Datapreparatie CAS

CAS maakt voorspellingen omtrent het waar en wanneer van criminaliteit, door patronen te herkennen in een historische dataset (twee jaar gescheiden door wekelijkse peilmomenten) en deze patronen vervolgens door te trekken naar de huidige stand van zaken.

De CAS-dataset wordt gevormd door een tabel waarbij de records kunnen worden geïdentificeerd door locatie (zoals bepaald door X- en Y-coördinaten op een raster; grid_x en grid_y) en periode (zoals bepaald door de periode_id).

Het gebruikte raster is beschikbaar voor heel Nederland. De vierkanten binnen het raster zijn 125x125 meter. Er is een tabel genaamd CAS_D_GRID waarin verschillende kenmerken van de vierkanten (RD-coordinaten van de centroides, basisteam, district en eenheid waartoe ze behoren etc) zijn opgeslagen.

De periode wordt bepaald door de *zaterdag* van een week. Dit komt omdat CAS in het weekend draait, en in de eerste instantie waren de periodes gebaseerd op BVCM-planperiodes (die in de regel op een zaterdag beginnen). De laatste tijd is er wel een roep om de CAS-periodes van maandag tot en met zondag te laten lopen ipv van zaterdag tot en met vrijdag, maar dit zou betekenen dat er niet genoeg tijd is om de meest actuele variant van de kaart te laten draaien. De periodes staan in CAS_D_PERIODE, deze bevat per periode de begindatum/tijd en einddatum/tijd.

CAS draait voor een bepaald gebied, bijvoorbeeld een basisteam of district. Het aantal records wordt bepaald door de grootte van het gebied (dwz het aantal vierkantjes dat hiertoe behoort), maal het aantal periodes dat wordt bekeken. Het aantal periodes is gelijk aan 104 (om de historie te bepalen), plus 1(om de trend door te kunnen trekken naar de komende periode). We hebben het dus over het aantal vierkantjes maal 105.

Een gebied kan maximaal 4 speerpunten opgeven, waar vervolgens de CAS-voorspelling voor wordt gemaakt. Aan het framework dat wordt bepaald door de vierkantjes en periodes worden vervolgens 4 soorten data gehangen:

1. Per speerpunt: informatie over incidenten in de betreffende periode in het betreffende vierkantje (zowel het aantal als of er minstens één incident heeft plaatsgevonden; deze laatste is de target-variabele voor de modelbouw).
2. Per speerpunt informatie over de historische incidenten:
    a. Het aantal incidenten in het vakje, in de periodes voorafgaand aan de periode_id van het record (tot en met 12 weken terug)
    b. Het aantal incidenten in de vakjes die grenzen aan het vakje, in de periodes voorafgaand aan de periode_id van het record (tot en met 12 weken terug)
    c. Het aantal maanden sinds het laatste incident in het vakje op de periode_id van het record.
    d. De lineaire trend (over 4 weken) in het vakje op de periode_id van het record.
    e. De lineaire trend (over 4 weken) in de vakjes die grenzen aan het vakje op de periode_id van het record.
3. Per speerpunt informatie over de nabijheid van woonadressen van bekende verdachten van dat speerpunt (voorwaarde: de personen moeten in het halve jaar voorafgaand aan de periode_id van het record betrokken zijn geweest bij een incident van het speerpunt)
    a. De kleinste (hemelsbrede) afstand van het woonadres van een bekende verdachte tot de centroide van het vakje
    b. Het aantal bekende verdachten dat binnen 500 meter van de centroide van het vakje woont
    c. Het aantal bekende verdachten dat binnen 1000 meter van de centroide van het vakje woont
4. Informatie van het CBS

Framework

Uit CAS_D_GRID worden de vierkanties die tot het gebied behoren geselecteerd. Uit CAS_D_PERIODE worden de hierboven beschreven 105 periodes geselecteerd. Het cartesisch product hiervan vormt het framework voor de rest van de dataset, alles wordt hier uiteindelijke dus aan gehangen. Dat ziet er als volgt uit:

Grid_x	Grid_y	Periode_id
XXX	YYY	250
XXX	YYY	251
XXX	YYY	252
XXX	YYY	253
XXX	YYY	254
…	...	…

Voorbereiding: ruwe data

De basis voor de target variabelen wordt gevormd door incident-informatie. Er is een tabel genaamd CAS_D_SP, waarin de speerpunten voor de CAS-gebieden geregistreerd staan. Dit kunnen enkele MK's zijn, of verzamelingen van MK's.

Met behulp hiervan wordt ruwe incidentinformatie voor het basisteam uit de tabel F_INCIDENT in de BV-laag van BVI gehaald (x- en y-coordinaten, begin- en einddatum en –tijd). Wat betreft de datums worden hier twee variabelen van afgeleid voor latere doeleinden: pleegduur (einddatum/tijd – begindatum-tijd) en geschatte pleegdatum (middelpunt tussen einddatum/tijd en begindatum/tijd).

Vervolgens wordt aan de hand van de RD-coordinaten het vierkantje bepaald waarin het incident heeft plaatsgevonden: grid_x = floor((LOC_X - 12999.5) / 125) + 1; grid_y = intof ((LOC_Y - 303999.5) / 125) + 1

Verder wordt ook de periode waarin het incident heeft plaatsgevonden bepaald. Dit wordt op het moment gedaan door cartesisch te koppelen met CAS_D_PERIODE en alles weg te gooien wat niet tussen de begin-en einddatum van de periode valt, maar dit kan ook met een formule gebeuren en dat is waarschijnlijk sneller.

Per speerpunt wordt een dataset bewaard van de ruwe data zoals hierboven beschreven is.

Dat ziet er als volgt uit:

Loc_x	Loc_y	Begindatum/tijd	Einddatum/tijd	Pleegduur	Geschatte pleegdatum	Grid_x	Grid_y	Periode_id
xxxx	yyyy	Ddmmyyy hh:mm	Ddmmyyy hh:mm	### uur	Ddmmyyy hh:mm	XXX	YYY	250
xxxx	yyyy	Ddmmyyy hh:mm	Ddmmyyy hh:mm	### uur	Ddmmyyy hh:mm	XXX	YYY	251
xxxx	yyyy	Ddmmyyy hh:mm	Ddmmyyy hh:mm	### uur	Ddmmyyy hh:mm	XXX	YYY	252
xxxx	yyyy	Ddmmyyy hh:mm	Ddmmyyy hh:mm	### uur	Ddmmyyy hh:mm	XXX	YYY	253
xxxx	yyyy	Ddmmyyy hh:mm	Ddmmyyy hh:mm	### uur	Ddmmyyy hh:mm	XXX	YYY	254
…	…	…	…	…	…	…	...	…

Targetvariabelen

Vervolgens wordt een en ander geaggregeerd op grid_x, grid_y en periode_id, en aan het framework gehangen. Er worden twee dingen afgeleid:

1. hoeveel incidenten hebben plaatsgevonden in het vakje in de periode; en
2. heeft er überhaupt een vakje plaatsgevonden in het vakje in de periode (0/1)

Voor elk van de speerpunten wordt een dergelijke tussen-tabel gemaakt. Als een basisteam minder dan vier speerpunten heeft worden er tabellen aangemaakt met nullen.

Dat ziet er als volgt uit:

Grid_x	Grid_y	Periode_id	N_INC	Target
XXX	YYY	250	2	1
XXX	YYY	251	0	0
XXX	YYY	252	0	0
XXX	YYY	253	1	1
XXX	YYY	254	0	0
…	...	…	…	…

Historische incidenten

Historische incidenten in het vakje

Dit wordt bepaald door de telling-variabele in de target-tabel te koppelen aan het framework waarbij de periode_id steeds met 1 wordt aangepast. Er wordt op deze manier tot en met 12 weken teruggekeken. Hierdoor zal de tabel er als volgt uit gaan zien:

Grid_x	Grid_y	Periode_id	His1	His2	His3	His4	…	His12
XXX	YYY	250	0	0	0	0	…	0
XXX	YYY	251	2	0	0	0	…	0
XXX	YYY	252	0	2	0	0	…	1
XXX	YYY	253	0	0	2	0	…	0
XXX	YYY	254	1	0	0	2	…	0
…	...	…	…	…	…	…	…	…

Van deze tabel worden ook vier stuks gemaakt, één per speerpunt.

Historische incidenten grenzend aan het vakje

CAS kijkt niet alleen naar historische incidenten binnen een vakje maar ook naar de incidenten in naburige vakjes. Dit wordt gedaan door de incidenten in aangrenzende vakjes op te tellen. De basis hiervoor is de historische incidententabel die in de voorgaande paragraaf is beschreven. Dit gaat op de volgende manier in zijn werk:

Stel voor een bepaalde periode_id zien we de volgende aantallen voor onderstaande locaties (tussen haakjes staan hypothetische gridcoordinaten):

(99, 101) 2	(100, 101) 0	(101, 101) 1
(99, 100) 1	(100, 100) 3	(101, 100) 0
(99, 99) 0	(100, 99) 1	(101, 99) 0

Het is de bedoeling dat per periode_id voor een vakje de naburige vakjes worden opgeteld, in dit geval is het onderhavige vakje nummer (100, 100). Deze moet dus de waarde (2 + 0 + 1 + 1 + 0 + 0 + 1 + 0) = 5 krijgen.

De structuur van de resulterende tabel is gelijk aan die in bovenstaande paragraaf.

Op het moment wordt dit gedaan door alles met alles te koppelen (enige sleutel is de periode_id) en alles weg te gooien wat geen buur is, en vervolgens aggregeren op periode_id, grid_x en grid_y. Een tussentabel waarin de buren staan geregistreerd is waarschijnlijk handiger.

Ook hier wordt één variant per speerpunt berekend.

Tijd sinds laatste incident

Hier wordt het aantal maanden sinds het laatste incident van een speerpunt uitgerekend, per vakje per periode_id. Dit wordt berekend door de ruwe incidenten te koppelen aan het framework, en alleen de gridcoordinaten als sleutel mee te geven. Vervolgens worden de incidenten die (volgens de geschatte pleegdatum) na de begintijd van de onderhavige periode zijn gepleegd uitgefilterd. Daarna vindt een aggregatie plaats om de meest recente (dus: maximum) pleegdatum per vakje uit te rekenen. Vervolgens wordt het verschil in maanden tussen deze meest recente pleegdatum en de datum van het peilmoment uitgerekend.

Er zullen vakjes zijn waar geen incidenten hebben plaatsgevonden, deze worden op een arbitrair hoog getal gezet (41, in dit geval). In de modelbouw wordt deze variabele categorisch gemaakt en als nominaal behandeld, welk getal hier wordt gebruikt is dus niet van belang zolang het maar hoog is.

De tabel ziet er als volgt uit:

Grid_x	Grid_y	Periode_id	TSLI
XXX	YYY	250	41
XXX	YYY	251	41
XXX	YYY	252	0.1
XXX	YYY	253	0.35
XXX	YYY	254	0.6
…	...	…	…

Lineaire trends

Deze worden per speerpunt bepaald, zowel voor de incidenten binnen een vakje als voor de incidenten in de aangrenzende vakjes. Hiervoor worden de meest recenten vier weken gebruikt. De volgende standaard-formule wordt toegepast:

LT(grid_x, grid_y, periode_id)= -(-1.5*His1 - 0.5* His2 + 0.5* His3 + 1.5* His4)/5

Dit wordt uitgerekend voor zowel de vakjes als voor de aangrenzende vakjes.

Beide tabellen zien er als volgt uit:

Grid_x	Grid_y	Periode_id	LT
XXX	YYY	250	0
XXX	YYY	251	0
XXX	YYY	252	0
XXX	YYY	253	0.3
XXX	YYY	254	0.1
…	...	…	…

Verdachtendichtheid

CAS gebruikt verschillende maten voor de mate waarin bekende verdachten van een incidentsoort in de buurt van een vakje wonen. Deze variabelen berekenen we omdat voor sommige delicten geldt dat plegers een voorkeur hebben voor locaties nabij hun woonadres. Dit betekent dat kan gelden dat dit soort informatie een verbetering van de voorspelling kan opleveren bovenop de incident-informatie. Verder beperken we ons tot verdachten die in de zes maanden voorafgaand aan het peilmoment als verdachte voor een speerpunt geregistreerd waren. We rekenen de variabelen voor alle speerpunten van een basisteam uit.

We rekenen de volgende zaken uit: 1) afstand van de centroide van het vakje tot de meest nabije bekende verdachte; 2) aantal bekende verdachten binnen 500 meter van de centroide van het vakje; 3) aantal bekende verdachten binnen 500 meter van de centroide van het vakje.

Dit wordt gedaan door per peilmoment de verdachten woonachtig in het gebied op te vragen (persoons_id, x-coordinaat en y-coordinaat van woonadres) die op het peilmoment maximaal een half jaar geleden als verdachte in verband zijn gebracht van een incident. Deze set wordt dan gekoppeld aan de incidentenset, met als enige sleutel de periode_id.

Vervolgens wordt de hemelsbrede afstand (in km) uitgerekend van de centroide van het vakje tot het woonadres van de verdachte. Daarna wordt uitgerekend of deze afstand minder is dan 500 m, en of deze minder is dan 1000m (allebei indicator-variabelen: 0 of 1). Vervolgens wordt twee keer geaggregeerd: eenmaal op periode_id, grid_x, grid_y en persoonsnummer (om personen die in het halve jaar vaker dan één keer zijn geregistreerd te ontdubbelen), en daarna alleen op periode_id, grid_x en grid_y. Wat wordt uitgerekend is per vakje en per periode het minimum van de afstands-variabele en de som van de indicator-variabelen.

Grid_x	Grid_y	Periode_id	Dist	Susp500	Susp1000
XXX	YYY	250	2.34	0	0
XXX	YYY	251	2.34	0	0
XXX	YYY	252	0.874	0	1
XXX	YYY	253	0.213	1	2
XXX	YYY	254	0.213	1	2
…	...	…	…	…	…

CBS-gegevens

Deze zijn berekend op basis van een dataset uit 2011 (Kerncijfers Postcodegebieden), deze is ondertussen dus aan verversing toe. De set is als volgt geprepareerd.

Ten eerste zijn de waarden in de CBS-tabel vervangen door ordinale klasses (5 stuks gecodeerd van 1 tot en met 5; missing values zijn gecodeerd met -1). De CBS-tabel is op postcodeniveau geregistreerd.

De CBS-gegevens zijn als volgt gekopieerd aan de vakjes.

We hadden vroeger in BinK een locatietabel voor heel Nederland. Daarin stonden voor ieder adres in Nederland onder meer de rijksdriehoekscoordinaten en de postcode. Voor deze adressen zijn de vakjes waarin de adressen vallen uitgerekend (grid_x en grid_y); op de manier zoals beschreven onder ”voorbereiding: ruwe data”). Vervolgens zijn de geklasseerde CBS-gegevens gekoppeld op basis van de postcode.

Vervolgens wordt geaggregeerd op grid_x en grid_y. Voor iedere CBS-waarde wordt de modus uitgerekend binnen een vakje. Dit betekent dat de waarden binnen een vakje worden bepaald door het postcodegebied dat de meeste adressen binnen het vakje heeft.

Een dergelijke tabel ziet er als volgt uit:

Grid_x	Grid_y	Var1	Var2	…
XXX1	YYY1	-1	3	…
XXX1	YYY2	1	-1	…
XXX1	YYY3	3	-1	…
XXX1	YYY4	5	1	…
XXX1	YYY5	2	4	…
…	...	…	…	…

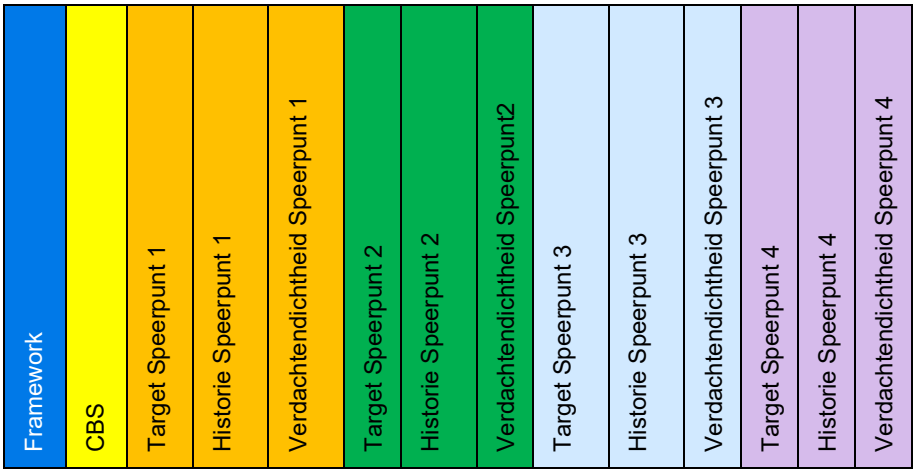
Analysesets

De ontstane datasets worden met elkaar gekoppeld om het uiteindelijke classificatiemodel te maken. Voor de target, historie en verdachtendichtheid gebeurt dit op basis van grid_x, grid_y en periode_id, voor de CBS-gegevens gebeurt dit op basis van grid_x en grid_y.

Er worden drie varianten van de analyseset gemaakt. De verschillen worden bepaald door de pauze die er zit tussen het peilmoment en de voorspelperiode:

1. geen pauze: dit betekent dat de komende week wordt voorspeld met de kenmerken die in het weekend daarvoor zijn verzameld. Dit noemen we de actuele kaart.
2. Pauze van een week: dit betekent niet de komende week, maar die week daarna wordt voorspeld met de kenmerken die in het weekend daarvoor zijn verzameld. Deze week wordt gebruikt om relevante kwalitatieve informatie te verzamelen en een plan te maken. Dit noemen we de een-weekse vroegplanning.
3. Dan is er ook nog een kaart met een pauze van 6 weken. Deze wordt gebruikt om een personeelsplanning te maken voor over 76 weken. Dit noemen we de 6-weekse vroegplanning.

Deze tabellen bestaan uit onderstaande blokken:



Wat natuurlijk ook kan is 3 varianten van de targetvariabelen op te nemen, is wsch met het oog op de opslagruimte handiger.