Befæstelsesdata som prototype på basis af AI

Ask Holm Carlsen[1] | Mogens Skov[2] | Andrew Flatman[3] | etc.

[1-3] Alle fra Jordobservations-kontoret (JOB) i Styrelsen for Dataforsyning og Effektivisering (SDFE)

Keywords: AI, maskinlæring, frie data, befæstelsesdata, kombinerbarhed, RandomForests, ortofotos, Geodanmark, SDFE LABS.

Abstract

I relation til Styrelsen for Dataforsyning og Effektiviserings (SDFE) fokus og målsætning om brug af kunstig intelligens og at skabe et sammenhængende datagrundlag præsenteres en prototype på et højtopløseligt befæstelseskort.

Befæstelseskortet er skabt ved hjælp af maskinlæringsalgoritmen RandomForests og er i dets nuværende form udelukkende blevet til ved brug af billeder fra GeoDanmarks forårsfotoflyvning. Planen er at inddrage andre datakilder såsom LiDAR eller sommerfotos samt forskellige former for filtre for at styrke algoritmen i dens klassifikation.

Prototypen vil blive præsenteret på SDFE LABS, der er en ny platform til introduktion af og brugerdialog om kandidater til nye og frie datasæt. Feedback ønskes til at videreudvikle prototypen i en retning, som matcher anvendernes ønsker.

1 | Indledning

SDFE har fokus på anvendelse af AI – kunstig intelligens. SDFE ønsker med AI at styrke muligheden for at generere nye og frie data ved smart brug af eksisterende og åbne geodata. Et konkret resultat af denne strategi er udviklingen af en prototype for befæstelsesdata.

Brugen af AI kommer i højere grad til at indgå som et værktøj i SDFEs målsætning om at skabe et sammenhængende og tværoffentligt datagrundlag.

SDFE håber gennem brugerfeedback at kunne tilpasse prototypen, så den kan fungere som et ensartet beregnings- og beslutningsgrundlag på tværs af blandt andet kommuner og forsyninger for bedre investeringer i klimatilpasning, vandforvaltning, anlægsarbejder og beredskab.

GeoDanmarks ortofotos og vektordata danner basis for beregningen og tilblivelsen af befæstelseskortet, der kan støtte et bredt, offentligt behov og øge anvendelse og kombinerbarhed.

2 | Metodebeskrivelse

SDFE har gjort brug af en klassisk maskinlæringsalgoritme ved navn RandomForests (RF). Dette er en såkaldt ”Supervised Learning”-algoritme. Det indebærer, at algoritmen skal trænes med data, såkaldt træningsdata.

Træningsdata giver algoritmen informationer om det, den betragter samt labels, der fortæller algoritmen, hvad det den betragter er eller kaldes. Labels er nødvendige for, at algoritmen kan differentiere mellem træningseksemplerne. Labels, der ofte også kaldes for klasser, er blot fagtermer inden for maskinlæring, der her hentyder til forskellige typer af befæstelse eller til en gruppering af befæstelsestyper. For eksempel er ”asfaltbelagte arealer” en klasse i vores

træningssæt, og en given pixel på en vej er blevet tildelt ”asfaltbelagte arealer” som sin label. Denne pixels befæstelsestype er således asfaltbelagt.

De anvendte data er billeder eller nærmere bestemt pixels i billeder, da det er de enkelte pixels, som algoritmen betragter, og det er ligeledes de enkelte pixels, der er tildelt en label. En pixel i et farvefoto har typisk tre værdier hæftet på sig, nemlig et tal for hhv. rød, grøn og blå (RGB). Billederne er fra GeoDanmarks forårsfotoflyvning, og disse indeholder yderligere et nær-infrarødt (NIR) bånd. Billederne indeholder således fire værdier, RGB-NIR. Dette kaldes også i maskinlæringssammenhænge for et ”Feature Space”, og billederne danner således et firedimensionelt Feature Space. Såfremt f.eks. forårsfoto for to forskellige år var blevet anvendt og lagt oven i hinanden, ville der være tale om et ottedimensionelt Feature Space.

Det er endvidere udbredt praksis at danne nye features eller bånd ved brug af forskellige filtre eller udregninger. Disse kunne f.eks. se på gennemsnitsværdien i en 3x3-matrix omkring hver enkel pixel i et bånd. Dette vil blive til et nyt bånd, et gråtonebillede, og dermed give information om nabolaget for en given pixels. Det kunne også være, at man udregnede et ”Normalized Difference Vegetation Index” (NDVI) og brugte dette som et ekstra bånd. Dette kaldes ”Feature Engineering” og bidrager til at give algoritmen mere information om det, den betragter.

I denne prototype er det valgt udelukkende at bruge de fire bånd fra forårsfotoflyvningen, men det tilstræbes yderligere at udforske brugen af sådanne filtre samt LiDAR-data og supplerende træningseksempler. Dels for at få et mere retvisende kort, dels for at afsøge mulighederne for at klassificere flere befæstelsestyper.

Udover billeddata skal algoritmen også have informationer om, hvad en given træningsdatapixel dækker over. Det træningsdatasæt, der er udarbejdet til denne prototype, var egentlig beregnet til at teste anvendelsen af data fra årets LiDAR-flyvninger, da disse indeholder nye features ved navn ”Extra Bytes”.

Extra Bytes

SDFE’s LiDAR-flyvninger resulterer i indsamling af et punktskydatasæt i formatet LAS. LAS-formatet understøtter en række prædefinerede datafelter på hvert punkt, og nyere versioner af formatet tillader derudover, at man tilføjer brugerdefinerede felter i ”Extra Bytes” på hvert punkt.

I SDFE’s nyere LIDAR-indsamlinger (2018-) indgår to nye datafelter vha. disse Extra Bytes. Felterne beskriver hhv. amplitude og pulsbredde for det ”ekko”, som et punkt er fremkommet ved, og disse værdier kan være interessante ifbm. at bestemme den ramte overflade beskaffenhed.

Der er derfor udelukkende udvalgt træningsdata i og omkring Københavnsområdet, da det er her, der er tilgængelige Extra Bytes, og der er ligeledes et stort befæstet areal med mange forskellige befæstelsestyper. Der er valgt seks klasser: Asfalt, brosten, fliser, grus/sand, jord og lav vegetation.

Indledningsvis er der foretaget en hurtig og grov segmentering af billederne, som er brugt som grundlag for den egentlige generering af træningspixels. Den grove segmentering gav et indtryk af, hvor stor en del af billederne, en given klasse optog, og hvor i billederne disse overfladetyper lå. Dette blev anvendt til at lave træningspixels i et størrelsesforhold, der nogenlunde stemte overens med en given klasses tilstedeværelse i træningsområderne. Herefter blev tilfældelige

punkter genereret for hver klasse i de områder, der stemte overens med klassen fra den grove segmentering.

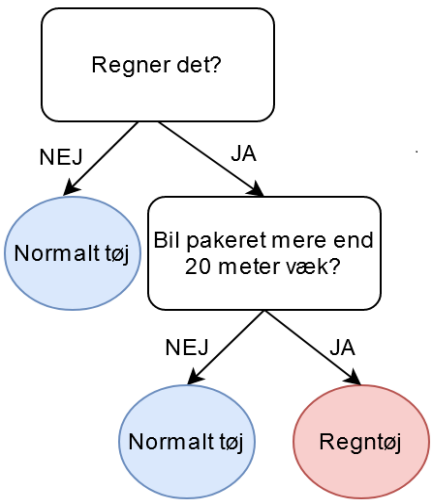
Dermed tilstræbtes en "Stratified Random Sampling"-metode. Dette blev gjort for at tilstræbe træningseksempler, der for hver klasse indeholdt den varians, der er i den virkelige verden. I stedet for at lave en masse punkter i én græsplæne, blev der lavet en masse punkter spredt over et stort antal græsplæner. Derudover giver den grove og indledende klassificering mulighed for at give hver træningspixel en formodet klasse. Dette letter arbejdet med at gennemgå og tildele en klasse til hvert træningseksempel. Sådan at det nu kun er, hvis den formodede klasse ikke stemmer overens med billedet, at man manuelt skal ændre klassen. Det vil være oplagt yderligere at anvende denne metode til genereringen af flere træningseksempler for hele landet, da vi nu har en formodet klasse for enhver pixels i hele Danmark.

Resultatet var omkring 3.500 klassificerede pixels. Det blev besluttet at reklassificere træningseksemplerne til kun at være "befæstet" og "ikke befæstet". Altså kun to klasser eller typer. Dette skyldes dels et ønske om at have et kort, der dækker hele landet, dels en vurdering af, at det ikke vil være muligt udelukkende på baggrund af de fire bånd fra forårsflyvningen at lave en så specifik kortlægning med seks forskellige klasser. Extra Bytes findes som nævnt ikke i hele landet.

At klassificere så specifikke befæstelsestyper vil sandsynligvis kræve: flere træningseksempler, mere Feature Engineering, bånd fra LiDAR og muligvis andre årgange af forårsflyvningen og/eller sommerfotoflyvningen.

Opløsningen på forårs-ortofotos blev ændret til 40 cm for at nedsætte maskinlæringsalgoritmens beregningstid og for at få et produkt, der matcher højdemodellens opløsning.

RandomForests er en relativt gammel, men stadig meget brugt algoritme, som i bund og grund er en kombination eller et ensemble af flere beslutningstræer, såkaldte "Decision Trees". Når metoden anvendes på pixels i et billede, bygges dette træ ved, at algoritmen tager et bånd og finder en pixelværdi, typisk i intervallet 0-255, som splitter træningsdata op i to grupper.

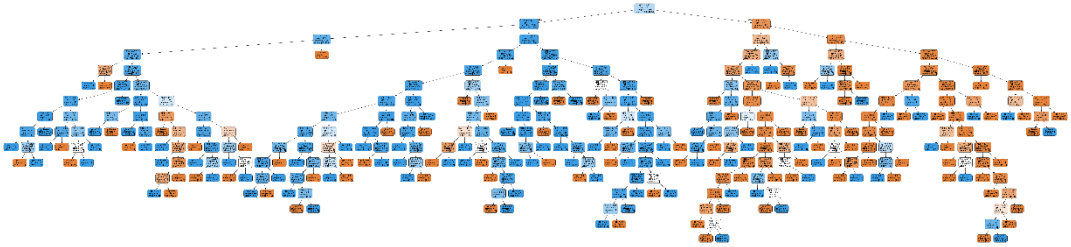


Decision Trees

Et Decision Tree kan anskues som et diagram over handlemåder i en given situation. Lad os sige, at du skal udenfor og du kan se, at det regner. Derfor vælger du at tage en regnfrakke på. Her er vejrsituationen dine data og tøjvalget er dine labels. Hvis man observerer dit tøjvalg og vejrsituationen, så vil man kunne opstille den antagelse, at hvis det regner, så tager du en regnfrakke på. Derfor kan man næste gang, du skal udenfor, og det regner eller ikke regner, komme med et kvalificeret bud på, om du tager en regnfrakke på eller ej. Træet kan selvfølgelig blive mere kompliceret fx med observationer om, hvor meget det regner, eller hvor langt bilen holder fra din hoveddør.

Figur 1: RandomForests er en relativt gammel, men stadig meget brugt algoritme, som i bund og grund er en kombination eller et ensemble af flere beslutningstræer, såkaldte "Decision Trees"

Lad os antage, at dit træningsdatasæt består af ca. 50 % befæstede pixels og 50 % ubefæstede pixels. Algoritmen tager et bånd og finder ud af, hvilket tal der vil være optimalt til at splitte datasættet med for dette bånd. Den opererer efter at maksimere renheden eller minimere urenheden i datasættet, og den værdi, der afføder de mest rene subset, er den værdi, den vælger for dette split. Det kunne være, at man efter splittet endte ud med to sæt, hvor der i det ene sæt var ca. 90 % befæstede træningseksempler og 10 % ubefæstede, og det andet sæt indeholdt således 90 % ubefæstede og 10% befæstede. Herfra udføres der så nye splits på de to nye sæt, og dette afføder så fire renere sæt. Dette gentager sig, indtil det ender ud med helt rene sæt, altså et subset bestående udelukkende af pixels tilhørende den ene eller anden klasse. Et konkret eksempel af kan ses i billedet nedenfor.

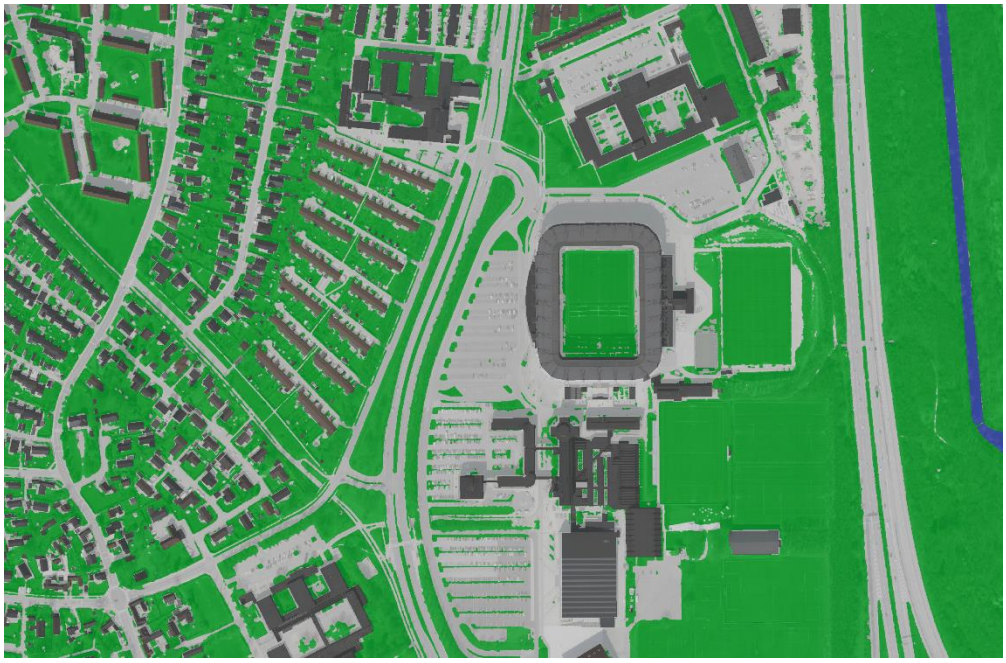


Figur 2: Eksempel på ét ud af hundredevis af træer fra RF-algoritmen. Farverne blå og orange refererer til træningseksemplernes klasse og nuancen antyder, hvor rent hvert subset er. Jo dybere rød eller orange jo renere er hvert subset. Læg mærke til, hvordan slutnoderne altid er den ene elle anden farve.

RandomForests er, som navnet antyder, en skov af Decision Trees, der er trænet på forskellige udgaver af træningssættet. Dette opnås ved den tilfældighed, hvilket også ligger i navnet, der kommer til udtryk i udvælgelsen af træningsdata til hvert træ i skoven. Der bruges kun en delmængde, måske 60 %, af træningssættet, til at træne hvert træ. Derudover kan træningseksemplerne også optræde flere gange i et træningssæt. Disse to egenskaber bevirker, at der ikke er to ens træer i skoven. Dette gør, at RF er meget stærkere end Decision Trees og bedre kan håndtere ny data, der afviger mere eller mindre fra træningsdata. Dette er blot én form af RF. Der findes flere forskellige udgaver af RF.

Outputtet fra RF-algoritmen er et rasterbillede med pixelværdier svarende til de valgte klasser. I dette tilfælde er det et binært output med klasserne: befæstet og ikke-befæstet.

Et af problemerne med RF og mange andre maskinlæringsalgoritmer er, at outputtet kan fremstå ”snavset” eller ”plettet”. For at gøre det lidt pænere er det efterbehandlet det med et filter, der for hver pixel returnerer den klasse, der oftest optræder i en 5x5-matrix omkring en pixel. Herefter er sø- og huspolygoner fra Geodanmark anvendt samt markpolygoner fra Landbrugsstyrelsen. Disse er brændt ind i kortet, fordi klasserne vand og huse ikke optræder i træningssættet. Desuden blev der observeret problemer med nogle marker, der stod uden afgrøder. Dette skyldtes formentlig, at marker heller ikke var repræsenteret i træningsdata, ligesom der også kun optrådte meget lidt bar jord i træningsdata. Det er ikke den langsigtede plan at benytte vektordata til dannelsen af befæstelseskortet, det skal snarere anses som et ”quick-fix” til problemer i prototypen.



Figur 3. Eksempel fra Brøndby af den endelige prototype af befæstelseskortet. Grå er befæstet, grøn er ubefæstet, mørkegrå er bygninger og blå er vand. Grå og grøn er outputtet fra RF, mens vand og bygninger er fra Geodanmark. Eksemplet er illustreret med GeoDanmarks forårsfoto i baggrunden.

Det ser umiddelbart ud til, at der er en del fejlklassificerede pixels i prototypen. Dette er specielt tydeligt i mange forskellige skov- og naturområder, som ofte fremstår "plettede" med mange små eller større områder, der er fejlklassificerede som værende befæstede. Derudover er der ikke brugt en kystlinje til at separere vand fra fastlandet, hvorfor land og vand ikke er adskilt i kortet. Desuden er der problemer i forbindelse med områder, der ligger henlagt i skygge. Dette gør sig gældende både med områder, der bliver fejlklassificeret som værende befæstet og omvendt. Der er også problemer med hustage, der grundet deres placering i forhold til kameraets optagepunkt, afviger fra GeoDanmarks bygningspolygoner og dermed ikke helt dækkes af disse. Specielt orange og røde tage bliver ofte klassificeret som værende ubefæstede arealer. Det samme gør gule bybusser og andre objekter i disse farver. Mange af disse fejl er at forvente med de nævnte begrænsninger, her tænkes specielt på det meget begrænsede træningsdatasæt og feature space.

3 | Udviklingsperspektiver

Som allerede berørt i metodebeskrivelsen er der mange muligheder for videre udvikling. Der findes adskillige metoder og datakilder, der supplerende kan forbedre prototypen. Her tænkes hovedsageligt på brug af Extra Bytes fra LiDAR-data, som dog i skrivende stund ikke findes for hele landet, samt brug af sommerortofoto, der kan bidrage til at beskrive en pixels udvikling hen over sæsonerne og dermed bidrage med værdifuld viden til algoritmens beslutningsproces. Derudover muligheden for at inddrage filtre, der beskriver nabolaget for en pixel. Ligeledes vil valg og udvidelses af træningsdata formentligt have en signifikant indflydelse, specielt på områder der på forskellige måder ikke ligner det nuværende træningsområde.

Denne prototype angiver hverken befæstelsestypen eller graden for, hvor befæstet en given pixel er. Det tidligere datasæt fra Miljøstyrelsens angiver en grad af befæstelse, og det kan tænkes, at dette også ønskes af SDFEs nye befæstelseskort. Umiddelbart ses to forskellige muligheder for at efterkomme et ønske om en befæstelsesgrad. Den ene mulighed ville være at downsample

befæstelseskortet og bruge forholdet mellem befæstede og ubefæstede pixels i de nye større pixels. Det ville være at foretrække at lave den første klassificering på højtopløselige billeder i 10-12,5 cm og derefter downloadsample resultatet til måske 40-50 cm GSD. Dette ville give en 4x4 matrix, som kunne danne grundlag for en grad af befæstelse. Den anden mulighed ville være at klassificere typer af befæstelse og bruge en vurdering af de forskellige typers befæstelsesgrad eller grad af vandgennemtræning. Denne metode vil også nemt kunne sammenkobles med den første metode.

SDFE vil tilstræbe at udvikle prototypen i retning af det produkt, som giver mest mening for vores brugere. I takt med, at vi får kortlagt de forskellige udfordringer, vil vi arbejde på at finde løsninger og producere mere retvisende datasæt, som vi, landsdækkende eller lokalt, vil gøre tilgængelige på SDFE LABS.

Hvordan videreudviklingen kan ske vil afhænge af den feedback, som vi modtager fra vores brugere. Vi forventer, at videreudviklingen vil foregå i perioden november 2020 - marts 2021.

Vi anbefaler, at prototypedata ikke benyttes til forvaltningsmæssige afgørelser, da klassifikationen i datasættet skal betragtes som vejledende.

4 | Præsentation af befæstelsesdata

I SDFE vil vi gerne samarbejde med vores brugere. Det gælder uanset, om det er nye data eller en ny distributionsløsning. Derfor lanceres SDFE LABS med planlagt start i november 2020, som bliver en indgang til at afprøve styrelsens innovative prototyper og give feedback til udviklerne bag.

I SDFE LABS kan du få visualiseret befæstelsesdata for hele landet. Ønsker du selv at visualisere data i egen GIS-klient, kan du med en bruger på dataforsyningen tilgå tjenesten via en URL.

Alternativt kan befæstelsesdata downloades fra FTP i 10x10 km tiles med tilhørende forslag til farveopsætning (.qml). 10x10 km tiles er navngivet efter kvadratnetsnotationen, dvs. at det er cellens SV-hjørne, der er angivet i hele 10-kilometer (EPSG 25832).

Du kan følge denne udvikling på SDFEs hjemmesides kommende nyheder.