

Machine Learning ideeën Kadaster

Dit document beschrijft wat verschillende innovatieve plannen die bij voldoende draagvlak uit kunnen groeien tot projecten.

Auteurs: [Rein van 't Veer](#), [Erwin Folmer](#)

[Ruimtelijke matcher voor koppeling BAG-BRT](#)

[Demonstrator voorspellen energielabels](#)

[Object classificatie: vector versus raster](#)

[Automatisch herkennen van fabrikant IMRO plangebied](#)

[Machine learning test data terugmeldingssysteem \(TMS\)](#)

[Bouwjaarcontrole BAG panden](#)

[Matcher voor BGT-BAG panden](#)

[Een betere matcher voor de LKO](#)

[Veranderingsdetectie voor de BRT op basis van 'street view' beelden](#)

[Encoding-detector voor tekstbestanden](#)

[Interessante tijdschriften om in te publiceren](#)

Ruimtelijke matcher voor koppeling BAG-BRT

Elk BAG pand zou een overeenkomend object in de BRT moeten hebben. De linkset tussen BAG en BRT is echter nog niet gerealiseerd, aangezien de BRT 'geaggregeerde' gebouwen bevat, waar de BAG enkelvoudige panden bevat - de relatie is niet 1-op-1. Aangezien ook de bron voor de gebouw/pandgeometrie verschilt, is de koppeling op landelijke schaal tussen de panden uit de BAG en de gebouwen uit de BRT daarmee non-triviaal.

Voorstel: maken van een trainingsset en machine learning model voor het koppelen van BAG panden aan de bijbehorende BRT gebouwen.

Demonstrator voorspellen energielabels

Op basis van experimenten is gebleken dat [decision trees](#) en [random forest](#) modellen vrij nauwkeurig energielabels van panden kan voorspellen. Er wordt daarnaast onderzocht hoe goed een deep learning model is voor het voorspellen van energielabels op basis van open

data. Met een dergelijk model is een live demonstrator te maken die op basis van de BAG een voorspelling kan doen van het pand.

Voorstel: maken van een *data story* waarin mensen op basis van een opgegeven adres zelf het energielabel kunnen laten schatten.

Object classificatie: vector versus raster

In een [eerste studie](#) is gebleken dat met machine learning algoritmes te redeneren valt over geometrieën: met deep learning kunnen geometrieën rechtstreeks geanalyseerd worden, met 'klassieke' algoritmes (decision trees, logistic regression, k-nearest neighbors etc.) kunnen analyses toepassen op afgeleide *signal processing* eigenschappen van geometrieën.

Onbekend is echter hoe beeldclassificatie-algoritmes presteren op gematerialiseerde rasterafbeeldingen van vector-geometrieën. Immers, iedere vector-geometrie is als een raster te visualiseren - dit is wat Google Maps bijvoorbeeld voor ons doet, of een topografische WMS. Wel is duidelijk dat het voor raster-analyse in combinatie met convnets op [bepaalde problemen kan stuiten](#).

Voorstel: herhaal een aantal classificatie-taken voor zowel een geo vector-representatie als voor zijn gematerialiseerde raster-representatie en onderzoek de prestatieverschillen.

Automatisch herkennen van fabrikant IMRO plangebied

De plangebieden voor de linked data van Ruimtelijke Plannen worden getransformeerd vanuit HTML - de HTML parser voor het Kadaster Data Platform. Soms is de vervaardiger hiervan niet bekend.

Voorstel: Een classificatiemodel trainen op de ruwe HTML tekst en de geëxtraheerde geometrie om automatisch een fabrikant-toewijzing te doen.

Machine learning test data terugmeldingssysteem (TMS)

Het Kadaster verzamelt al enige jaren met succes verbetersuggesties via het terugmeldingssysteem. Het is mij (RV) nog niet duidelijk om hoeveel succesvolle terugmeldingen het op dit moment gaat, maar wellicht is er inmiddels genoeg data beschikbaar om te onderzoeken of het mogelijk is om een model te trainen op deze data. Kunnen er automatisch fouten uit de basisregistraties worden gehaald op basis van de teruggemelde gevallen?

Voorstel: trainen van een classificatiemodel dat BRT-, BGT- of BAG-objecten kan aanmerken als akkoord of 'mogelijk incorrect' op basis van geverifieerd juiste records en geverifieerd onjuiste records. Extra gewicht toekennen aan het herkennen van onjuiste objecten. In meer formele termen: lagere precisie maar hogere recall op onjuiste objecten.

Bouwjaarcontrole BAG panden

Er zijn opvallend veel gebouwen in de BAG met een onwaarschijnlijk vroeg bouwjaar. Zie bijvoorbeeld: <http://bag.basisregistraties.overheid.nl/bag/id/pand/0363100012177438> een pand in Amsterdam aan de 17e-eeuwse grachtengordel met een bouwjaar volgens de BAG in 1005. Wikipedia [kent 9 artikelen over 11e-eeuwse architectuur](#), geen hiervan liggen in Noord-Holland. Het leeuwedeel betreft delen van de Sint Servaasbasiliek te Maastricht. Het is duidelijk dat er nogal wat pand-bouwjaaren in de BAG niet kloppen. De evidente fouten zijn eruit te halen op basis van eenvoudige filters (bouwjaar "1005" in Amsterdam), maar de minder evidente fouten zijn er op basis van eenvoudige logica moeilijker uit te halen.

Voorstel: trainen van een regressiemodel dat op basis van een luchtfoto, evt. cyclorama en omgevingsinformatie (omliggende panden) het bouwjaar van een pand voorspelt.

Matcher voor BGT-BAG panden

Niet alle BGT panden hebben een match met een BAG pand. Dat kan komen omdat er geen equivalent in de BAG is (bijvoorbeeld voor windturbines), maar ook omdat er mogelijk een administratieve fout is gemaakt.

Voorstel: trainen van een matcher die voor BGT panden het bijbehorende BAG pand opzoekt, indien dit van toepassing is.

Een betere matcher voor de LKO

De LKO koppeltabel tussen nummeraanduidingen uit de BAG en percelen uit de BRK-DKK wordt nu gevuld op basis van 'business logica': regels die koppeling tussen een nummeraanduiding en een perceel goed- of afkeuren. De koppeling wordt gelegd op basis van zowel administratieve (adres-)gegevens en geometrie.

Voorstel: vullen van de koppeltabel op basis van deep learning

Nog uitzoeken: wat is de [recall](#) van de business rules? Welk percentage van matches missen nog?

Veranderingsdetectie voor de BRT op basis van 'street view' beelden

Momenteel wordt algoritmisch uit luchtfotografie bepaald of er topografische veranderingen hebben plaatsgevonden die verwerkt dienen te worden in de BRT. Pas indien de verandering in luchtfoto's is waargenomen, worden cyclorama's benut om de wijziging te verifiëren.

Voorstel: het bepalen van wijzigingen rechtstreeks uit de cyclorama's op basis van beeldherkenningstechnieken uit de Deep Learning hoek.

Encoding-detector voor tekstbestanden

Er zijn in de wereld veel encoding-standaarden in omloop voor het encoden van Bulgaars, Russisch, Japans etcetera. Soms zijn de verschillen tussen deze encodings subtiel. Dit leidt vaak tot inleesproblemen van tekstbestanden, zoals CSVs.

Voorstel: Een recurrent neurale netwerk kan mogelijk beter onderscheid maken tussen encodings dan bijvoorbeeld echardet (in C), ced (C++) of [chardet](#) (python).

2. SAP - Linked Data ideeën. (opgesteld voor Hans Hendriks)

- * SAP laat zien dat ze linked data (sparql) aankunnen...
- * En daarmee een combinatie te leggen is tussen interne SAP data, en de buitenwereld...
- * De toegevoegde waarde die het SAP BI platform daar vervolgens bovenop kan leggen (machine learning, visualisaties), vooral in de combinaties van datasets (want dat is "nieuw" door linked data).

Jouw case is in ieder geval al bullet 1 en 3...

In die hoek zou je ook een case kunnen bedenken rond kwaliteit: Neem de BRT, alle objecttype in de BRT zijn slecht gevuld (sorry). Met Linked Data hebben we al meermaals aangetoond dat andere open datasets vaak veel meer (en andere) scholen, parkeergarages, moskeeen, etc. bevatten....en dat die prima bruikbaar zijn om de kwaliteit te verbeteren....daar hielden die experimenten vaak op.....maar het is niet helemaal triviaal, dat die andere dataset in zijn compleetheid beter is, in het geval van de parkeergarages hadden we er zelfs drie...maar alle drie bevatte parkeergarages die weer niet in de anderen zaten.....dus er moet nog intelligentie overheen: machine learning achtig....zou toch gaaf zijn als een (SAP) BI achtige toepassing met een voorstel zou kunnen komen: dit zijn de parkeergarages voor de BRT....

(ps. en hier zou je misschien ook nog de luchtfoto's in kunnen betrekken; of cyclomedia beelden; herkennen we de parkeergarage)

Andere case waarbij ik ook de 2e bullet probeer te betrekken....

Even denken wat zit er in onze sap systemen:

- * Klantopdrachten webshop?: linken met cbs demografie (die publiceren we waarschijnlijk vanaf volgende week vanuit PDOK als linked data); om te bepalen of we bv. regionale marketing moeten doen.
- * Opdracht maatwerk: welke gemeentes, met welk collega (politieke kleur) staan garant voor meer opdrachten...oftewel een voorspelling nu met de nieuwe gemeenteraden waar de komende vier jaren de meeste opdrachten gaan ontstaan op basis van onze data uit het verleden.

En wat algemener:

- * Kadaster Open Data BI dashboard...kun je op basis van SAP een soort open BI dashboard maken waar je eenvoudig (linked data / api) datasets kan toevoegen...(in de browser)

Interessante tijdschriften om in te publiceren

[EPJ Data Science](#)

[European Journal of Information Systems](#)

[Journal of Geographical Systems](#)
[Government information quarterly](#)
[Semantic Web journal](#)