EGU24-5791, updated on 23 Jan 2025
https://doi.org/10.5194/egusphere-egu24-5791
EGU General Assembly 2024

Detection of noise in supervised label data: a practical approach in the Amazonas region of Brazil using land use and land cover maps

Maximilian Hell[1] and Melanie Brandmeier[1,2]
[1]Technical University of Applied Sciences Würzburg-Schweinfurt, Würzburg, Germany (maximilian.hell@thws.de)
[2]Esri Deutschland GmbH

The region of Amazônia Legal in Brazil is in constant change due to deforestation and degradation of forest and conversion into arable land to use for farming or cattle ranching. It is important to monitor these changes with respect to global climate change and to aid political decision makers. These changes are best captured and analyzed through openly accessible satellite data, such as the products of ESA's Sentinel Missions. Land use and land cover (LULC) classification is often performed on remotely sensed data through supervised learning algorithms that rely on precise labels to produce accurate results. However, this kind of data is often not available and it is a time consuming task to create such data at the required accuracy level through image interpretation. This can be alleviated by using existing LULC maps from other sources such as the classification maps produced by the *MapBiomas Brasil* project used in our project. These maps are created using Landsat time series data and multiple machine and deep learning models to classify the whole of Brazil into five macro and multiple micro classes. This data has it's own bias and is not correct in all places or highly inaccurate, especially compared to data which is higher in spatial resolution -- as the aforementioned Sentinel data -- and reveals more detail in the land coverage. Thus, it is a critical step to investigate the noise in the label data. There are multiple approaches in the relevant literature to tackle learning with noisy labels, most of these approaches rely on robust loss functions or learned models to identify the noise. We present a novel approach where the satellite imagery is split pixel-wise in the prior given five macro class labels. For each class, a self-organizing map (SOM) is learned to cluster the data in the spectral domain and thus identify representative prototypes of each class. Each class is represented by the same number of prototypes, which overcomes the problem of imbalanced classes. The labels are then checked through neighborhood rules if they belong to their given class or are labeled as unsure or even switch classes otherwise.
In our study, approx. 79.5% of pixels keep their given class, while the rest is reassigned or even discarded. To validate the approach, the results are compared to a manually created validation set and inspected visually for qualitative correctness. The MapBiomas LULC maps reach an overall accuracy of 62.6% in the created validation areas. After relabeling the data with the presented approach, the overall accuracy reached a score of 81.3%, showing a significant increase. This approach is independent of a specifically learned model and only leverages on the relationship between the training data and the given label data --- the Sentinel-2 imagery and the MapBiomas LULC map, respectively.