
Octavian Dumitru¹, Gottfried Schwarz¹, Mihai Datcu¹, Dongyang Ao, Zhongling Huang, and Mila Stillman²
¹German Aerospace Center, Germany (corneliu.dumitru@dlr.de)
²Technical University of Munich (TUM), Germany

During the last years, much progress has been reached with machine learning algorithms. Among the typical application fields of machine learning are many technical and commercial applications as well as Earth science analyses, where most often indirect and distorted detector data have to be converted to well-calibrated scientific data that are a prerequisite for a correct understanding of the desired physical quantities and their relationships.

However, the provision of sufficient calibrated data is not enough for the testing, training, and routine processing of most machine learning applications. In principle, one also needs a clear strategy for the selection of necessary and useful training data and an easily understandable quality control of the finally desired parameters.

At a first glance, one could guess that this problem could be solved by a careful selection of representative test data covering many typical cases as well as some counterexamples. Then these test data can be used for the training of the internal parameters of a machine learning application. At a second glance, however, many researchers found out that a simple stacking up of plain examples is not the best choice for many scientific applications.

To get improved machine learning results, we concentrated on the analysis of satellite images depicting the Earth's surface under various conditions such as the selected instrument type, spectral bands, and spatial resolution. In our case, such data are routinely provided by the freely accessible European Sentinel satellite products (e.g., Sentinel-1, and Sentinel-2). Our basic work then included investigations of how some additional processing steps – to be linked with the selected training data – can provide better machine learning results.

To this end, we analysed and compared three different approaches to find out machine learning strategies for the joint selection and processing of training data for our Earth observation images:

- One can optimize the training data selection by adapting the data selection to the specific instrument, target, and application characteristics [1].
- As an alternative, one can dynamically generate new training parameters by Generative Adversarial Networks. This is comparable to the role of a sparring partner in boxing [2].
One can also use a hybrid semi-supervised approach for Synthetic Aperture Radar images with limited labelled data. The method is split in: polarimetric scattering classification, topic modelling for scattering labels, unsupervised constraint learning, and supervised label prediction with constraints [3].

We applied these strategies in the ExtremeEarth sea-ice monitoring project (http://earthanalytics.eu/). As a result, we can demonstrate for which application cases these three strategies will provide a promising alternative to a simple conventional selection of available training data.

