

Identifikaasje en monitoring fan DNA-groepen yn ûngelokken- en partijdata

Sjoerd Broos, Daan Kolkman, Giedo Donkers · 26 november 2020

Mei data science-techniken liede wy it DNA ôf fan natuerlike groepen ûngelokken en partijen, sadat gemeenten gelyksoartige ûngelokken oer de jierren hinne monitore kinne.

Resultaat en metoade foar it groepearjen en yndielen fan ûngelokken en partijen yn groepen.

S. Broos (VIA), D. Kolkman (MKB datalab), G. Donkers (VIA)

Gearfetting

Mei help fan data science-techniken is it DNA fan natuerlike groepen ûngelokken en partijen te bepalen. Mjitten wurde kin hoe't gelyksoartige ûngelokken en belutsen partijen har oer de jierren hinne ûntwikkelje: monitoring. It kin per gemeente ynsichtlik meitsje wat foar groep gelyksoartige ûngelokken oer de jierren hinne krekt in soad of net folle foarkomt. Der kin fergelike wurde mei itselde soarte gemeenten, gemeenten mei deselde problemen of mei hiel Nederlân. Dêrneist lit it sjen wat foar gelyksoartige partijen der foarkomme by de ûngelokken. Dit ûndersyk brûkt "unsupervised learen" lykas clustering om groepen ôf te liden en heuristiken om DNA te bepalen.

Ynlieding

Data science

Data science (datawittenskip) wurdt hieltyd populêrder en Davenport en Patil (2012) hawwe it berop fan datawittenskipper sels útroppen ta de meast sexy baan fan dizze ieu. Mei nije techniken kin data brûkt wurde om nije ynsjoggen te krijen. Dizze ynsjoggen binne ferskillend, bygelyks it foarsizzen fan de fraach (bgl. bôle) foar in bakker of in mear kompleks fraachstik as it oanrikkemandearjen fan produkten op in webshop. Data wurdt útlêzen, skjinmakke, omfoarme en dêrnei brûkt om in spesifyk doel te berikken. In oantal techniken dy't brûkt binne foar dit ûndersyk wurde yn de folgjende paragrafen útljochte.

Supervised learen

Earst fan alles bestiet der supervised learen; hjir besiket men de relaasje tusken input en output te learen (Kaptein, 2019a). De input is in set fan fariabelen (ek wol features of skaaimerken neamd) dêr't in output tsjinoer stiet. Nim bygelyks in foto wêrfan't jo witte dat it in hûn is. De input binne dan de piksels fan de foto en de output is de groep/kategory hûn. In oar foarbyld is it foarsizzen fan hûsprizen. As input nimme jo de eigenskippen of skaaimerken fan it hûs en as output nimme jo de hûspriis. Dizze foarbylden binne te kategorisearjen as klassifikaasje- en regresjeproblemen. Dit ûndersyk fokust op in klassifikaasje probleem.

Unsupervised learen

Unsupervised learen is in oar type fan machine learning. Hjir hawwe jo wol in input, mar gjin output (Kaptein, 2019b). Hjir hawwe jo bygelyks de skaaimerken fan brûkers op in online webshop, mar witte jo net yn hokker groep brûkers dy krekt falle. Der kinne aktive en net-aktive brûkers wêze, mar ek kompleksere relaasjes lykas brûkers dy't ynteressearre binne yn bepaalde produktsegmenten. Dizze data kin brûkt wurde om natuerlike groepen te ûntdekkjen yn de data. Hjirfoar wurde clustering-algoritmen brûkt. Om it clustering-algoritme te optimalisearjen (wat de rintiid oanbelanget), wurdt geregeld gebrûk makke fan dimensjereduksje-algoritmen. Dizze dimensjereduksje-algoritmen kinne it oantal skaaimerken (oftewol: features, fariabelen) mei tûke techniken reducearje.

Ferkearskundich probleem

It idee is om op basis fan data-analyse gemeenten de mooglikheid te jaan om gelyksoartige ûngelokken en byhearrende partijen rjochter oan te pakken. It úteinlike doel is om ynstânsjes oan te setten ta maatregels. Sa kin in gemeente yn de takomst nei oare soortgelikense gemeenten sjen dy't gelyksoartige ûngelokken effektyf oanpakke. De algemiene term hjir is monitoring: troch de tiid hinne de status fan gemeenten bepale op basis fan ferskillende skaaimerken. Troch it yn kaart bringen en monitoren fan de ferskillende soarten DNA (groepen) kin de gemeente net allinne ynsichtlik krije wat foar kompleksiteit oan ûngelokken en partijen der yn syn gemeente plakfint, mar ek effektiver en effisjinter te wurk gean by it doelmjittich reducearjen fan it oantal ûngelokken. Dit liedt ta de folgjende ûndersyksfraach dy't wy beäntwurdzje sille yn dizze white-paper:

Falle ûngelokken & partijen yn natuerlike groepen (DNA)? As dat sa is, kinne wy dizze groepen (DNA) oer de jierren hinne monitore?

Data science en it ferkear

Data

Foar dit ûndersyk is gebrûk makke fan in part fan de STAR-database[1] mei ûngelokken- en partijdata. Dizze data omfettet skaaimerken fan in ûngelok[2] en skaaimerken fan de belutsen partijen[3] (allinne bestjoerders, fuotgongers en objekten) by in ûngelok. Mocht by in ûngelok in skaaimerke ûnbekend wêze,

dan is foar dat ûngelok gjin DNA ôf te lieden. Dat jildt ek op it stuit dat der in ûngelok is dêr't allinne objekten of bisten by foarkomme. Dêrtroch hawwe wy in dataset fan 205.653 ûngelokken en 520.284 partijen foar de jierren 2017 – 2019.

Data tarieding en ferfolch

Om de ûndersyksfraach te beäntwurdzjen is de hjirboppe beskreaune data analysearre. Guon fariabelen binne ordinaal (folchoarder is wichtich) en guon nominaal (folchoarder is net wichtich). De nominale fariabelen binne útsplitst (one-hot encoding | dummies) en de ordinale fariabelen binne numeryk makke.

Foar it bepalen fan DNA yn de data is der keazen foar clustering (unsupervised learen). Dizze clustering kin ferskillende soarten ûngelokken en partijen yn groepen ûnderferdiele. Dizze groepen (clusters) hawwe elk in eigen betsjutting. Dizze betsjutting kin letter efterhelle wurde. Troch de útsplitsing (one-hot encoding) fan in oantal skaaimerken binne der in soad dimensjes ûntstien. Om it clusterjen makliker ferrinne te litten (koartere rintiid fan it algoritme) is der keazen foar in dimensjereduksje. Dizze dimensjereduksje reducearret it oantal skaaimerken nei in X-oantal dimensjes (dêr't X stiet foar in sels te kiezen oantal dimensjes).

De dimensjereduksjemetoade is principal component analysis (PCA); dy transformearret en kombinearret lineêre relaasjes nei in spesifisearre X-oantal dimensjes. Dêrnei binne de oerbleaune X-dimensjes clustere. Foar de clustering is keazen foar it K-means-algoritme (sawol foar de ûngelokken- as partijdata) (MacQueen, 1967). Nei de clustering is de data transformearre om de wichtichste (dominantste) skaaimerken per groep efter te heljen. Hjirfoar is in heuristyk brûkt. De krêft fan de heuristyk is om domein- en case-spesifike kennis ta te foegjen. Yn dit gefal is de útkomst de dominante fariabelen dy't in cluster definiearje.

Resultaten

Dizze seksje beskriuwt de resultaten fan it ûndersyk. Earst fan alles sil de dimensjereduksje besprutsen wurde, dêrnei de clustering, dan it DNA en as lêste de monitoring.

Dimensjereduksje

De earste stap yn it proses is de dimensjereduksje. De fariabelen binne reducearre nei 25 dimensjes. De keazen 25 dimensjes joegen de bêste resultaten, sjoen nei de úteinlike útkomst (de betsjutting fan de groepen).

Clustering

De data fan de dimensjereduksje wurdt dêrnei brûkt om te clusterjen. It doel is om groepen ûngelokken en partijen ûnderskiede en monitore te kinnen. Earst fan alles binne de ûngelokkeclusters ôflaat mei it K-

means-algoritme. Hjirfoar is k (it oantal clusters) fan 40 bepaald oan de hân fan ferskillende prestaasje-yndikatoaren:

- **Inertia:** jout de ôfstân binnen it cluster oan; dizze skoare moat sa leech mooglik útfalle.
- **Calinski-Harabasz index:** jout de ratio tusken de ôfstân binnen it cluster en tusken clusters. Dizze skoare moat sa heech mooglik útfalle.
- **Davies-Bouldin:** jout oan hoe fier clusters fan elkoar ôf lizze. Dizze skoare moat sa leech mooglik útfalle.

Foar de partijdata is ek it K-means-algoritme brûkt mei $k=48$. Dy binne ek basearre op de hjirboppe neamde prestaasje-yndikatoaren. Dêrneist is de dominânsje (betsjutting) fan it DNA ek meiwage. Wichtich yn de oerweging is dat lytse gebieten ek de mooglikheid hawwe moatte om in cluster te monitoren. It moat fansels net sa wêze dat der safolle clusters binne dat elk cluster in yndividueel ûngelok befettet.

DNA

Elke groep (cluster) hat syn eigen betsjutting. It iene cluster kin in soad ûngelokken befetsje op in krúspunt, wylst it oare cluster in soad ûngelokken befettet dy't krekt net op in krúspunt bard binne. Itselde jildt foar de partijen: it iene cluster kin bestean út útslutend fytters en it oare út útslutend fuotgongers. Om efter te heljen wat elk cluster betsjut, is der gebrûk makke fan in heuristyk. Der binne meardere dominânsjemetoaden ôfwage, lykas klassifikaasje (supervised learen) mei it random forest-algoritme of allinne fisualisaasjes mei bgl. in heatmap. Úteinlik is de heuristyk ûntwikkele om domein- en case-spesifike kennis meinimme te kinnen. Yn de heuristyk binne "aktive" fariabelen en "yllustrative" fariabelen meinommen. De aktive fariabelen binne opnommen om de clusters ôf te lieden en de yllustrative binne allinne bedoeld om de dominânsje te bepalen. Twa foarbylden fan de ûngelokken- en partijclusters binne werjûn yn tabel 1 en 2. Wichtich om hjirby rekken mei te hâlden is dat der sjoen is nei itselde soarte ûngelokken. It hoecht dus net te betsjutten dat elk ûngelok krekt itselde is (oars krije jo ûneinich folle clusters). Dêrtroch binne termen as "faak", "hiel faak", "geregeld", ensfh. tige wichtich. Dit lit de dominânsje sjen fan in bepaald skaaimerke binnen in cluster. As foarbyld: in cluster mei de betsjutting "faak 50 km/h-wegen" wol sizze dat it binnen dit cluster faak foarkomt dat ûngelokken op 50 km/h-wegen plakfine, mar dit is net altyd it gefal.

Tabel 1: Ferliking ûngelokkeclusters

Cluster 9	Cluster 22	Cluster 25
Mix krúspunt/gjin krúspunt	Faak net op in krúspunt	Faak op in krúspunt
Yn de jûnen en moarns (foar de spits)	Oerdei	Oerdei

Geregeld heftiger ûngelokken	Geregeld heftiger ûngelokken	Faak UMS-ûngelokken
Faak 50-wegen	Faak 50 km/h-wegen	Geregeld 60, 80-wegen
Faak tsjusternis	Hiel faak deiljocht	Hiel faak deiljocht
Faak oar asfalt	Faak oar asfalt	Altyd oar asfalt
Hiel faak rein	Hiel faak droech	Altyd droech
Hiel faak 2 partijen	Hiel faak 2 partijen	Hiel faak 2 partijen
Botsing stadich – fluch ferkear	Botsing stadich – fluch ferkear	Botsing fluch ferkear mei in objekt
It hiele jier troch mei pyken yn kwartaal 1 en 4	It hiele jier troch mei pyken yn kwartaal 2 en 3	It hiele jier troch mei pyken yn kwartaal 2
Faak alle leeftiden by belutsenen wêrby't 1 fan de partijen in jongere is	Altyd folwoeksenen by belutsenen	Alle leeftiden dy't tsjin in objekt oan botse

Tabel 2: Ferliking partijclusters

Cluster 1	Cluster 3	Cluster 11
Fytters	Motorfytsen	Auto's en bustsjes
Leeftyd tusken de 58 en 70 jier	Leeftyd tusken de 18 en 21 jier	Leeftyd tusken de 69 en 76 jier

Monitoring

De jûne clusters en harren byhearrende DNA kinne oer de tiid hinne mjitten wurde, sadat ynstânsjes binnen harren gebiet better sicht krije op wat der spilet en dit analysearje kinne. Dit kin bestimpele wurde as monitoring[4]. Tabel 3 jout ynsjoch yn de clustergrutte (absolute en relative wearden) oer de tiid hinne foar ferskillende gebieten: (1) Nederlân, (2) 's-Hertogenbosch, (3) Vught. Yn tabel 3a is te sjen dat cluster 22 absolút en relatyf tanimt. Cluster 25, oan de oare kant, nimt absolút en relatyf ôf. Cluster 9 is stabyl oer de jierren hinne.

Foar in ynstânsje kin it ynteressant wêze om syn gegevens te fergelykjen mei in soartgelikense ynstânsje of mei de trend yn hiel Nederlân. Tabel 3b lit de clustergrutte foar de gemeente 's-Hertogenbosch sjen. Hjiryn is te sjen dat cluster 9 stabyl is en cluster 22 oer de jierren hinne skommelet. Cluster 25 is ôfnommen. Dus cluster 9 en 25 ûntwikkelje har itselde yn 's-Hertogenbosch as yn hiel Nederlân.

Yn Vught binne de oantallen ûngelokken (folle) leger (tabel 3c). Wy kinne lykwols op guon clusters wol monitore. Cluster 9 hat lege oantallen en skommelet oer de jierren hinne. Hjir kinne jo (op dit stuit) net goed op monitore. Cluster 22 en 25 litte lykwols in dúdliker trend sjen. Cluster 22 nimt sterk ta, wylst cluster 25 skommelet. Dus cluster 22 en 25 ûntwikkelje har itselde yn Vught as yn hiel Nederlân.

Tabel 3a: Monitoring ûngelokken Nederlân per cluster (absolút: links, relatyf: rjochts)

Cluster	2017	2018	2019	2017	2018	2019
9	995	980	1109	1,5	1,4	1,6
22	2448	2823	2815	3,6	4,1	4,1
25	3212	3074	2832	4,7	4,5	4,1

Tabel 3b: Monitoring ûngelokken ynstânsje 's-Hertogenbosch per cluster (absolút: links, relatyf: rjochts)

Cluster	2017	2018	2019	2017	2018	2019
9	10	10	10	1,6	1,7	1,7
22	28	31	22	4,3	5,4	3,7
25	30	16	17	4,7	2,8	2,9

Tabel 3c: Monitoring ûngelokken ynstânsje Vught per cluster (absolút: links, relatyf: rjochts)

Cluster	2017	2018	2019	2017	2018	2019
9	2	0	2	2,4	0,0	2,1
22	0	4	6	0,0	4,0	6,2
25	3	2	3	3,6	2,0	3,1

Op dit stuit binne der 3 "random" clusters keazen om in gefoel te krijen foar de betsjutting fan in cluster. Bettere analysestrategyen binne lykwols mooglik op bygelyks gemeentlik nivo. In ynstânsje soe him rjochtsje kinne op groepen dy't fluch groeie of krekt op de grutste groepen oer meardere jierren hinne. De strategy is ek ôfhinklik fan de gemeente. Om it foar lytse gemeenten ynteressant te hâlden, is der in lykwicht socht tusken it oantal clusters en de hoemannichte ûngelokken binnen it cluster. Mei sa'n analyse kinne ynstânsjes bepaalde ûngelokstypen (DNA) identifisearje, monitore en (ûteinlik) maatregels nimme.

Monitoring mei in strategy

Strategy: grutste groepen, ynstânsje: gemeente Vught

Yn de gemeente Vught falle relatyf net folle ûngelokken en is monitoring sa no en dan lestich. In strategy lykas it pakken fan de grutste groepen heart lykwols wol ta de mooglikheden. Clusters 2 (20 ûngelokken [5]), 21 (21 ûngelokken[5]) en 36 (15 ûngelokken[5]) binne yn totaal goed foar 20% fan alle ûngelokken yn de jierren 2017-2019 yn de gemeente Vught (as referinsje: dit binne mar 3 fan de 40 clusters, oftewol 7,5%). Cluster 2 binne benammen ûngelokken dy't faak net op in krúspunt foarkomme, oerdei mei deiljocht, faak 50 km/h-wegen, geregeld heftiger, faak oar asfalt, hiel faak 2 partijen, en stadich ferkear dat botst mei fluch ferkear (figuer 1).

Figuer 1: Foarbyld hoe't de ûngelokken fan cluster 2 der útsjen kinne soene.

Figuer 2: Foarbyld hoe't de ûngelokken fan cluster 21 der útsjen kinne soene.

Figuer 3: Foarbyld hoe't de ûngelokken fan cluster 36 der útsjen kinne soene.

Cluster 21 binne benammen ûngelokken dy't oerdei mei deiljocht barre, geregeld heftiger, geregeld 30 km/h-wegen, hiel faak op klinkers, hiel faak droech, faak 2 partijen, hiel faak in botsing tusken stadich en fluch ferkear (figuer 2). Cluster 36 binne benammen ûngelokken dy't faak op krúspunten barre, hiel faak droech, faak oerdei mei deiljocht, geregeld 70, 80 of 100 km/h-wegen, faak UMS, faak oar asfalt, geregeld 3 of mear partijen oanwêzich, faak in kettingbotsing/kop-sturt (figuer 3).

Troch dizze strategy kin konkludearre wurde dat der problemen binne mei stadich en fluch ferkear (botsingen) en botsingen mei meardere fluchferkearspartijen yn de gemeente Vught. Hjir soene maatregels foar nommen wurde kinne.

Strategy: grutste stiger, ynstânsje: gemeente 's-Hertogenbosch

Yn de gemeente 's-Hertogenbosch falle relatyf mear ûngelokken, wêrtroch't monitoring hast altyd mooglik is. Yn dit gefal is sjoen nei de grutste stiger oer de jierren 2017 oant en mei 2019. Hjiryn sjogge wy 2 clusters: cluster 12 (26 ûngelokken yn 2017, 37 ûngelokken yn 2018, en 38 ûngelokken yn 2019) en cluster 33 (8 ûngelokken yn 2017, 10 ûngelokken yn 2018, en 20 ûngelokken yn 2019).

Figuer 4: Foarbyld hoe't de ûngelokken fan cluster 12 der útsjen kinne soene.

Figuer 5: Foarbyld hoe't de ûngelokken fan cluster 33 der útsjen kinne soene.

Cluster 12 binne benammen ûngelokken oerdei yn deiljocht, faak UMS, altyd oar asfalt, faak droech, faak 1 partij, faak fluch ferkear, alle leeftiden (figuer 4). Cluster 33 binne benammen ûngelokken oerdei yn deiljocht, hiel faak UMS, faak oar asfalt, altyd rein, faak âlderen en/of folwoeksenen belutsen, faak 2 partijen, botsing tusken fluch ferkear en fluch ferkear (figuer 5). Dúdlik is dat der in tanimmend probleem is mei iensidige ûngelokken fan fluch ferkear oerdei en botsingen tusken fluch ferkear en fluch ferkear yn de rein yn de gemeente 's-Hertogenbosch.

Konklúzjes/oanbefellingen

Yn dit paper litte wy it gebrûk fan data science yn kombinaasje mei ferkearsdata sjen. Wy toane oan dat it mooglik is om troch middel fan unsupervised learen in unyk profyl of "DNA" te ûnderskieden yn de data oer ûngelokken en de belutsen partijen. De clusters mei ferlykber DNA kinne troch wegbehearders monitore wurde, om sa opfallende ûntwikkelingen te sinjlearjen, yn kaart te bringen en dêr't it nedich is te yntervenearjen. Troch de clusters op te nimmen yn de software makket VIA Software it mooglik foar wegbehearders om hjir sels mei oan 'e slach te gean. Dit ûndersyk demonstrearret de mearwearde fan

data science yn it ferkearsdomein, mar is mar in earste stap. De ferfolchstappen binne twaledich. Yn it foarste plak sil der sjoen wurde nei it ferbetterjen fan modellen troch bgl. it ûndersykjen fan (nije) modellen, lykas gaussian mixture models (GMMs) of Birch (Zhang, et al. 1996), oare dimensjereduksjes lykas UMAP (McInnes & Healy, 2018), de fierdere spesifikaasje fan dominânsjebepaling en it gebrûk fan mear of minder fariabelen. As twadde sil der sjoen wurde nei de tapasberens fan it algoritme yn de VIA-software, lykas it fergelykjen mei itselde soarte gemeenten of fergelykjen mei gemeenten dy't itselde probleem hawwe.

Literatuer/referinsjes/boarnen

Davenport, T. H., & Patil, D. J. (2012). Data scientist. *Harvard Business Review*, 90(5), 70-76.

Kaptein, M. (2019a, 20 maaie). *Hoe doen computers slimme dingen? Zo werkt machine learning*. Rieplachte fan <https://www.emerce.nl/best-practice/machine-learning-werkt-dankzij-supervised-leren>

Kaptein, M. (2019b, 5 juny). *Hoe doen computers slimme dingen? Ze kunnen ook prima data samenvatten*. Rieplachte fan <https://www.emerce.nl/achtergrond/hoe-doen-computers-slimme-dingen-ze-kunnen-prima-data-samenvatten>

MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, pp. 281-297.

McInnes, L., Healy, J. (2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *ArXiv e-prints* 1802.03426.

Zhang, T., Ramakrishnan, R., & Livny, M. (1996). BIRCH: an efficient data clustering method for very large databases. *ACM Sigmod Record*, 25(2), 103-114.

Noaten

1. <https://www.star-verkeersongevallen.nl>
2. Foarbylden fan meinommen ûngelokskaaimerken: ljochtomstannichheden, dykferhurding, maksimumsnelheid.
3. Foarbylden fan meinommen partijskaaimerken: leeftyd, ferfierwize.
4. Monitoring is it systematysk folgjen fan de ûntwikkelingen oer de tiid hinne fan spesifike groepen mei foarôf bepaalde yndikatoaren. It wurdt brûkt om te sjen wat, wannear en wêrom it spilet (analyse).
5. Data: 2017 oant en mei 2019.