Limperg Instituut

Zelflerende algoritmen bij het verbeteren van datakwaliteit

W.J. Willemse (DNB)

Limperg Instituut

Overzicht

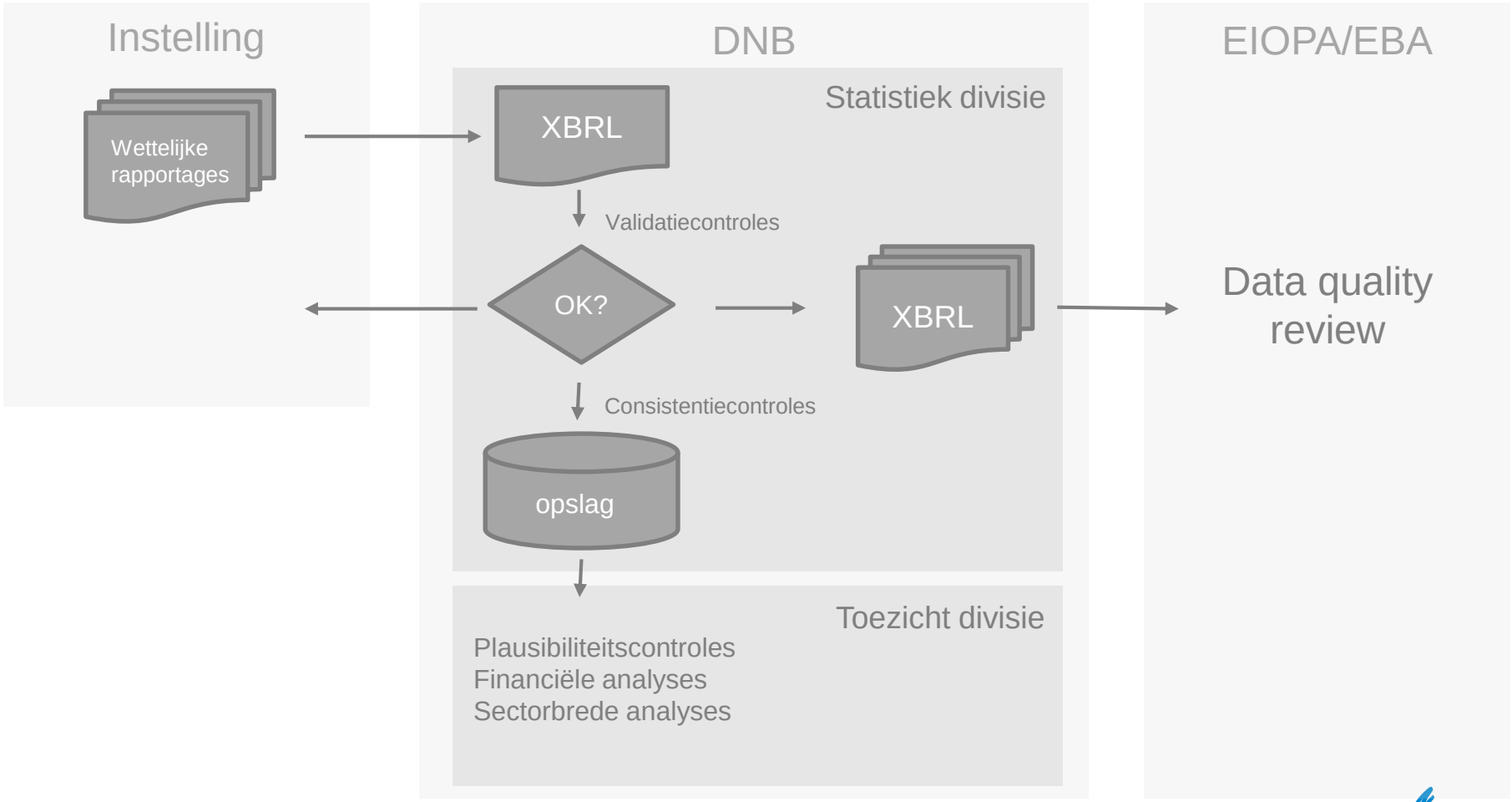
1. Toezichtrapportages bij DNB
   - zeer veel datapunten per entiteit
   - veel domeinkennis nodig bij interpretatie
   - (te) veel werk om alle controles uit te schrijven

2. Zelflerende algoritmen bij plausibiliteitscontroles
   - in staat om verbanden en patronen in data herkennen zonder kennis vooraf
   - genereren signalen die (kunnen) wijzen op datafouten, uitbijters en contra-intuïtieve veranderingen
   - toepassing is beslissingsondersteunend

3. Uitdagingen bij toepassing open source software binnen organisaties

Limperg Instituut

Workflow toezichtrapportages



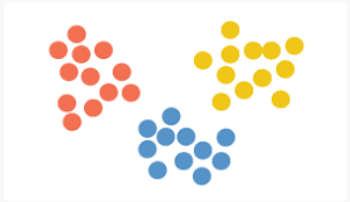
Limperg Instituut

Data workflow plausibiliteitscontroles

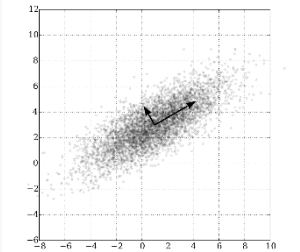
Data clustering

Referentiegroepen

k-means, t-sne

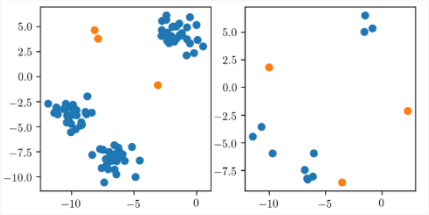


PCA

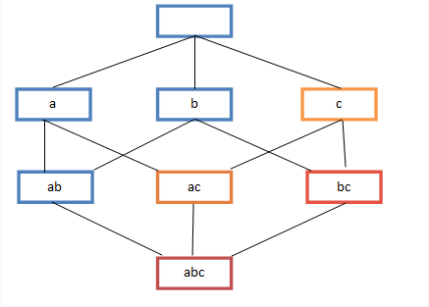


Data analyse

Outlier detection



Patroonherkenning



Expert regels

Resultaten

Signalen

Trendanalyse

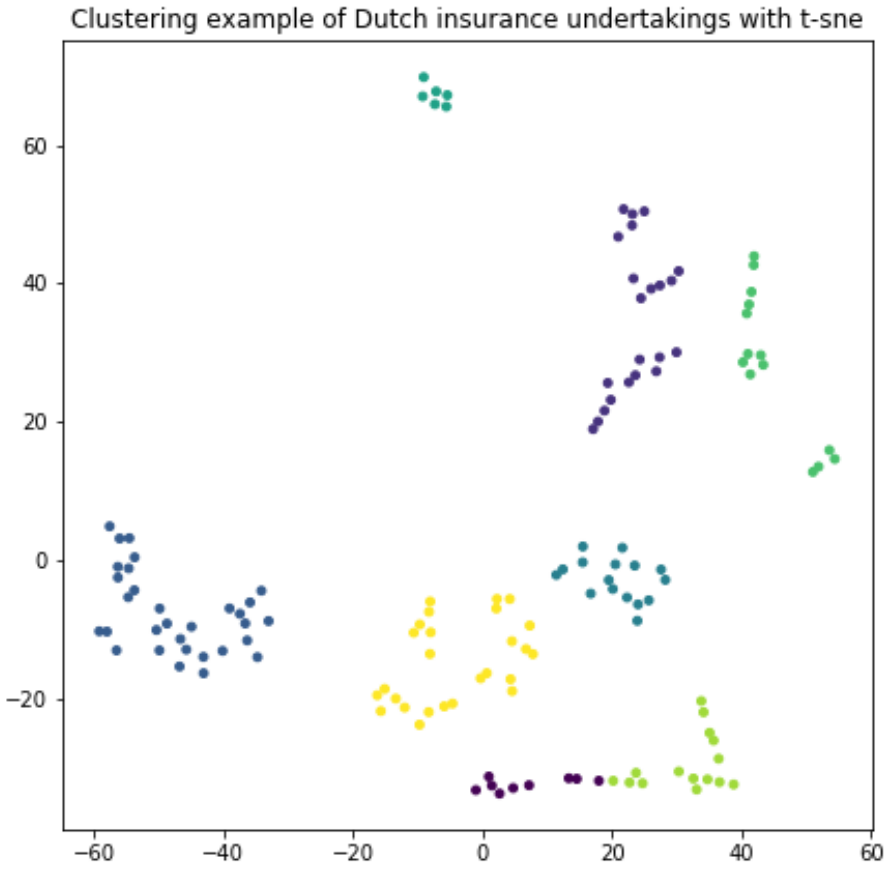
4

VU

Limperg Instituut

Dimensiereductie en clusteren

- Automatische classificatie met dimensiereductie en clustering
- Het t-sne algoritme vat n-dimensionale data samen in 2 dimensies met behoud van overeenkomsten en verbanden
- De afbeelding geeft het resultaat op basis van de balansen van Nederlandse verzekeraars (SII: 80 dimensies)

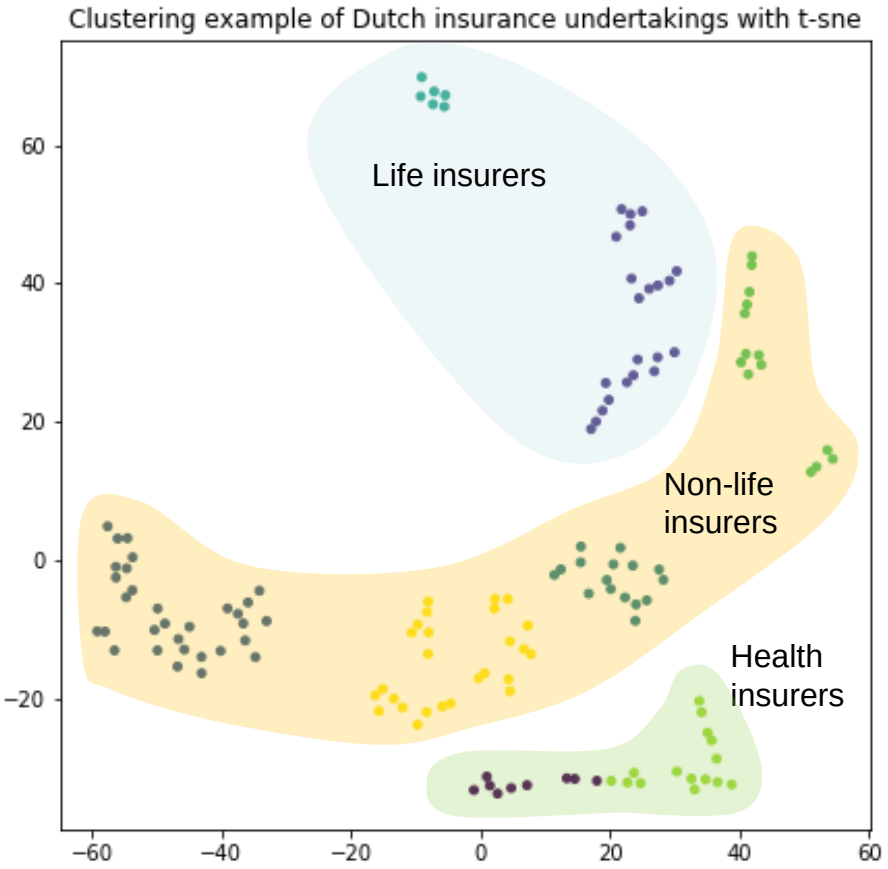


VU

Limperg Instituut

Dimensiereductie en clusteren

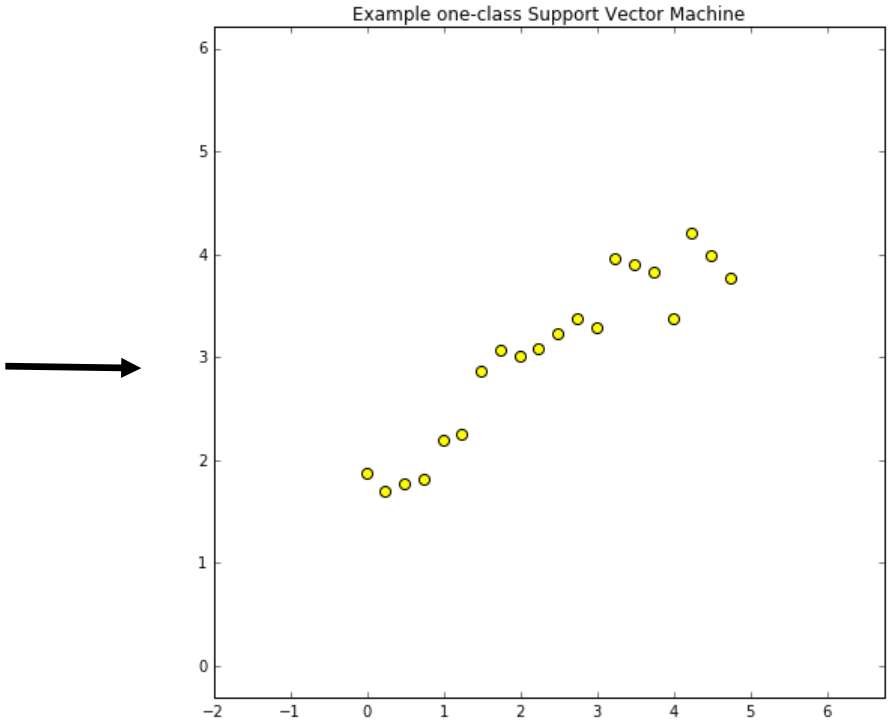
- Automatische classificatie met dimensiereductie en clustering
- Het t-sne algoritme vat n-dimensionale data samen in 2 dimensies met behoud van overeenkomsten en verbanden
- De afbeelding geeft het resultaat op basis van de balansen van Nederlandse verzekeraars (SII: 80 dimensies)



VU

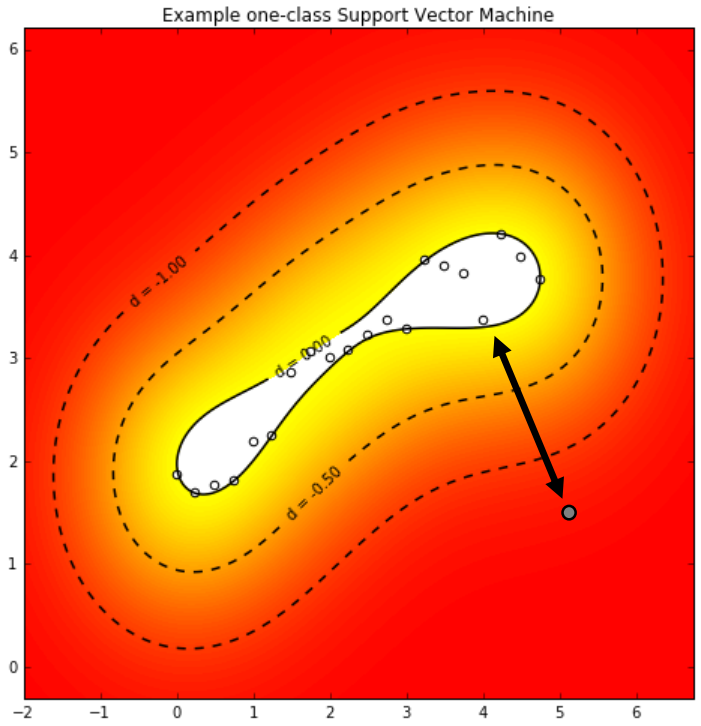
Uitbijter detectie (1)

	X	Y
Data point 1	0.00	1.86
Data point 2	0.25	1.68
Data point 3	0.50	1.77
…	…	…
Data point 20	4.75	3.76



Uitbijter detectie (2)

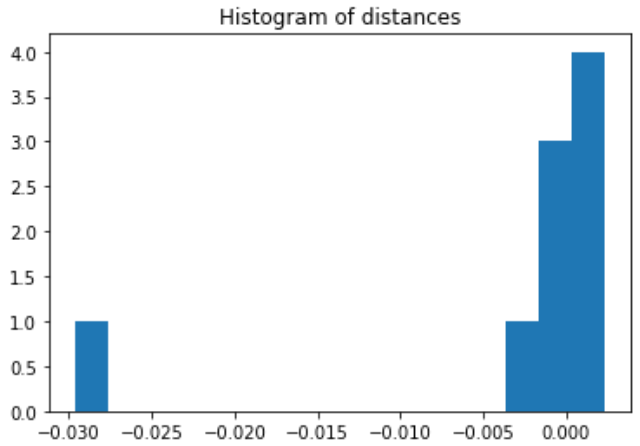
- Een Support Vector Machine bepaalt een *decision boundary* op basis van een gegeven *training set*, bijvoorbeeld historische rapportages
- Uitbijters in nieuwe rapportages worden herkend aan de hand van de afstand tot de *decision boundary*
- De afbeelding laat een 2-dim voorbeeld zien; het algoritme werkt voor *n*-dim data



Limperg Instituut

Uitbijter detectie (3)

Periode	CET-1 vermogen	Total capital ratio	T1 capital ratio	CET-1 capital ratio
2014H2	40543	14,6%	13,5%	13,5%
2015H1	39874	17,0%	14,3%	12,8%
2015H2	41554	16,9%	14,5%	12,9%
2016H1	41986	18,0%	15,1%	13,2%
2016H2	44466	19,3%	16,3%	14,2%
2017H1	44888	19,6%	16,3%	14,5%
2017H2	45581	18,5%	16,2%	14,7%
2018H1	44833	18,4%	15,7%	14,1%
Nieuwe rapportage (hypothetisch)				
2018H2	35000	18,2%	9,7%	13,1%



(Op basis van: individuele gegevens banken (jaar), statistiek.dnb.nl)

Patroonherkenning - associatieregels (1)

- Een associatie regel beschrijft de associatie tussen twee delen in een dataverzameling of tussen dataverzamelingen

	Shopping list
1	bread, milk, eggs
2	bread, milk, eggs, bacon, cheese, muesli
3	milk, muesli
4	bread, eggs

- De associatie {melk} → {muesli} heeft
  - Een ondersteuning / support van 2 / 4 (twee van de vier lijstjes hebben zowel melk als muesli)
  - Een betrouwbaarheid / confidence of 2 / 3 (twee van de drie lijstjes met melk heeft ook muesli)

Patroonherkenning - associatieregels (2)

- Binnen toezichtrapportages zoeken we naar patronen zoals samenhangende datapunten in de rapportages en verbanden met eerdere rapportages

- Het associatie-algoritme ontdekt regels; vervolgens kijken we naar data die de regels bevestigen of juist met de regels in strijd zijn

- Voorbeelden:

Support	Confidence	Rule
134	96%	risk_margin ≤ technical_provisions
132	95%	assets held for unit-linked contracts ≤ technical provisions – unit-linked
122	88%	net scr – market risk = gross scr – market risk
93	89%	delta(reinsurance recoverable from non-life) ≥ 0 and delta(technical provisions – non-life) ≥ 0

(Op basis van: individuele gegevens verzekeraars (jaar), statistiek.dnb.nl)

Limperg Instituut

Uitdagingen bij open source software

- Gebruik van open source software vereist aandacht
  - noodzaak tot gestructureerd beheer van omgevingen en codeversies
  - afspraken nodig om reproduceerbaarheid te waarborgen
  - ondersteuning (zelf) organiseren
- Implementatie in bestaande processen en infrastructuur niet altijd eenvoudig
  - uitlegbaarheid en transparantie
  - feedbackloop met gebruikers
  - toegang tot ongestructureerde data