confidence statistieken. De support geeft het aantal transacties weer waarin zowel A als B samen voorkomen. Dit wordt uitgedrukt als de waarschijnlijkheid (probability) dat een transactie de unie van sets A en B bevat, of zowel A als B.

$$supp(A \Rightarrow B) = P(A \cup B)$$

De confidence geeft aan in hoeveel gevallen dat A voorkomt, B ook voorkomt (Czora et al., 2017). Ofwel, wat is de kans dat de consequent voorkomt wanneer de antecedent voorkomt? Uitgedrukt als een conditionele kansfunctie:

$$confidence(A \Rightarrow B) = P(B \mid A)$$

Dit valt tevens te noteren gebruikmakend van de support:

$$confidence(A \Rightarrow B) = \frac{supp(A \cup B)}{supp(A)}$$

Om de associatie te beoordelen en interessante verbanden te kunnen identificeren stellen Czora et al. (2017) de Kulczynski (Kulczynski, 1928) maat en de imbalance ratio (IR) voor. De Kulczynski maat

baseert zich op het gemiddelde van de twee conditionele kansfuncties voor de associatieregels $A \Rightarrow B$ en $B \Rightarrow A$. De Kulczynski maat is als volgt gedefinieerd:

$$Kulc(A,B) = \frac{1}{2}(P(A|B) + P(B|A))$$
$$= \frac{1}{2}(confidence(B \Rightarrow A) + confidence(A \Rightarrow B))$$

Een Kulczynski score van 0 of 1 geeft respectievelijk een sterke negatieve of positieve correlatie aan tussen de gevonden regels en maken de regel interessant. Een waarde van 0.5 duidt aan dat er geen verband is.

Czora et al. (2017) stellen dat bij een Kulczynski score van 0.5 de imbalance ratio kan worden toegepast om te kijken of een associatieregel mogelijk toch interessant is. De imbalance ratio wordt als volgt berekend:

$$IR(A,B) = \frac{|supp(A) - supp(B)|}{supp(A) + supp(B) + supp(A \cup B)}$$

De imbalance ratio geeft een waarde van 0 tot 1 van de onbalans tussen patronen A en B waar 0 volledig gebalanceerd is en 1 zeer scheve verhouding aanduidt.

Naast deze maten wordt binnen het association rule mining vaak gebruikt gemaakt van de Lift of Interest maat (Tan, Kumar,& Srivastava, 2002), (Merceron & Yacef, 2008). De lift is de ratio van de geobserveerde support van de association rule ten opzichte van wat verwacht wordt wanneer de antecedent en consequent onafhankelijk waren. De Lift is als volgt gedefinieerd:

$$Lift(A \Rightarrow B) = \frac{supp(A \cap B)}{supp(A) \times supp(B)}$$

4.3.3 Analyse

Er zijn verschillende manieren gevonden om frequente itemsets te vinden in databases van transacties. Omdat dit onderzoek voornamelijk geïnteresseerd is in het vinden van zeldzame patronen in het storingsgedrag van de bruggen, is ook gezocht naar manieren om zeldzame patronen, ofwel rare itemsets te minen. De algoritmes die zijn gevonden om itemsets te minen uit transactie databases zijn opgenomen in Tabel 5. Binnen dit onderzoek worden methoden als constraint based mining en multi level mining niet toegepast vanwege tijdsoverwegingen.

Tabel 5: Itemset mining algoritmes

Algoritme
Apriori
Apriori Inverse
FP-Growth
Eclat

Omdat de SCADA logs van de bruggen ook het functioneren van alle deelsystemen beslaan geeft het onderzoek van Czora et al. (2017) een richting om gebruikmakend van ARM verbanden te vinden in alarmen en storingen. Gebaseerd op de gevonden algoritmes en de werking hiervan is gekozen om

het vinden van de associaties te benaderen gebruikmakend van een Apriori Inverse algoritme. Dit algoritme geeft het meeste vrijheid voor het vinden van itemsets omdat het zowel een minimum als een maximum support voor het vinden van itemsets accepteert als parameters. Op deze manier kan verwacht gedrag van de brug, ofwel patronen die altijd bij elkaar voorkomen en dus niet als sporadisch of 'rare' worden gezien, worden geschrapt uit de gevonden itemsets. Dit bespaart geheugen tijdens het uitvoeren van het algoritme.

4.4 Deelvraag 4

Om deelvraag 4, "Hoe kunnen deze data mining technieken worden ingezet op de SCADA gegevens?", te kunnen beantwoorden is het algoritme gevonden in deelvraag drie toegepast op de SCADA gegevens van de eenentwintig bruggen. Het algoritme wordt als prototype opgeleverd en vormt de basis voor het experiment van de toepasbaarheid van predictive maintenance op SCADA log gegevens gebruikmakend van association rule mining. Eerst worden de resultaten van het realiseren van het model samengevat en worden de gebruikte parameters en samples verantwoord. Daarop volgt een analyse van de resultaten uit het model.

4.4.1 Resultaten

Om de rare patterns te minen worden vanuit een database met SCADA events van alle bruggen eerst een transactie database opgebouwd. In pseudocode ziet dit proces er als volgt uit:

```
INPUT: DB, brug_id (int), alarms_only (bool), time_window (int)
OUTPUT: T

T = {}
for alarm in DB.get_alarms(brug_id) do
....if alarm.tijd < tijd_previous then continue
....if alarms_only then
.... ...t = DB.get_alarms where tijd between(alarm.tijd –
time_window, alarm.tijd)



....else
.... ...t = DB.get_events(brug_id) where tijd <= between(alarm.tijd,
alarm.tijd + time_window)
....endif
....tijd_previous = alarm.tijd + time_window
....T.append(t)
end
return T
```

Iedere transactie wordt opgebouwd door over de alarmen van een brug met een door de gebruiker gespecificeerd ID te itereren. Het algoritme voor het opbouwen van de transactiedatabase neemt een parameter alarms_only waarin gespecificeerd kan worden of enkel alarmen worden meegenomen in de transactie of dat alle meldingen (procesmeldingen) worden meegenomen. Voor ieder alarm wordt vervolgens een tijdsvenster gemaakt van de tijd dat het alarm opdoet tot een gebruiker gespecificeerde time_window uitgedrukt in uren vóórdat het alarm opdoet. Om een

overlapping in de transacties te voorkomen wordt gecontroleerd of een alarm al voorkomt in een transactie door te kijken of het alarm binnen het tijdsvenster van de vorige transactie valt. Op deze manier wordt ruis voorkomen die veroorzaakt wordt door overlappende transacties. Alle alarmen die binnen het tijdsvenster voorkomen worden beschouwd als dezelfde storingsaanloop. De keuze voor een tijdsinterval is genomen na overleg met een expert data scientist van de Hogeschool Zeeland.

Deze transactiedatabase wordt vervolgens gebruikt door het Apriori Inverse algoritme zoals beschreven in Bijlage D: Overzicht pattern mining algoritmes. Van het Apriori Inverse algoritme is een implementatie gevonden in Python ("AprioriMLxtend-Extension," 2018) die zich baseert op het Apriori Algoritme van de library MLExtend, een python library voor data mining algoritmes ("*mlxtend," 2018)*.

Vanuit de gevonden itemsets zijn vervolgens association rules gegenereerd gebruikmakend van de MLExtend library. In de analyse is beschreven welke parameters gebruikt worden en waarom deze zijn gekozen. Voor het beoordelen van de gevonden association rules zijn ook de imbalance ratio en de Kulczynski maat toegevoegd aan het algoritme. Zo wordt de regel beoordeeld volgens de volgende meetinstrumenten:

```
// Zekerheid dat A voorkomt ten opzichte van intersectie A ∩ B
Confidence (A -> B) = support(A,B) / support(A)

// Verhouding van onafhankelijkheid tussen A en B
Lift (A -> B) = confidence(A,B) / support(B)

// Samenhang tussen regels A->B en B->A (a->b impliceert b->a en
vise versa.)
Kulczynski (A,B) = 0.5*(confidence(B,A) + confidence(A,B))

// Graad van onbalans tussen itemsets A en B
Imbalance Ratio (A,B) = abs(support(A)-support(B)) / (support(A) +
support(B) + support(A,B))
```

4.4.2 Analyse

Voor het uitvoeren van het algoritme zijn eerst alle parameters vastgesteld waar de itemsets aan moeten voldoen. Daarna is voor iedere brug het algoritme uitgevoerd voor alle gevonden storingen van het object. Op deze manier wordt de invloed van alle deelsystemen die events loggen naar het SCADA systeem van het object op het storingsgedrag in kaart gebracht. Storingen worden gezien als alarm meldingen met prioriteit 'WAARSCHUWING' en 'URGENT'. Er is gekozen om de procesmeldingen buiten beschouwing te laten voor deze stap in het onderzoek omdat deze meldingen dusdanig grote sets $\gg 50000$ met association rules genereren, met allen zeer hoge confidence en lift (omdat het gaat om verwacht gedrag van de brug) dat deze geen zinvolle regels genereren.

Om patronen te kunnen herkennen is een absolute minimale support nodig voor itemsets. Deze minimale support is na overleg met een expert data science aan de Hogeschool Zeeland vastgesteld op vijf, $minsup = 5$. Iedere itemset moet minstens vijf keer voorkomen om als significant frequent gezien te worden. Een lagere minsup zou associaties creëren met hoge confidence, maar door de

lage support zijn deze regels nietszeggend omdat de toevalligheid te groot wordt. Voor het vinden van association rules is besloten voor een $minconf = 0.6$. Dit resulteert in association rules die een minimale confidence hebben van zestig procent.

De maximale lengte van itemsets die worden gegenereerd is ingesteld op twee om de kandidaatset generatie te verkleinen en omdat één op één verbanden tussen items als interessant gezien worden. Doordat een lage support nodig is voor het vinden van zeldzame patronen benadert het algoritme de theoretische maximale tijds- en geheugencomplexiteit $O(2^{|I|})$, waar I de set is van items in de database (Tan et al., 2005). Dit wordt voorkomen door de lengte van de itemsets te beperken.

4.3.2.1 Tijdsvenster t=1 uur

Voor het experiment is gekozen voor een tijdsvenster van één uur. Per brug is een transactie database opgesteld met transacties die urgente- en waarschuwingsmeldingen bevatten van een uur voordat een storing (tevens een urgente alarmmelding) wordt gelogd. De parameters voor het experiment zijn $minsup = 5, minconf = 0.6, time\_window = 1$.

In Tabel 6 is opgenomen hoeveel regels de bruggen genereren en daarbij het aantal transacties dat is gevonden in de database. Dit geeft een beeld van de verhoudingen in storingsaanloop tot het aantal patronen dat herkend kan worden.

Tabel 6: Gevonden transacties en regels per brug

ID	Brug Naam	Aantal Transacties	Gevonden Regels
01	Amertakbrug	511	2
02	Dr. Deelenlaan	49	0
03	Waalstraat	38	0
04	Lijnsheike	87	14
05	Heikantsebaan	43	37
06	Enschotsestraat	48	55
07	Bosscheweg	212	91
08	Oisterwijksebaan	48	25
09	Trappistenbrug	27	0
10	Biesthoutakker	43	0
11	Holenakker	37	0
12	Stad van Gerwen	37	0
13	Heuvel	119	0
14	Houtens	46	2
15	Sonse brug	31	0
16	Hooijdonk	21	0
17	Groenewoud	41	0
18	Brug over Sluis V	76	65
19	Oranjelaan	45	0
20	Biesterbrug	76	30
21	Stadsbrug Weert	99	7

In Tabel 7 is een observatie te zien van een aantal association rules gegenereerd voor verschillende bruggen. De gevonden regels illustreren de toepasbaarheid van association rule mining voor

predictive maintenance op de huidige informatie uit de SCADA systemen. De association rules gegenereerd voor brug 14, Houtens, geeft een sterk verband aan tussen een waarschuwing en een consequentie die beiden binnen een uur van elkaar gebeuren in meer gevallen dan de minsup parameter. Het opdoen van de “Camera 3 Laag Detail” melding leidt in honderd procent van de gevallen tot de melding “Camera 3 Urgente Storing”. De Kulczynski maat geeft tevens een hoge correlatie aan tussen de twee meldingen.

Wanneer wordt gekeken naar andere bruggen geven de gegenereerde association rules enkel verbanden aan die onlosmakelijk gedrag van de brug weergeven. Ter illustratie, de regels van Brug 21, Stadsbrug Weert (Tabel 7), geven aan dat een verzamelstoring van de brug altijd een directe stop van het object als resultaat hebben. De Kulczynski maat geeft ook een correlatie van 1 en de associatie is volledig gebalanceerd.

Tabel 7: Samenvatting association rules verschillende bruggen

RID	antecedents	consequents	confidence	lift	Kulczynski	IR
Brug 14, Houtens						
0	Camera 3 Laag Detail	Camera 3 Urgente Storing	1.000000	3.285714	0.928571	0.052632
1	Camera 3 Urgente Storing	Camera 3 Laag Detail	0.857143	3.285714	0.928571	0.052632
Brug 21, Stadsbrug Weert						
0	Brug Directe Stop	Brug Verzamelstoring	1.000000	16.500000	1.000000	0.000000
1	Brug Verzamelstoring	Brug Directe Stop	1.000000	16.500000	1.000000	0.000000
2	Brug Discrepantie Fout Verzameling	Brug Einstanden Dicht Discrepantie Fout	1.000000	19.800000	1.000000	0.000000
3	Brug Einstanden Dicht Discrepantie Fout	Brug Discrepantie Fout Verzameling	1.000000	19.800000	1.000000	0.000000
4	Luidspreker LS-02 Link Down	Luidspreker LS-04 Link Down	0.615385	2.175824	0.593407	0.028571
5	Luidspreker LS-02 Link Down (NCT)	Luidspreker LS-04 Link Down	0.600000	2.121429	0.407143	0.409091
Brug 7, Bosscheweg						
27	Afsluitboom 1 Storing	Afsluitboom 4 Storing	1.00	26.500000	1.000	0.0
28	Brug Noodstop	Afsluitboom 1 Storing	1.00	26.500000	0.875	0.1
29	Afsluitboom 1 Storing	Brug Noodstop	0.75	26.500000	0.875	0.1

30	Brug Verzamelstoring	Afsluitboom 1 Storing	1.00	26.500000	0.875	0.1
31	Afsluitboom 1 Storing	Brug Verzamelstoring	0.75	26.500000	0.875	0.1
32	Noodstop Actief	Afsluitboom 1 Storing	1.00	26.500000	0.875	0.1
33	Afsluitboom 1 Storing	Noodstop Actief	0.75	26.500000	0.875	0.1
34	Afsluitboom 2 Storing	Afsluitboom 2 Noodstop	0.75	26.500000	0.875	0.1
35	Afsluitboom 2 Noodstop	Afsluitboom 2 Storing	1.00	26.500000	0.875	0.1

4.3.2.2 Tijdsvenster t>1 uur

In gevallen waar het tijdsvenster groter wordt gemaakt dan één uur produceert het algoritme geen regels, te wijten aan het afnemend aantal transacties. Er zijn te weinig alarmen gelogd om een transactie database op te bouwen van significante grootte (waar minsup > 5).

4.5 Totaalbeeld

Op basis van een verkenning van de doelstellingen van het vitale assets project, specifiek het Tilburg 3 contract, is een lijst met data mining requirements opgesteld. Deze requirements zijn geprioriteerd en de belangrijkste is gekozen om uit te werken en het onderzoek op te vervolgen. Daarnaast is de kwaliteit van de beschikbare SCADA gegevens in kaart gebracht door deze te toetsen aan datakwaliteitsdimensies. Deze stap van het onderzoek is uitgevoerd als data analyse van de SCADA log gegevens van alle bruggen vanaf februari tot en met september. Hieruit blijkt dat een groot aantal bruggen gegevens mist in de maanden juli en augustus en de compleetheid van de gegevens dus niet volledig is.

Vervolgens is naar de toepassingen van patroonherkenning op SCADA gegevens, gebaseerd op de data mining requirement (DMR 2) welke in deelvraag één als hoogst geprioriteerd is, gezocht. In het literatuuronderzoek is de toepassing van frequent pattern mining op SCADA gegevens onderzocht. Daaruit blijkt dat association rule mining kan worden ingezet op SCADA logs. Deze methode is vervolgens verder uitgewerkt en verschillende algoritmes voor association rule mining zijn onderzocht. Op basis van het Apriori Inverse algoritme is vervolgens een methode opgesteld om itemsets te minen uit de SCADA data door alarm meldingen te groeperen in tijdsvensters voorafgaande aan een alarm. Deze tijdsvensters worden als transactie gezien. De transacties worden gevoed aan het Apriori Inverse algoritme welke itemsets opstelt die vaker voorkomen dan minimale support 5. Hieruit zijn association rules opgesteld volgens het Apriori algoritme welke vervolgens zijn beoordeeld aan de meetwaarden zoals beschreven in paragraaf 4.3.

5. Discussie

Op basis van het uitgevoerde onderzoek en de verkregen resultaten worden in dit hoofdstuk de deelvragen beantwoord. Daarnaast worden de resultaten van het onderzoek geïnterpreteerd en wordt op basis van de interpretatie een conclusie gevormd. Per deelvraag wordt een conclusie gevormd die leidt tot een algehele conclusie. Op basis van deze conclusie worden suggesties opgesteld voor vervolgonderzoek en worden aanbevelingen gedaan met betrekking tot het vervolg van het Vitale Assets project binnen het Tilburg 3 contract.

5.1 Deelvraag 1

5.1.1 Discussie

Het in kaart brengen van de doelstelling van het project heeft inzichtelijk gemaakt welke richting het onderzoek kon nemen. Dit is gedaan doormiddel van overleg met experts binnen Rijkswaterstaat betrokken bij het Tilburg 3 contract. Hiervoor zijn twee verschillende stakeholders geïnterviewd die beiden een eenduidig beeld van de doelstelling schetsten. Op basis hiervan kan gesteld worden dat de gevonden doelstelling valide is. Bij herhaling van het onderzoek zouden de doelstellingen daarom hetzelfde zijn. Binnen deze stap is de keuze gemaakt om één requirement uit te diepen in plaats van alle requirements te verkennen. Dit is besloten vanuit een tijdsoverweging en om beter te kunnen verdiepen in het vakgebied om zo een beter onderbouwde conclusie op te kunnen stellen.

5.1.2 Deelconclusie

Er is een beeld geschetst van de oplossingsrichtingen die het project kon aannemen. Dit is gedaan door het vastleggen van de doelstelling die Rijkswaterstaat heeft aan het toepassen van predictive maintenance. Deze doelstelling is, vanuit onderhoudsmanagement oogpunt, het op een slimmere manier inrichten van onderhoud. Daarbinnen wordt de eis aan gesteld om verbanden te vinden binnen de SCADA meldingen om zo het storingsgedrag in kaart te brengen.

5.2 Deelvraag 2

5.2.1 Discussie

Voor de tweede deelvraag is de datakwaliteit van de eenentwintig bruggen in kaart gebracht aan de hand van kwaliteitsdimensies onderverdeeld in scenario's. Omdat deze gegevens statisch en de scenario's eenduidig zijn, is dit proces herhaalbaar en zal het altijd hetzelfde resultaat opleveren. De vraag over de consistentie van de gegevens zoals beschreven in paragraaf 4.2.2 blijft onbeantwoord doordat er geen validatie plaats heeft kunnen vinden met de verantwoordelijke voor het ontsluiten van de SCADA gegevens. Het is hierom dat aannames moeten worden gemaakt over deze uitkomsten.

5.2.2 Deelconclusie

Er wordt van uitgegaan dat de gaten in de logging veroorzaakt zijn door inconsistente logging. Dit vormt een substantieel gat in de data logging van 12.5% op een dataset van acht maanden. Daarnaast is een gat in de compleetheid van de gegevens ontdekt met in het geval van de Sonse brug een kwaliteitsverlies in compleetheid van 10.93%. Op een dataset van acht maanden is dit een significante hoeveelheid missende gegevens. Het voorspellen of zoeken naar verbanden in storingen om te sturen op onderhoud wordt bemoeilijkt door de hoeveelheid missende gegevens.

5.3 Deelvraag 3

5.3.1 Discussie

Om deelvraag drie te beantwoorden is gezocht naar literatuur over het toepassen van data mining methodieken op SCADA systemen. Dit is gedaan met als doel het verkrijgen van inzicht in de conditie en functioneren van civieltechnische bouwwerken en de toepassing hiervan binnen predictive maintenance. Gegeven de zoektermen die ontleend zijn uit de doelstelling van Rijkswaterstaat en de sleutelwoorden uit de methode vormt dit een herhaalbare stap in het onderzoek. De zoektermen zijn beperkt gebleven tot het vinden van patronen in SCADA gegevens. Vanuit tijdsoverweging is besloten geen andere methodes te onderzoeken die aansloten bij andere gevonden requirements. De gekozen richting is overlegd met begeleidend docent en de informatiemanager Zee en Delta bij Rijkswaterstaat welke de onderzoeksrichting goed bevonden.

5.3.2 Deelconclusie

De gevonden methode die ingezet is binnen dit onderzoek is association rule mining. Dit wordt gebruikt om inzicht te krijgen in het functioneren op systeemniveau. Op deze wijze kunnen patronen in alarmmeldingen en storingen worden gevonden. Daarnaast is gezocht naar een methode om associatieregels te toetsen op validiteit. Hiervoor zijn meetinstrumenten onderzocht welke samen met implementaties van itemset mining algoritmes een framework vormden voor het in kaart brengen van patronen in storingsgedrag.

5.4 Deelvraag 4

5.4.1 Discussie

Om deelvraag vier te beantwoorden is een implementatie van het Apriori Inverse algoritme toegepast op de SCADA gegevens. De SCADA gegevens zijn onderverdeeld in transacties met tijdsvensters van één uur. Alle alarmmeldingen die binnen het uur voordat de storing optrad vallen binnen deze transactie. Deze implementatie leidt tot de limitatie dat er vanuit gegaan wordt dat er waarschuwingsmeldingen of andere alarmen plaatsvinden binnen het uur voordat de storing optreedt. Daarom geeft het geen volledig beeld van een storingsaanloop waarvan de waarschuwing eerder kwam dan dat uur. Om dit te toetsen zijn verschillende tests gedaan met een groter tijdsvenster die tegen het probleem aanliepen dat er dan té weinig transacties zijn om te voldoen aan de minimale support nodig om de itemset te kwalificeren voor het association rule mining algoritme. Dit kan met zekerheid worden toegeschreven aan het gebrek aan data op langere termijn. Het aantal storingen verdeeld over verschillende deelinstallaties dat op een periode van acht maanden wordt gelogd lijkt te weinig om nauwkeurige patronen te ontdekken.

5.4.2 Deelconclusie

Het is niet mogelijk om met de gevonden implementatie interessante patronen te ontdekken die inzicht geven in het functioneren van het object en het storingsgedrag hiervan, gegeven de beperkte hoeveelheid storingen die per brug worden waargenomen. De gevonden association rules geven enkel verwacht gedrag van de objecten weer.

5.5 Conclusie

De hoofdvraag van het onderzoek is als volgt gedefinieerd:

"In hoeverre is het mogelijk om met de beschikbare SCADA gegevens predictive maintenance toe te passen?"

De conclusie van het onderzoek is, gegeven de huidige dataset van eenentwintig bruggen in Zuid Nederland, periode februari tot en met september, dat het niet mogelijk is om verbanden te vinden in het storingsgedrag van de bruggen. De verkende onderzoeksrichting, association rule mining in log gegevens, biedt geen bruikbare uitkomsten voor een predictive maintenance strategie. De uitkomst van het onderzoek geeft geen uitsluitsel dat met grotere, meerjarige datasets, ook geen bruikbare resultaten worden gevonden.

5.5.1 Vergelijking met andere theorie

Czora et al. (2017) produceerden onderzoek dat aantoonde dat rule mining succesvol kan worden toegepast op SCADA systemen voor petrochemische installaties. Zij gebruikten in plaats van tijdsvensters een vast aantal events voordat een alarm werd geproduceerd. Ook werd hun onderzoek verricht op een conditie monitoring systeem. Verma (2012) maakte gebruik van association rule mining om patronen te vinden in het functioneren van windturbines die naast alarmen ook periodieke logging genereren over hun staat van functioneren. De eenentwintig bruggen doen dit niet.

5.5.2 Aanbevelingen

Vanuit de conclusie wordt aanbevolen dat meerjarige data wordt vergaard om zo het zoekgebied te vergroten voor de toepassing van association rule mining op SCADA logs. Op deze manier kan uitsluitsel gegeven worden over de toepassing van association rule mining om inzicht te krijgen in het storingsgedrag. Wanneer er succesvol association rules worden gevonden bij vergroten van de dataset wordt aanbevolen om multi level mining (Han & Fu, 1999) toe te passen om associaties te kunnen aggregeren, bijvoorbeeld op deelinstallatieniveau. Daarnaast wordt aanbevolen de verbositeit van de logs te vergroten. Dit kan worden gedaan door alarm thresholds te verlagen en eerder waarschuwingen te genereren vanuit de PLC systemen van de bruggen. Het toevoegen van conditie monitoring sensoren binnen deelinstallaties van bruggen, welke communiceren met de SCADA systemen, kan het inzicht ook vergroten in het functioneren van het object, zoals de onderzoeken van Verma (2012) en Czora et al. (2017) aantonen. Het periodiek loggen van conditie monitoring informatie kan mogelijk ook bijstaan aan de toepasbaarheid van association rule mining op de SCADA systemen.

Mogelijkheden voor vervolgonderzoek naar de mogelijkheden van predictive maintenance zijn het zoeken naar verbanden tussen bedienhandelingen en storingsgedrag zoals data mining requirement (DMR 3). Er wordt vanuit dat oogpunt aanbevolen om te kijken naar sequentiële patronen (Pei et al., 2000) in de SCADA proces- en alarmgegevens. Daarnaast wordt aanbevolen om onderzoek te doen naar het combineren van SCADA log gegevens en andere databronnen zoals weer en energieverbruik om de invloed van deze factoren op het functioneren van de objecten in kaart te brengen.

5. Literatuur

AprioriMLxtend-Extension [Github Repository]. (2018, 26 juli). Geraadpleegd op 6 november 2018, van https://github.com/Al2a2d2m/AprioriMLxtend-Extension

Askham, N., Cook, D., Doyle, M., Fereday, H., Gibson, M., Landbeck, U., ... & Schwarzenbach, J. (2013). The six primary dimensions for data quality assessment. DAMA UK Working Group, 432-435.

Bayardo, R. J., Agrawal, R., & Gunopulos, D. (1999, March). Constraint-based rule mining in large, dense databases. In Data Engineering, 1999. Proceedings., 15th International Conference on (pp. 188-197). IEEE.

Borgelt, C. (2012). Frequent item set mining. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2(6), 437-456.

Carnero, M. (2006). An evaluation system of the setting up of predictive maintenance programmes. Reliability Engineering & System Safety, 91(8), 945-963.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. & Wirth, R. (2000). CRISP-DM 1.0 Step-by-step data mining guide (). The CRISP-DM consortium.

Czora, S., Dix, M., Fromm, H., Klöpper, B., & Schmitz, B. (2017). Mining Industrial Logs for System Level Insights. Datenbanksysteme für Business, Technologie und Web (BTW 2017)-Workshopband.

Eickmeyer, J., Li, P., Givehchi, O., Pethig, F., & Niggemann, O. (2015). Data Driven Modeling for System-Level Condition Monitoring on Wind Power Plants. In DX@ Safeprocess (pp. 43-50).

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996a). From data mining to knowledge discovery in databases. AI magazine, 17(3), 37.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996b). The KDD process for extracting useful knowledge from volumes of data. Communications of the ACM, 39(11), 27-34.

Gartner. (2013, 13 mei). Analytics Framework [Illustratie]. Geraadpleegd van https://zdnet1.cbsistatic.com/hub/i/r/2016/11/23/2610d41b-e3d0-475d-8fe6-a784c7bc78c6/resize/470xauto/4c9802081ab69615b06a4358ccbe77f2/analytic-maturity.jpg

Gaushell, D. J., & Darlington, H. T. (1987). Supervisory control and data acquisition. Proceedings of the IEEE, 75(12), 1645-1658.

Han, J., & Fu, Y. (1999). Mining multiple-level association rules in large databases. IEEE Transactions on Knowledge & Data Engineering, (5), 798-805.

Han, J., Kamber, M. (2006). data mining: Concepts and Techniques. University of Illinois at Urbana-Champaign. (Foto)

Hashemian, H. M., & Bean, W. C. (2011). State-of-the-art predictive maintenance techniques. IEEE Transactions on Instrumentation and measurement, 60(10), 3480-3492.

Khan, D. M., Mohamudally, N., & Babajee, D. K. R. (2013). A unified theoretical framework for data mining. Procedia Computer Science, 17, 104-113.

Koh, Y. S., & Rountree, N. (2005, May). Finding sporadic rules using apriori-inverse. In Pacific-Asia Conference on Knowledge Discovery and Data Mining (pp. 97-106). Springer, Berlin, Heidelberg.

Kulczynski S. (1928) Die Pflanzenassoziationen der Pieninen. Bull. Int. Acad. Pol. Sci. Lett. Cl. Sci. Math. Nat. Ser. B, (Suppl. II) 1927, 57–203.

Kurgan, L. A., & Musilek, P. (2006). A survey of knowledge discovery and data mining process models. The Knowledge Engineering Review, 21(1), 1-24.

Leskovec, J., Rajaraman, A., & Ullman, J. D. (2014). Mining of massive datasets. Cambridge university press.

mlxtend [Github Repository]. (2018, 23 november). Geraadpleegd op 23 november 2018, van https://github.com/rasbt/mlxtend

Pei, J., Han, J., Mortazavi-Asl, B., & Zhu, H. (2000, April). Mining access patterns efficiently from web logs. In Pacific-Asia Conference on Knowledge Discovery and Data Mining (pp. 396-407). Springer, Berlin, Heidelberg.

Rijkswaterstaat. (2017, 3 mei). I-Visie 2025. Geraadpleegd op 10 oktober 2018, van http://corporate.intranet.rws.nl/Content/Media/92ad2b01-a14c-4c01-8159-681dfaf85fb9/i-Visie%20RWS%202025.pdf

Sipos, R., Fradkin, D., Moerchen, F., & Wang, Z. (2014, August). Log-based predictive maintenance. In Proceedings of the 20th ACM SIGKDD international conference on knowledge discovery and data mining (pp. 1867-1876). ACM.

Stouffer, K., & Falco, J. (2006). Guide to supervisory control and data acquisition (SCADA) and industrial control systems security. National institute of standards and technology.

Tan, P. N., Steinbach, M., & Kumar, V. (2005). Association analysis: basic concepts and algorithms. Introduction to Data mining, 327-414.

Verma, A. P. (2012). Performance monitoring of wind turbines: a data-mining approach.

Wang, J., Li, C., Han, S., Sarkar, S., & Zhou, X. (2017). Predictive maintenance based on event-log analysis: A case study. IBM Journal of Research and Development, 61(1), 11-121.

Wang, K. (2016). Intelligent predictive maintenance (IPdM) system–Industry 4.0 scenario. WIT Transactions on Engineering Sciences, 113, 259-268.

Waters, K. (2009). Prioritization using moscow. Agile Planning, 12, 31.

What is Predictive Maintenance? - Definition from Techopedia. (z.d.). Geraadpleegd op 8 november, 2018, van https://www.techopedia.com/definition/32027/predictive-maintenance

Yao, Y. Y. (2001). On modeling data mining with granular computing. In Computer Software and Applications Conference, 2001. COMPSAC 2001. 25th Annual International (pp. 638-643). IEEE.

Bijlage A: Requirementsdocument

Requirements Document

Business Understanding fase Project 21 bruggen

Roel van Endhoven

In opdracht van Rijkswaterstaat regio Zee en Delta

Versie 1.0

Inhoud

1. Inleiding

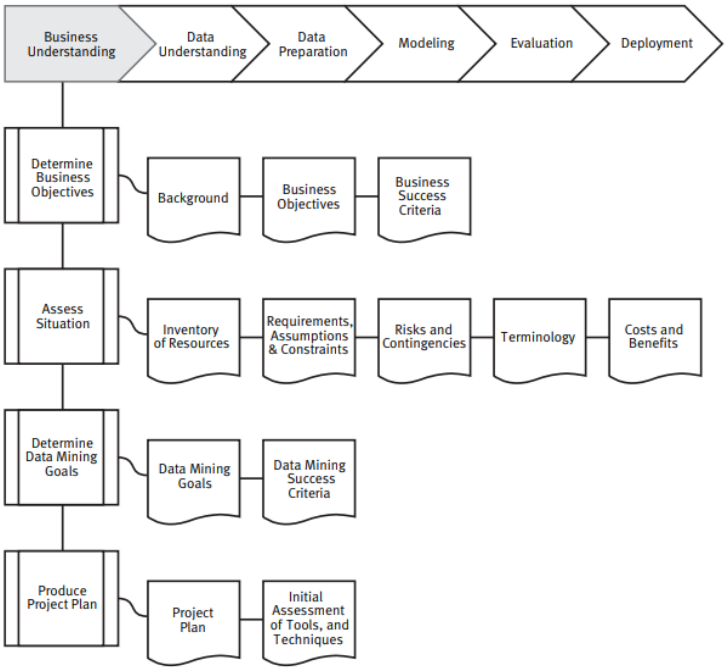
Dit document is opgesteld om de gebruikseisen (requirements) van het project 21 bruggen in kaart te brengen. Het requirementsdocument legt verslag van overleg met stakeholders welk is gepleegd om de informatiewens en doelstellingen van Rijkswaterstaat te inventariseren. De rol van de SCADA gegevens staat hierbij centraal. De Business Understanding fase van het CRISP-DM proces schrijft een aantal stappen voor om tot een inventarisatie van doelstellingen en eisen te komen. Deze stappen vormen de methode voor het ontsluiten van de requirements aan het project.

Het document bevat de deelproducten van de Business Understanding fase van de CRISP-DM cyclus. In paragraaf 1.1 wordt omschreven welke producten er worden meegenomen in het proces en hoe deze binnen het requirementsdocument zijn opgenomen. Binnen het CRISP-DM proces worden ook Software requirements opgesteld vanuit de uitkomsten van de Business Understanding fase. Deze requirements worden opgesteld om naast het data mining proces ook een dashboard te ontwikkelen waarin bevindingen gepresenteerd en gevisualiseerd kunnen worden.

1.1 Business Understanding

In figuur 1 wordt schematisch weergeven hoe de Business Understanding fase van de CRISP-DM methode eruit ziet en welke deelproducten hierbij horen. Binnen dit requirementsdocument worden de business goals vastgelegd aan de hand van overleg met betrokken stakeholders aan het Vitale Assets project en de SCADA gegevens van het project 21 bruggen. De output hiervan zijn verslagen van het overleg met Robert van Lakwijk en Arnoud de Kruijf. De uitkomsten van dit overleg zijn omgevormd tot functionele requirements aan het data mining proces.

Het tweede deel van de Business Understanding bestaat uit het prioriteren volgens het MOSCOW systeem. Moscow staat voor Must, Should, Could en Won’t. Aan de hand van deze schaal kunnen kritische requirements worden opgesteld en kan de scope van het project ingekaderd worden. Daarbij wordt de focus gelegd op requirements welke zich lenen aan datamining technieken en zich richten op het verkrijgen van inzicht over het functioneren van de bruggen en de mogelijke toepassing van predictive maintenance.



Figuur 5: Overzicht Business Understanding

2. Verslag overleg Robert van Lakwijk

In dit hoofdstuk wordt verslag gedaan van het overleg met Robert van Lakwijk, gehouden op 12 september 2018. Dit overleg is gehouden om de informatiebehoefte en de doelstellingen aan het predictive maintenance project Vitale Assets in kaart te brengen.

2.1 Huidige situatie

Het Tilburg 3 contract is een iteratie op het contract dat betrekking heeft op bruggen in Zuid Nederland. Het project Tilburg loopt sinds 2011 met als huidig contract Tilburg 2.5. De bruggen worden bediend vanuit de Nautische Centrale Tilburg (NCT).

Op dit moment worden storingen handmatig ingevoerd in het meldsysteem Ultimo. De meldregel komt binnen op de NCT en wordt ingevoerd in het systeem zodat de aannemer hierop kan reageren, en indien nodig, onderhoud plegen. Dit is een vorm van reactief onderhoud. Het doel van het project is, vanuit onderhoudsmanagement oogpunt, hoe onderhoud op een slimmere manier ingericht kan worden.

Het beheersysteem van de bruggen is ontwikkeld door Istimewa Elektro. Dit systeem baseert zicht op een Supervisory control en data acquisition (SCADA) systeem en vormt het Bediening op Afstand (BoPa) systeem. De SCADA gegevens zijn afkomstig van de objecten (bruggen) zelf.

2.2 Doelstellingen

De belangrijkste punten die uit het overleg naar voren kwamen, met betrekking tot het onderzoeksproject en de informatiewens, zijn de volgende:

5. Creëer inzicht in wat er gebeurt met een object aan de hand van de SCADA gegevens.
6. Achterhaal waarom dingen gebeuren.
    a. Leg vast hoe analyseerbaar de SCADA gegevens zijn.
    b. Leg vast welke eisen er gesteld moeten worden om dit inzicht te verkrijgen.

    c. Leg vast welke andere databronnen gebruikt kunnen worden om de analyseerbaarheid te vergroten.
7. Kunnen de storingsmeldingen specifieker worden gemaakt door events uit de SCADA gegevens te analyseren?
    a. Is er mogelijkheid tot het automatiseren van storingsmeldingen op basis van de SCADA gegevens?

De vragen en richtlijnen die uit het overleg zijn gekomen vormen de basis van waar de requirements op worden toegespitst om zo aan de data mining doelstellingen te kunnen voldoen.

3. Verslag overleg Arnoud de Kruijf

Dit hoofdstuk legt verslag van overleg met Arnoud de Kruijf, gehouden op 10 oktober 2018. Bij dit overleg is gekeken welke verdere doelstelingen er zijn binnen het 21 bruggen project en hoe deze ingepast kunnen worden in een dashboard.

3.1 Doelstellingen

Uit het overleg kwamen de volgende punten naar voren met betrekking tot de doelstellingen aan het data mining proces en de informatiebehoefte van Rijkswaterstaat:

1. Breng het gedrag van de brug in kaart.
    a. Hoe gedraagt de brug zich in totaliteit.
        i. Is het storingsgedrag verklaarbaar?
        ii. Zitten er patronen in?
    b. Breng doorlooptijden in kaart.
        i. Storingen
        ii. Brughandelingen
2. Kan worden herleid wat de oorzaak (bron) is van een storing?
    a. Zo ja, kan worden herleid wat de oplossing is?
    b. Wat is de graad van complexiteit van een storing?
    c. Kan de aanloop naar een storing in kaart worden gebracht?
3. Kunnen oneigenlijke bedienhandelingen worden vastgesteld?
    a. Kunnen deze in verband gebracht worden met storingen?

Met betrekking tot het **dashboard** zijn de volgende wensen naar voren gekomen:

1. Dashboard voert analyses uit op de achtergrond.
2. Dashboard toont correlaties tussen storingen en andere factoren.
3. Invloed leeftijd object op storingen.
4. Clusteren op overeenkomstige areaaldelen.
5. Van alle objecten storingen meten (bijvoorbeeld alle storingen op een aflsuitboom).
6. Dashboard toont abnormaliteit storingen (duur).
7. Dashboard toont frequenties en duur van handelingen (Brugopeningen).
8. Dashboard moet doorsneden kunnen maken op areaalniveau en type object.
9. Dashboard moet doorsneden kunnen maken op tijdsperiode.

4. User Stories

Aan de hand van het gehouden overleg zijn User Stories opgesteld. Deze User Stories geven een kader voor de requirements die worden opgesteld voor het project. De doelstellingen die zijn gevonden in het overleg worden uitgedrukt in User Stories met het formaat:

“Als <rol> wil ik <actie/kenmerk> om <reden>”.

(“Als data-analist wil ik het aantal storingen kunnen zien van brug x op dag x om inzicht te krijgen in het functioneren op dag niveau“).

Op deze manier wordt de doelstelling achter het data mining proces op middels een standaard methode in kaart gebracht.

4.1 Opgestelde User Stories

De User Stories zijn vastgelegd in tabel 1 met daarbij de User Story en om de traceerbaarheid te waarborgen staat vermeldt op welke doelstelling, gevonden in de overleggen met Robert van Lakwijk en Arnoud de Kruijf, deze gebaseerd is. Daarbij zijn ook de aanvullende eisen over het maken van doorsneden gekoppeld aan User Stories zodat deze uigewerkt kunnen worden naar requirements.

Tabel 8: User Stories

#	User Story	Tracability
U01	Als gebruiker wil ik openings- en sluitingstijden van een object kunnen zien om inzicht te krijgen in doorlooptijden van processen.	Arnoud de Kruijf, Dashboard punt 7, 8, 9 Robert van Lakwijk punt 1
U02	Als gebruiker wil ik de storingen van een object kunnen zien om inzicht te krijgen in het functioneren van de brug.	Arnoud de Kruijf, Dashboard punt 5, 8, 9 Robert van Lakwijk, punt 1
U03	Als gebruiker wil ik abnormaliteiten in storingen kunnen inzien om beter te kunnen sturen op onderhoud.	Arnoud de Kruijf, Dashboard punt 6, 8
U04	Als gebruiker wil ik aanvullende informatie zien over een brug om zo verbanden te zien tussen kenmerken van de brug en het gedrag.	Arnoud de Kruijf, Dashboard punt 3,8
U05	Als gebruiker wil ik statistische gegevens m.b.t. het gedrag van een object zien om inzicht te krijgen in het functioneren van een object.	Arnoud de Kruijf, Doelstellingen punt 1. Dashboard punt 2, 4, 8, 9
U06	Als gebruiker wil ik kunnen zien welke andere factoren invloed hebben op het storingsgedrag van een object om beter te kunnen sturen op onderhoud.	Arnoud de Kruijf, Dashboard punt 2

5. Requirements

Uit het opgestelde User stories zijn requirements opgesteld voor zowel het data mining proces, als het dashboard dat wordt ontwikkeld om de SCADA gegevens inzichtelijk te maken. Deze requirements zijn opgenomen in twee requirements matrices welke per requirement de prioriteit weergeeft en deze koppelt aan de opgestelde User stories zodat traceerbaarheid wordt gewaarborgd.

Requirements worden opgedeeld in categorieën Functioneel (FR) en Niet functioneel (NFR, volgens de ISO-25010 norm). Daarnaast wordt vanuit CRISP-DM oogpunt ook een voor het onderzoek specifieke Data-mining (DM) categorie onderscheiden. DM requirements zijn een subset van functionele requirements die apart worden benoemd om een overzicht te creëren van specifieke data mining eisen.

5.1 Functionele Requirements

In tabel 2 zijn de functionele requirements aan het storingsdashboard weergeven. De requirements zijn geprioriteerd aan de hand van MOSCOW. De functionele requirments zijn onderverdeeld in deelrequirements die onderdeel vormen van de gehele requirements.

Tabel 9: Functionele Requirements

Requirement	Omschrijving	Tracability	Prioriteit
FR 1	Het systeem moet doorlooptijden van brugopeningen weergeven. 1.1 Het systeem moet doorlooptijden kunnen aggregeren op tijdsperiode. 1.2 Het systeem moet doorlooptijden kunnen aggregeren op objecttype. (Ophaalbrug, Draaibrug, Hefbrug)	U01	MUST
FR 2	Het systeem toont storingen per object 2.1 Het systeem toont duur storingen. 2.2 Het systeem toont frequentie storingen. 2.3 Het systeem moet storingen per objecttype kunnen aggregeren.	U02	MUST
FR 3	Het systeem toont abnormaliteit storingen. 3.1 Het systeem toont uitschieters in doorlooptijd storing. 3.2 Het systeem toont uitschieters in frequentie van storingen. 3.3 Het systeem toont storingsgedrag ten opzichte van andere bruggen.	U03	MUST
FR 4	Het systeem toont informatie over de brug. 4.1 Leeftijd 4.2 Type brug	U04, FR3	SHOULD

	4.3 Areaal		
FR 5	Het systeem toont statistische informatie over het functioneren van bruggen. 5.1 Aantal brugopeningen 5.2 Aantal meldingen vanuit SCADA 5.3 Gemiddelden van bovenstaande 5.4 Doorsneden op dag, maand, en jaarniveau	U05	MUST
FR 6	Het systeem toont correlaties tussen storingen en andere factoren. 6.1 Temperatuur 6.2 Neerslag 6.3 Windkracht 6.4 Windrichting	U06	COULD

5.2 Data mining requirements

Tabel 3 geeft een overzicht van de data mining requirements gesteld aan het project. Binnen CRISP-DM vallen deze onder de data mining goals van de organisatie. Deze doelstellingen zijn opgesteld aan de hand van de business goals die zijn gevonden bij de overleggen met Robert van Lakwijk en Arnoud de Kruijf.

De prioritering van deze requirements is afgeleid aan een overleg met Martijn Koole en Martijn van den Boor, werkzaam bij het Datalab van Rijkswaterstaat Delft. Dit overleg vond plaats tussen het overleg met Robert van Lakwijk en Arnoud de Kruijf. In dit overleg zijn de doelstellingen die uit het eerte overleg kwamen besproken en is de richting van het onderzoek gescopet zodat er afstemming is met de projecten van het Datalab. Hieruit is gekomen dat de richting patroonherkenning een goede weg is naar het in kaart brengen van het gedrag van bruggen in relatie tot storingen. Deze richtlijn is vervolgens aangehouden om de data mining requirements te prioriteren.

Tabel 10: Data mining requirements

Requirement	Omschrijving	Tracability	Prioriteit
DMR 1	Het systeem legt aanloop naar storing vast. 1.3 Het systeem legt vast welke alarmen leiden tot een storing 1.4 Het dashboard toont de storingen en aanloop in een overzicht	Arnoud de Kruijf, punt 2c, U03, U06	SHOULD
DMR 2	Dashboard toont patronen in Storingsgedrag. 2.3 Het systeem zoekt verbanden tussen voorkomende alarmen en storingen. 2.4 Het systeem toont verbanden in overzicht.	Arnoud de Kruijf, 1a, ii, U03, U06	MUST

DMR 3	Systeem brengt oneigenlijke bedienhandelingen in kaart. 3.1 Het systeem zoekt verbanden tussen bedienpatronen en storingen 3.2 Het systeem toont storingen die gevolg (kunnen) zijn van afwijkend bediengedrag.	Arnoud de Kruijf, punt 3, U03, U06	COULD
DMR 4	Het systeem classificeert storingen. 4.1 Systeem maakt onderscheid in storingen die uitval veroorzaken en storingen die geen uitval tot gevolg hebben.	Robert van Lakwijk, punt 3a, U03, U06	COULD
DMR 5	Het systeem zoekt verbanden tussen storingen en andere factoren. 5.1 Weersgegevens ten opzichte van storingsgedrag	Arnoud de Kruijf, Dashboard punt 2, U03, U06	COULD

5.4 Non-functionele requirements

Niet functionele requirements zijn opgesteld volgens de kwaliteitsattributen gespecificeerd in ISO 25010. Per requirements is aangegeven onder welke kwaliteitsattribuut deze valt. De requirements zijn zo opgesteld dat zij de inrichting van een dashboard dat data mining analyses uitvoert faciliteert. Zo kan dit project in volgende iteraties van vervolgonderzoek gebruikt worden om verdere analyses in weer te geven.

Requirement	Omschrijving	Tracability	Prioriteit
NFR 1	Het systeem moet analyses uit kunnen voeren als achtergrond taak.	Arnoud de Kruijf, Dashboard punt 1	MUST
NFR 2	Het systeem sluit aan op de architectuur gebruikt door het Datalab in Delft.	Roel van Endhoven	SHOULD

Bijlage B: Datakwaliteitsrapport

Data Kwaliteitsrapportage

SCADA Gegevens eenentwintig bruggen

Roel van Endhoven
HZ University of Applied Science
Versie 2.0

Inhoud

1. Inleiding

In dit document is een overzicht opgenomen van de gegevenskwaliteit van de SCADA gegevens van 21 bruggen in Regio Zuid Nederland. De kwaliteitsrapportage vormt de basis van de bruggen die worden gekozen in het verdere onderzoek m.b.t het toe passen van data mining modellen. Het document is opgedeeld in verschillende invalshoeken om naar datakwaliteit te kijken. Eerst wordt de structuur van de gegevens in kaart gebracht en getoetst op uniekheid en accuraatheid. Dan wordt de compleetheid van de gegevenssets verkend. Tot slot wordt gecontroleerd of ook de consistentie van de gegevens klopt en of de beschikbare gegevens valide zijn.

1.1. Keuze Kwaliteitsdimensies

Om de kwaliteit van de gegevens vast te leggen wordt gebruikt gemaakt van kwaliteitsdimensies zoals gedefinieerd door Askam et al. (2013). Deze dimensies zijn:

- Completeness
- Uniqueness
- Timeliness
- Validity
- Accuracy
- Consistency

De kritieke dimensies worden geoperationaliseerd aan de hand van scenario's welke definiëren aan welke criteria een dimensie moet voldoen. Zo kan een score gegeven worden aan de data per kwaliteitsdimensie. De kritieke dimensie voor het nauwkeurig kunnen doen van uitspraken over het functioneren van de assets is compleetheid van de gegevens. Wanneer er records ontbreken, kan er geen volledig beeld van de waarheid worden vastgelegd.

2. Kwaliteitsdimensies

De SCADA event logs produceren gegevensregels altijd in een vast formaat. Omdat er geen menselijke handeling bij betrokken is zijn deze gegevens altijd hetzelfde formaat. De events zijn voorzien van een timestamp, welke ze uniek maakt voor de tijd waarop de events opdoen.

2.1 Uniekheid en Nauwkeurigheid

Het formaat van de SADA gegevens is te vinden in tabel 1.

Tabel 11: Structuur SCADA gegevens

Kolom	Type	Verplicht	Waarden
Tijd	Timestamp in ISO-8601	JA	timestamp
Type	String	JA	{PROCES, CAME, CAME-ACK, WENT, WENT-ACK, PROCES-DETAIL , LOG-IN, LOG-UIT}
Klasse	integer	NEE	Zie klassen
Waarde	{BOOL, float, int}	NEE	meerdere
ID	int	NEE	Bij alarmen corresponderend met DPE, anders NULL
Omschrijving	String	JA*	n.v.t.
Status	String	NEE	{Openstaand, In Behandeling}
Prioriteit	String	NEE	***
Tijd came	Timestamp in ISO-8601	NEE	timestamp
Tijd went	Timestamp in ISO-8601	NEE	timestamp
Tijd ack	Timestamp in ISO-8601	NEE	timestamp
Gebruiker ID	integer	NEE	integer
DPE	String	JA	string

* De Bediencentrale Tilburg kent records met een lege omschrijving

2.1.3 Klassen en Prioriteiten

De SCADA systemen maken onderscheid tussen de volgende zeven klassen:

1. Klasse 1 (urgent)
2. Klasse 2 (niet urgent)
3. Klasse 3 (proces en bediening)
4. Klasse 4 (onderhoud)

5. Klasse 5 (onderhoud bericht)
6. Klasse 6 (attentie)
7. Klasse 7 (bewaking)

Daarnaast maken de SCADA systemen onderscheid tussen de volgende prioriteitsniveaus:

1. Urgent
2. Bewaking
3. Waarschuwing
4. Urgent onderhoud
5. Onderhoud
6. Attentie

2.2 Compleetheid

Een vraag die zich opdeed tijdens overleg met Istimewa Electro en RWS is wat de kwaliteit van de huidige gegevens is met betrekking tot volledigheid van de logging. Tot op de maand van augustus werden logfiles geschreven tot aan een maximale grootte, waarna een nieuw logfile werd geschreven. Het kon bij die implementatie voorkomen dat beschikbare ruimte voor logging al volgeschreven stond voordat er een volledige maand doorlopen is. Om deze reden is de compleetheid van de gegevenssets meegenomen als kwaliteitsdimensie. Compleetheid wordt gedefinieerd door missende gegevens in kaart te brengen. Voor deze dimensie zijn scenario's gespecificeerd, voorzien van implementatie in de vorm van pseudocode. Deze scenario's zijn te vinden in tabel 2.

Tabel 12: Scenario's compleetheid

Scenario ID	Dimensie	Beschrijving	Implementatie
001	Completeness	Compleetheid wordt gedefinieerd door missende gegevens in kaart te brengen. Dit gaat in het geval van de SCADA gegevens om de dagen waarop events niet zijn gelogd of waar data ontbreken. Dit geeft inzicht in de kwaliteit van de logging op dag niveau	Compleetheid = totaal aantal dagen zonder gegevenslogging / aantal dagen van eerste record naar laatste record
002	Completeness	Een ander geval van missende gegevens is waanneer het geheugen van een brug vol loopt. Op dit moment stopt de brug met gegevens loggen tot er weer geheugen beschikbaar is.	Compleetheid = timedelta(geheugen-vol-melding, eerstvolgende-melding) for geheugen-vol-melding in log-events

003	Completeness	Inzicht in compleetheid kan ook verkregen worden door simpelweg te tellen hoe vaak het geheugen vol loopt.	Compleetheid = COUNT(geheugen-vol-melding)
004	Consistency	De mate van consistentie wordt gemeten door te kijken naar of de data voldoet aan de verwachte trends of patronen.	Consistency = Geregistreerde logging is conform het patroon van voorgaande gelogde gegevens

2.2.1 Resultaten scenario 001 en 003

In tabel 3 is een overzicht van dagen waarop het aantal gelogde gegevens kleiner is dan tien. Eerst werden alleen compleet missende dagen geteld maar daar bleek uit dat er dagen waren waarop enkel de back-up van het systeem werd gelogd, waardoor er wel events binnen die dag vielen. Daarentegen produceerde de brug geen enkele andere events. Dit laatste wordt getoetst binnen de dimensie consistentie.

Tabel 13: Score Bruggen compleetheid data

Brug	Missende dagen	Score	Geheugen Vol Meldingen
Amertakbrug	2	99.16%	0
Dr. Deelenlaan	3	98.73%	0
Sonse brug	26	89.03%	3
Stadsbrug Weert	3	98.73%	0

2.2.2 Resultaten scenario 2

Wanneer gekeken wordt naar de intervallen tussen de momenten waarop de geheugen vol meldingen werden gelogd is dit in twee gevallen significant:

02/05/2018 08:22:33 tot 02/05/2018 11:27:06 = 3.08 uur

06/06/2018 17:53:27 tot 08/06/2018 09:00:39 = 39.12 uur

In dit laatste geval is deze missende data zeer significant omdat op het moment dat de logging weer begint er verschillende storingen opdoen. Een fragment hiervan is te zien in tabel 4.

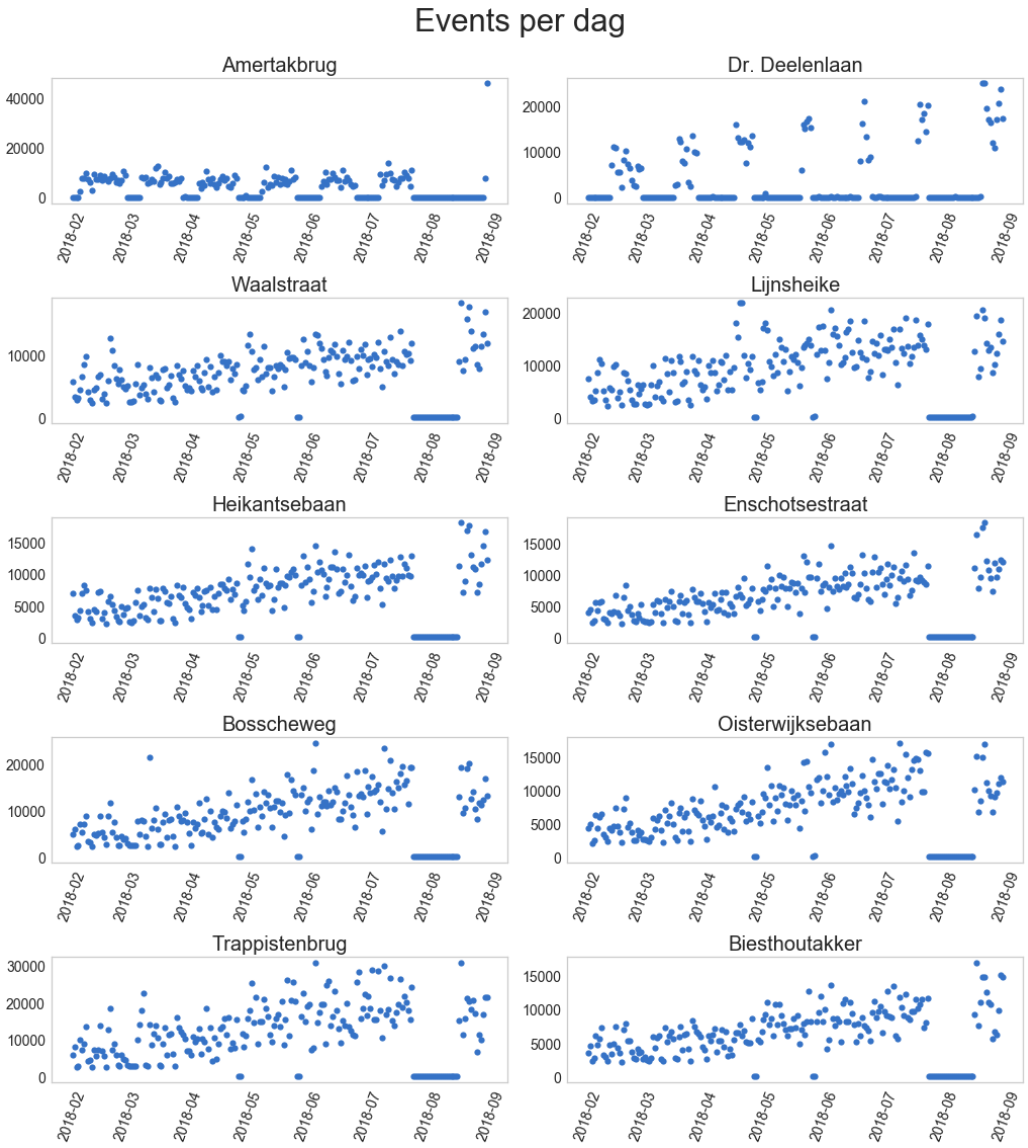
Tabel 14: Storingen na memory reset

tijd	type	omschrijving
2018-06-08 09:00:00.000000	CAME	Geen communicatie met audio-rack

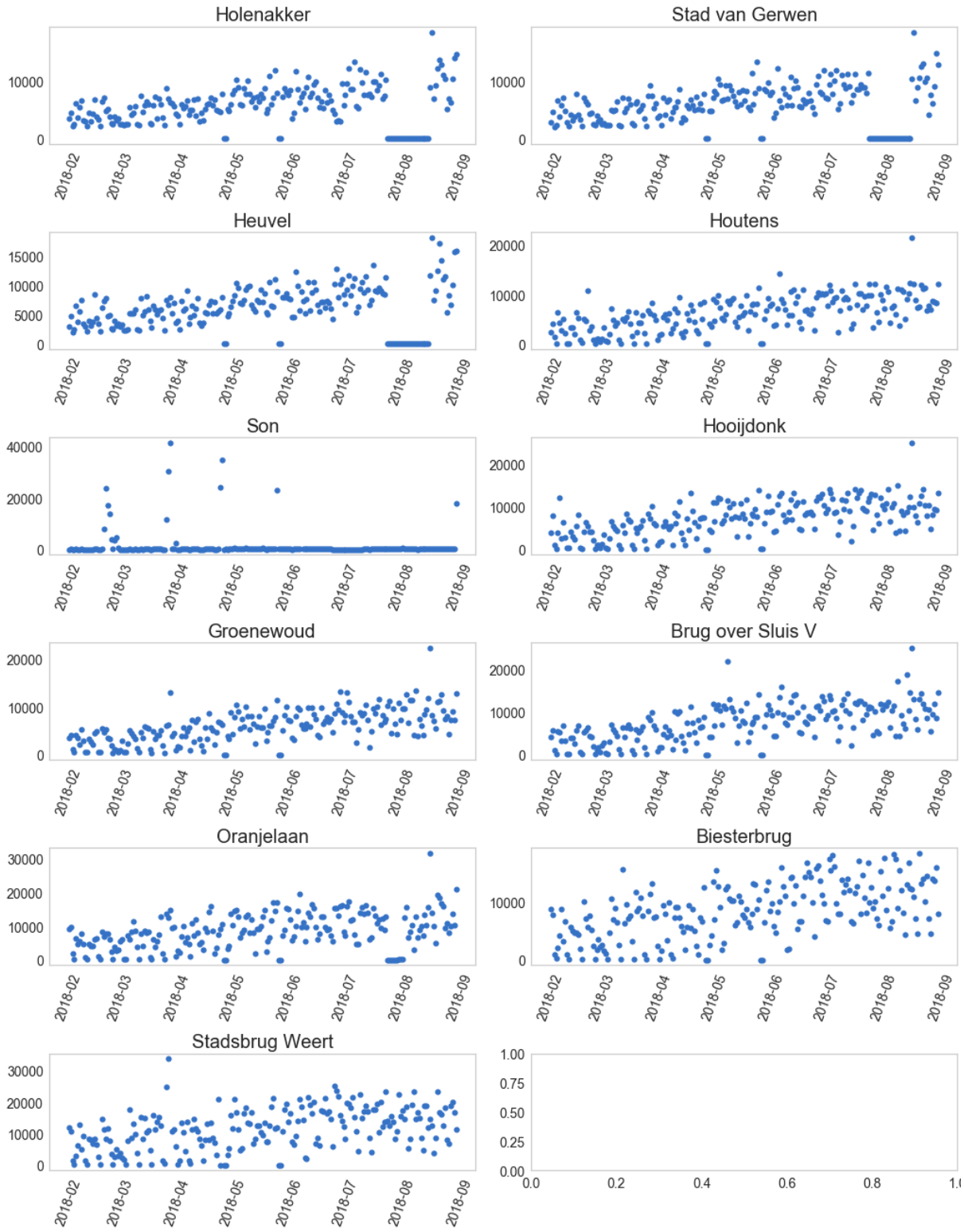
2018-06-08 09:00:30.000000	CAME	NCT-OPC verbinding storing
2018-06-08 09:00:30.000000	CAME	Client Lokaal SCADA Heartbeat Bewaking Aangesp...
2018-06-08 09:00:30.000000	CAME	Client Afstand SCADA Heartbeat Bewaking Aanges...
2018-06-08 09:00:30.000000	CAME	Geen communicatie met audio-rack
2018-06-08 09:00:34.887000	CAME	PLC verbinding storing
2018-06-08 09:00:34.887000	CAME	NO storing bediening en besturing
2018-06-08 09:00:34.887000	CAME	UPS 1 Temperatuur Te Hoog

2.3 Consistentie en nauwkeurigheid

Om de consistentie (Scenario 004) en validiteit van de SCADA logs in kaart te brengen zijn de gegevens per brug geplot waarin het aantal records per dag is genomen als meetwaarde. In de gegevens is een duidelijk seizoenspatroon te zien bij alle bruggen welke meer gegevens loggen in de zomermaanden. Daarnaast is bij een groot aantal bruggen te zien dat er gaten in de logging lijken te zijn. In de periode half juli tot half augustus worden heel weinig gegevens gelogd. Dit is niet consistent met de rest van de gegevens welke een groeiend aantal events voorspelt. In figuur 1 en 2 is voor iedere brug een grafiek uitgezet met het aantal events per dag tegenover de tijd.



Figuur 6: Events per dag brug 1-10



Figuur 7: Records per dag bruggen 11-21

Om te verifiëren of deze gegevens correct zijn en daarmee voldoen aan de nauwkeurigheidsdimensie is om expert opinie gevraagd. Bij gebrek aan bevestiging over de juistheid van deze gegevens is de aanname gemaakt dat de gegevens niet consistent zijn gelogd en wordt beschouwd dat de bruggen Brug Waalstraat tot en met Brug Heuvel een gat hebben in consistentie en validiteit van één op acht maanden, ofwel 12.5%.

3. Bibliografie

(Askham, N., Cook, D., Doyle, M., Fereday, H., Gibson, M., Landbeck, U., ... & Schwarzenbach, J. (2013). The six primary dimensions for data quality assessment. DAMA UK Working Group, 432-435.)

Bijlage C: Zoekplan Bronnenonderzoek: Toepassing Data Mining op SCADA gegevens en beschikbare modellen

Zoekplan Bronnenonderzoek

Toepassing Data Mining op SCADA gegevens en beschikbare modellen

Dit document bevat een overzicht van de gevonden bronnen voor de literatuurstudie naar toepassingen van data mining technieken richting predictive maintenance. Eerst is een lijst opgesteld met Trefwoorden m.b.t data mining technieke. Vanuit de gevonden bronnen zijn vervolgens modellen gezocht die toepasbaar zijn op de SCADA gegevens. De taal die is aangehouden voor het bronnenonderzoek is Engels door de voornamelijk Engelstalige terminologie binnen het vakgebied.

Trefwoorden vastgelegd in onderzoeksmethode
Data Mining
SCADA
Event log
Logging
Log mining
Trefwoorden afgeleid uit data mining requirements
(DMR 2) Patronen -> Pattern
(DMR 2) Verbanden -> Relations
(DMR 4) Classificeren -> Classification
Trefwoorden gevonden in literatuur
Pattern Mining
Frequent Pattern Detection
Association Rules
Rare Pattern Mining
Association Rule Mining
Apriori
Apriori-Inverse
FP-Growth
Sporadic Rules
Market basket analysis
Eclat

Hierbij zijn de volgende bronnen gevonden over het toepassen van datamining op log gegevens en eventueel de toepassing op SCADA logs.

Auteur	**Titel**	**Informatie (uitgever, tijdschrift , etc.)**	**Waar gevonden**

Agarwal, R., & Srikant, R.	Fast algorithms for mining association rules.	In Proc. Of the 20th VLDB Conference 1994, September Pagina's 487-499	Scholar.google.com
Bayardo, R. J., Agrawal, R., & Gunopulos, D.	Constraint-based rule mining in large, dense databases.	In Data Engineering, 1999. Proceedings., 15th International Conference IEEE. Pagina's 188-197 1999, Maart	Scholar.google.com
Borgelt, C.	Frequent item set mining.	Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, Volume 2 Issue 6, Pagina's 437-456. 2012	Scholar.google.com
Borgelt, C.	An Implementation of the FP-growth Algorithm.	Proceedings of the 1st international workshop on open source data mining: frequent pattern mining implementations ACM. Pagina's 1-5 2005, Augustus	Scholar.google.com
Czora, S., Dix, M., Fromm, H., Klöpper, B., & Schmitz, B	Mining Industrial Logs for System Level Insights	Datenbanksysteme für Business, Technologie und Web 2017	Scholar.google.com
Eickmeyer, J., Li, P., Givehchi, O., Pethig, F., & Niggemann, O.s	Data Driven Modeling for System-Level Condition Monitoring on Wind Power Plants.	In DX@ Safeprocess Pagina's 43-50 2015	Mining Industrial Logs for System Level Insights
Han, J., & Fu, Y.	Mining multiple-level association rules in large databases.	IEEE Transactions on Knowledge & Data Engineering Issue 5 Pagina's 798-805 1999	Scholar.google.com

Koh, Y. S., & Rountree, N.	Finding sporadic rules using 52priori-inverse	In Pacific-Asia Conference on Knowledge Discovery and Data Mining 2005, Mei Pagina's 97-106 Springer, Berlijn, Heidelberg	Researchgate.com
Pei, J., Han, J., Mortazavi-Asl, B., & Zhu, H.	Mining access patterns efficiently from web logs.	In Pacific-Asia Conference on Knowledge Discovery and Data Mining Springer, Berlin, Heidelberg. Pagina's 396-407 2000, April	Scholar.google.com
Schmidt-Thieme, L.	Algorithmic Features of Eclat.	FIMI. (2004, November).	Scholar.google.com
Sipos, R., Fradkin, D., Moerchen, F., & Wang, Z.	Log-based predictive maintenance.	In Proceedings of the 20th ACM SIGKDD international conference on knowledge discovery and data mining ACM. Pagina's 1867-1876 2014, augustus	Scholar.google.com
Vaarandi, R.	(2004). A breadth-first algorithm for mining frequent patterns from event logs.	In Intelligence in Communication Systems Pagina's 293-308 Springer, Berlin, Heidelberg. (2004)	Scholar.google.com
Verma, A. P.	Performance monitoring of wind turbines: a data mining approach	Graduate College of the University of Iowa, Iowa City, ABD. 2012	Scholar.google.com
Wang, J., Li, C., Han, S., Sarkar, S., & Zhou, X.	Predictive maintenance based on event-log analysis: A case study.	IBM Journal of Research and Development, Volume 61 uitgave 1, Pagina's 11-121 2017	Scholar.google.com

Bijlage D: Overzicht Pattern mining algoritmes

Itemset Mining en Association Rules

Overzicht en werking

Roel van Endhoven
HZ University of Applied Sciences
Versie 1.0

Inhoud

1. Inleiding

In dit document geeft een overzicht van verschillende itemset mining algoritmes. Deze algoritmes worden gebruikt binnen association rule mining voor het vinden van itemsets. De algoritmes zijn vastgelegd in het bronnenonderzoek naar toepassing van patroonherkenning op SCADA gegevens. De gevonden algoritmes zijn te vinden in tabel 1. Dit document bespreekt de probleemstelling van association rule mining en frequent itemset mining. Daarna geeft het een overzicht van verschillende gevonden itemset mining algoritmes. Tot slot geeft het een overzicht van interssantheidsmaten welke aangeven hoe belangrijk een association rule is.

Tabel 15: Gevonden Algoritmes

Algoritme
Apriori
Apriori Inverse
FP-Growth
Eclat

1.2 Probleemstelling

Frequent itemset mining is een proces dat begint met een set van transacties $T = \{t_1 \dots t_n\}$ waar iedere transactie een set is van items in een database $I = \{i_1 \dots i_m\}$, volgens de regel $t \subseteq I$. Een voorbeeld hiervan zijn winkelmanden met producten in een supermarkt. Een frequente itemset wordt gedefinieerd als een itemset welke in meer transacties voorkomt dan een vooraf gespecificeerde support:

$$supp(X) = \frac{|t \in T; X \subseteq t|}{|T|}$$

De support is het aantal keer dat een itemset X als subset voor komt in T. Wanneer de support hoger is dan gespecificeerd door de gebruiker wordt het gezien als frequent *(Leskovec, Rajaraman, & Ullman, 2014).* Rare itemset mining is een variant van itemset mining die zich richt op het vinden van zeldzame verbanden die een zeer lage support hebben (Koh & Rountree, 2005).

Association rule mining bestaat uit het vinden van verbanden vinden in frequente itemsets. Het is één van de meest gebruikte vormen van data mining (Koh & Rountree, 2005). Deze verbanden, genaamd association rules, zijn regels in de vorm van $A \Rightarrow B$ waar A het antecedent is en B de consequent waartoe het antecedent leidt. De significantie van deze regels kan worden beschreven door de support en confidence statistieken. De support geeft het aantal transacties weer waarin zowel A als B samen voorkomen. Dit wordt uitgedrukt als de waarschijnlijkheid (probability) dat een transactie de unie van sets A en B bevat, of zowel A als B.

$$supp(A \Rightarrow B) = P(A \cup B)$$

De confidence geeft aan in hoeveel gevallen dat A voorkomt, B ook voorkomt (Czora et al., 2017). Ofwel, wat is de kans dat de consequent voorkomt wanneer de antecedent voorkomt? Uitgedrukt als een conditionele kansfunctie:

$$confidence(A \Rightarrow B) = P(B \mid A)$$

Dit valt tevens te noteren gebruikmakend van de support:

$$confidence(A \Rightarrow B) = \frac{supp(A \cup B)}{supp(A)}$$

2. Algoritmes

Dit hoofdstuk geeft een beschrijving van de gevonden algoritmes en hun implementatie. Daarbij wordt uitgelegd hoe het algoritme werkt.

2.1 Apriori

Het Apriori algoritme voor het vinden van association rules bestaat uit twee stappen. Eerst worden de frequent voorkomende itemsets gevonden in een transactie database. Dit zijn de items in de dataset met een support hoger dan het gespecificeerde minimum. De tweede stap is het genereren van association rules vanuit de frequente itemsets. Hierbij is het vinden van de frequente itemsets vaak de belangrijkste stap omdat het de meeste procestijd kost. Om het genereren van frequente itemsets efficiënt te maken gaat Apriori uit van het feit dat geen enkele superset van een niet frequente itemset ($support(itemset) < minsup$), frequent kan zijn. Dit is omdat de set de support aan neemt van de subset met de minste frequentie. Ofwel als set $\{3\}$ vijf keer voorkomt kan set $\{1, 3, 4\}$ nooit meer dan vijf keer voorkomen (Borgelt & Kruse, 2002).

Agarwal & Srikant (1994) beschrijven de procedure dat Apriori doorloopt om frequente itemsets te vinden als volgt:

De input van het proces is de database met sets van transacties.

Procedure:

- Eerst telt het algoritme hoe vaak ieder item voorkomt om zo de frequente sets van 1 item te vinden (1-itemsets). Dit herhaalt zich voor ieder item.
- Kandidaat Itemsets van lengte (k+1) worden gegenereerd vanuit itemsets met lengte k.
- Subsets van lengte k die niet groter zijn dan de minsup worden geschrapt.
- De support van iedere kandidaat itemset wort geteld door de database te scannen.
- Kleine kandidaat itemsets worden geschrapt.

2.2 Apriori-Inverse

Apriori Inverse is een variant van het Apriori algoritme voor het vinden van sporadische association rules. Sporadische association rules worden gedefinieerd als association rules die lage support hebben maar een hoge confidence (Koh & Rountree, 2005). Sporadische association rules hebben een support lager dan een door de gebruiker gedefinieerd maximum en een confidence hoger dan een dan door de gebruiker gedefinieerd. De procedure voor het vinden van sporadische itemsets is als volgt:

- Eerst wordt door de database gegaan en wordt een omgekeerde index van unieke items (1-itemsets) en de transactie ID's waarin zij voorkomen gemaakt. De support van de itemset is de lengte van de lijst transacties waarin deze voorkomt.
- Als de support van deze itemsets kleiner is dan de maximum support of groter dan de absoluut minimale support wordt deze geschrapt.
- Vervolgens worden k-itemsets gezocht waar k > 2.
    - Daarbinnen worden alle itemsets gezocht die extensies zijn van itemset k-1. Bijvoorbeeld {1,4,6} is een extensie van itemset {1,4}, ze hebben dezelfde prefix {1,4}.

    - Dan wordt gecontroleerd of de extensies een support hebben die hoger is dan de absolute minimum support.
    - Dit wordt vervolgens toegevoegd aan de set k-itemsets
- De unie van 1 … k –itemsets wordt geretourneerd.

2.3 FP-Growth

Eén van de snelste en populairste algoritmes voor het vinden van frequente itemsets is FP-Growth (Borgelt, 2005). FP-Growth baseert zich op een representatie van een prefix tree van een gegeven database $T = \{t_1 \dots t_n\}$. FP-Growth kan op deze manier geheugen besparen ten opzichte van andere algoritmes voor het vinden van frequente itemsets. Het FP-Growth algoritme wordt omschreven als een recursief eliminatie schema. Binnen een preprocessing stap worden, net als bij de Apriori (Borgelt & Kruse, 2002) en Eclat algoritmes alle items van de transacties verwijderd die niet frequent zijn ($supp(i) < minsup$). Het proces om een FP-Tree te bouwen is als volgt:

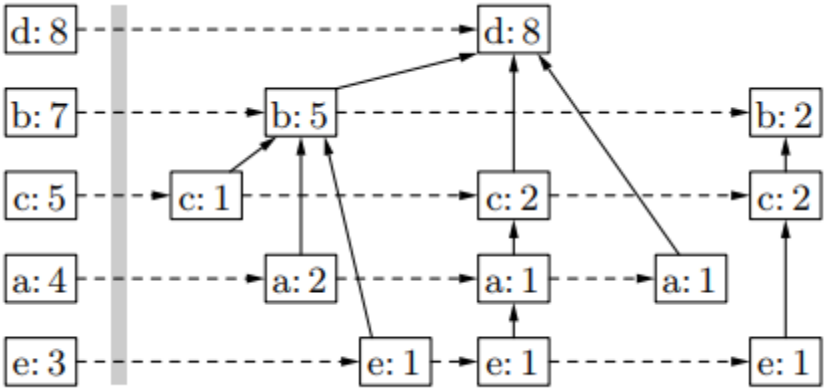
- Niet frequente items worden verwijderd uit transacties omdat ze nooit deel kunnen zijn van een frequente itemset.
- Daarnaast worden de items in iedere transactie in aflopende volgorde gesorteerd met betrekking tot de frequente waarop ze voorkomen in de database. Zie figuur 1.

a d f	d a
a c d e	d c a e
b d	d b
b c d	d b c
b c	b c
a b d	d b a
b d e	d b e
b c e g	b c e
c d f	d c
a b d	d b a

d	8
b	7
c	5
a	4
e	3
f	2
g	1

Figuur 8: Preprocess FP-Tree met gesorteerde items (Borgelt, 2005)

- Nadat alle individueel infrequent voorkomende items zijn verwijderd wordt een FP-Tree opgebouwd als prefix tree. Ieder pad binnen een FP-Tree geeft een set van transacties aan met dezelfde prefix (bijvoorbeeld {d, b, a} en {d, b, c} delen de prefix {d, b}). Formeel gezien leidt dit tot $prefix(A, B) = A \cap B$. In figuur 2 is de implementatie van een FP-Tree te zien van de database in figuur 1.

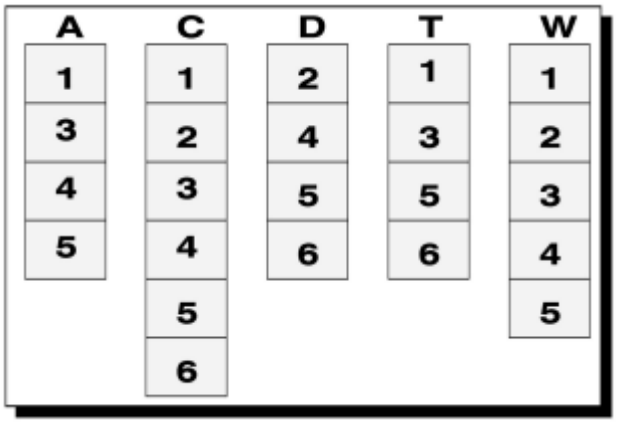


Figuur 9: FP-Tree gebaseerd op database in figuur 1 (Borgelt, 2005)

- Daarnaast wordt een lijst opgesteld met de alle nodes in de FP-Tree die verwijzen naar hetzelfde item. Op deze manier kunnen de paden die het item bevatten snel worden opgezocht. Dit wordt ondergebracht in een head element, welke ook aangeeft hoe vaak het item voorkomt in de database (Borgelt, 2005).

2.4 Eclat

Het Eclat algoritme is een depth-first search algoritme voor het vinden van frequente itemsets. Het opereert op een verticale transactie database. (Kaur, 2014). De meeste algoritmes baseren zich op een horizontale database zoals weergeven in de linkerkant van figuur 1, waar iedere regels gezien wordt als een transactie. Eclat baseert zich op een verticale database zoals weergeven in figuur 3 (Zaki, 2000). In plaats van een database van transacties waar items in voorkomen, bestaat de database uit items en in welke transacties ze voorkomen.



Figuur 10: Eclat Verticale Database (TID-list) (Zaki, 2000)

Het Eclat algoritme werkt als volgt:

- De frequente 1-itemsets worden gevonden met een enkele database scan. Voor ieder item wordt de TID-lijst gelezen en wordt de counter voor ieder item met 1 verhoogd.
- Vervolgens wordt met een depth-first algoritme gezocht naar verdere k+1 itemsets.

3. Meetinstrumenten Association Rules

Om de associatie te beoordelen interessante verbanden te kunnen identificeren stellen Czora et al. (2017) de Kulczynski (Kulczynski, 1928) maat en de imbalance ratio (IR) voor. De Kulczynski maat baseert zich op het gemiddelde van de twee conditionele kansfuncties voor de associatieregels $A \Rightarrow B$ en $B \Rightarrow A$. De Kulczynski maat is als volgt gedefinieerd:

$$Kulc(A,B) = \frac{1}{2}(P(A|B) + P(B|A))$$
$$= \frac{1}{2}(confidence(B \Rightarrow A) + confidence(A \Rightarrow B))$$

Een Kulczynski score van 0 of 1 geeft respectievelijk een sterke negatieve of positieve correlatie aan tussen de gevonden regels en maken de regel interessant. Een waarde van 0.5 duidt aan dat er geen verband is.

Czora et al. (2017) stellen dat bij een Kulczynski score van 0.5 de imbalance ratio kan worden toegepast om te kijken of een associatieregel mogelijk toch interessant is. De imbalance ratio wordt als volgt berekend (Han et al., 2011):

$$IR(A,B) = \frac{|supp(A) - supp(B)|}{supp(A) + supp(B) + supp(A \cup B)}$$

De imbalance ratio geeft een waarde van 0 tot 1 van de onbalans tussen patronen A en B waar 0 volledig gebalanceerd is en 1 zeer scheve verhouding aanduidt.

Naast deze maten wordt binnen het association rule mining vaak gebruikt gemaakt van de Lift of Interest maat (Tan, Kumar,& Srivastava, 2002), (Merceron & Yacef, 2008). De lift is de ratio van de geobserveerde support van de association rule ten opzichte van wat verwacht wordt wanneer de antecedent en consequent onafhankelijk waren. De Lift is als volgt gedefinieerd:

$$Lift(A \Rightarrow B) = \frac{supp(A \cap B)}{supp(A) \times supp(B)}$$

4. Bibliografie

Agarwal, R., & Srikant, R. (1994, September). Fast algorithms for mining association rules. In Proc. of the 20th VLDB Conference (pp. 487-499).

Borgelt, C. (2005, August). An Implementation of the FP-growth Algorithm. In Proceedings of the 1st international workshop on open source data mining: frequent pattern mining implementations (pp. 1-5). ACM.

Borgelt, C., & Kruse, R. (2002). Induction of association rules: Apriori implementation. In Compstat (pp. 395-400). Physica, Heidelberg.

Han, J., Kamber, M. and Pei, J. (2011). "Data mining: concepts and techniques." Elsevier

Kaur, M. (2014). ECLAT Algorithm for Frequent Itemsets Generation. International Journal of Computer Systems, 1(3), 82-84.

Koh, Y. S., & Rountree, N. (2005, May). Finding sporadic rules using apriori-inverse. In Pacific-Asia Conference on Knowledge Discovery and Data Mining (pp. 97-106). Springer, Berlin, Heidelberg.

Merceron, A., & Yacef, K. (2008, June). Interestingness measures for association rules in educational data. In Educational Data Mining 2008.

Schmidt-Thieme, L. (2004, November). Algorithmic Features of Eclat. In FIMI.

Tan, P. N., Kumar, V., & Srivastava, J. (2002, July). Selecting the right interestingness measure for association patterns. In Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 32-41). ACM.

Zaki, M. J. (2000). Scalable algorithms for association mining. IEEE transactions on knowledge and data engineering, 12(3), 372-390.