A Walk-Through on Machine Learning Techniques for Sentinel Big Data Fusion

Sara Aparício
ESA ESRIN, EOP-Φ

Φ-Week - 12th November 2018
The world’s most comprehensive suite of dedicated EO missions.

20 Tb daily (2017)
Major trends

a) The advent of cloud computing & increase computing power,
b) proliferation of open-access satellite data streams,
c) growing use of machine-learning algorithms

- Supervised Learning:
  - Develop predictive model based on both input and output data
  - Classification
  - Regression

- Unsupervised Learning:
  - Group and interpret data based only on input data
  - Clustering
Sentinel data fusion with Machine Learning Techniques

Multisensor/multitemporal data

Classification

Time

Best technique?
Best band selection?
Motivation I – Merging different datatypes

A promising direction of machine learning in Earth Observation is its pairing with data fusion.

- Sentinel-2
  - Sensitive to soil properties
  - Day & night capability
  - Easy interpretation due to true colour image
  - Gives information on soil features

- Gives information on spectral signatures
This document addresses elements of what needs to be done at European level, some specifically by ESA, to harness the full potential of artificial intelligence to exploit earth observation data and the data supply chain and vice versa.

It captures the recommendations of a community-led AI4EO workshop held at ESA/ESRIN on 27 Mar 2018 (c.f. participant list in Annex 1). The aim of the workshop was to informally assess progress in the development and application of AI techniques to the world of EO and to explore the potential value of a concerted Research and Innovation (R&I) effort on this topic at European level. It is envisaged to be a dynamic report capturing the evolving needs of the community. It will be reviewed and updated in a follow-on workshop at ESA/ESRIN on November 14th, 2018.

**Classification/ Recognition**
- updating land-cover maps

**Detection**
- large scale in automatic basis
- use SAR in machine learning

**Data Fusion**
- Merging diverse EO data
Objectives/Goal

1. To develop a pixel-based classification, reproducible, scalable with a machine learning-based approach of large-area mapping/land cover of high resolution (10m) based on a multi-sensor & multi-temporal approach;

2. To evaluate the additive value of open-access satellite optical and radar variables, processed using cloud computing, to a topographic baseline model.

3. To explore/understanding/address efficiency of Google Earth Engine to effectively execute big data workflows using machine learning techniques on Google Earth Engine (and accuracy) for multi-temporal land use mapping.
METHODOLOGY
Main reasons for selection:

1) Ground truth data available
2) Relatively plain area > since ground range products (GRD) were already terrain corrected
3) A good variety of land cover types to access > agriculture, forest, water and urban areas.

Landcover classes:
- Agriculture
- Coniferous forest
- Mixed Forest
- Grassland
- Bare soil
- Wetland
- Urban fabric
- Water body
Pre-processed on EE
- Apply orbit file
- Border noise removal
- Thermal noise removal
- Radiometric calibration
- Terrain corrected

Filtering data
- Filter data:
  - IW
  - Both orbits
- Time selection
- Clip to ROI

Further processing & data preparation
- Speckle Filter Lee 3x3
- Band ratios creation
  - Mean, std Dev calc.

Filtering data
- Filter data:
  - Cloudy Pixel %<30
  - Time selection
  - Clip to ROI

Data processing
- Cloud masked
- Cirrus masked
- NDVI, NDWI

Stacking

Image
Workflow – input data

Image

Sample regions

Training Data

ML #1
ML #2
ML #3
ML #4
ML #5
ML #6

Validation Data
Workflow – classification

Training Data

- CART
- RF
- SVM
- PER
- NB
- GMO

Classification & Regression trees
Random Forest
Support Vector Machines
Perceptron
Naïve Bayes
GMO Maximum Entropy

Trained

Validation Data

Test model performances:
- Overall Accuracy
- Kappa
- Producers' acc.
- Users's acc.

Classification:

Image

Output

Output

Output

Output

Output

Output
Machine Learning performances – a first overview!
Classifier - CART
Bands – VV $\mu$ (lee)
A closer look

Classifier - CART
Bands - B4 B8 B11
Classifier - RF
Bands – S1 + S2 bands

Classifier - RF
Bands – All except…

Classifier - RF
Bands – All S2 bands

Classifier - RF
Bands – B4 B8 B11

Classifier - RF
Bands - NDVI NDWI
Behaviour to data input
Averaged from best performing models
Improving model performances...

Changing temporal coverage
- updating land-cover maps

Changing training data
In size or coverage

Trying new band selection
Different combinations of S1+S2
Best performing models with S1+S2 variations
Other accuracy parameters for best band & models:

- Random Forest (Tuned)
- CART
- Support Vector Machines
- Random Forest
Further accuracy metrics...

<table>
<thead>
<tr>
<th>Classified</th>
<th>Water body</th>
<th>Wetland</th>
<th>Urban Fabric</th>
<th>Agriculture</th>
<th>Coniferous Forest</th>
<th>Mixed Forest</th>
<th>Grassland</th>
<th>Bare soil</th>
<th>PA</th>
<th>UA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Water body</td>
<td>193</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Wetland</td>
<td>0</td>
<td>29</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.97</td>
<td>0.94</td>
</tr>
<tr>
<td>Urban Fabric</td>
<td>0</td>
<td>0</td>
<td>239</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.99</td>
<td>1.00</td>
</tr>
<tr>
<td>Agriculture</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>428</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0.99</td>
<td>0.99</td>
</tr>
<tr>
<td>Coniferous Forest</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>154</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0.99</td>
<td>0.99</td>
</tr>
<tr>
<td>Mixed Forest</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>157</td>
<td>0</td>
<td>0</td>
<td>0.99</td>
<td>0.98</td>
</tr>
<tr>
<td>Grassland</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>60</td>
<td>0</td>
<td>0.95</td>
<td>1.00</td>
</tr>
<tr>
<td>Bare soil</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>1.00</td>
<td>1.00</td>
</tr>
</tbody>
</table>

**Kappa**

0.9881

**Overall Accuracy**

0.9906
Ground Truth/Reference VS Final model
Limitations...bear in mind that....

So.... Is this then the best model?

- Training data (plays a huge role) affecting:
  - Models performances
  - Smaller size – the more data input...not always is the best
  - Bigger size – more difference in ‘additive power’ but models behaved + similarly
- Tuning hyperparameters of ML are done manually
  - Some models were inserted by default & data was not normalized
What potential...?
Example of straightforward applications...

- Analysis of temporal land use and land cover change...

Attention!
This graph moves 😊
Larger scale + less classes

Mapping water bodies
Larger scale + less classes

Mapping forest cover
Large scale + data input

Elevation
(DEM STRM)
Large scale + data input + classes

CORIN LAND COVER (100m)

\[ S1 + S2 = + \text{accurate} + \text{frequent} + \text{higher resolution} \]
Overall Conclusions

Google Earth Engine

• GEE offers **powerful** capabilities in **handling large volumes** of remote sensing imagery
• GEE contains state-of-art machine learning algorithms achieving high accuracies and excellent tool for rapidly prototype AI applications.
• A big limitation is the need of manually tune the machine learning algorithms

**Added value of fusing Sentinel data**

• The integration of texture and spectral information for pixel-based classification improves classification accuracy (S2 outperforms S1 alone but together detect finer structures).
• Data: the more, the merrier!

**Large scale mapping – further work?**

• Results can be used to calculate/estimate use cover and land change dynamics
• Normalize data and test more combinations!
Thank you for your attention!