Active Learning with Constrained Virtual Support Vector Machines for Classification of Earth Observation Data

Christian Geiß[*,1,2], Lorenzo Carlassara[2,3], Ludovico Giorgio Aldo Biagi[3], Hannes Taubenböck[1,4], and Patrick Aravena Pelizari[1]

[1] German Aerospace Center (DLR), German Remote Sensing Data Center (DFD), Münchener Straße 20, 82234 Weßling, Germany

[2] University of Bonn, Department of Geography, Meckenheimer Allee 166, 53115 Bonn, Germany

[3] Department of Civil and Environmental Engineering, Politecnico di Milano, Italy.

[4] University of Würzburg, Institute of Geography and Geology, Chair of Remote Sensing, Am Hubland, 97074 Würzburg, Germany

Developing methods for thematic information extraction from remote sensing imagery and other image data has been one of the major tasks of the scientific remote sensing community over the past few decades. Nowadays, very large deep learning and foundation models are being created, which are often built upon billions of parameters and are very costly in terms of required input data and the energy required for training, among other factors. In contrast, here we follow the idea of learning invariant decision functions for remote sensing image classification using Support Vector Machines (SVM). To achieve this, we generate artificially transformed samples (i.e., virtual samples) from very limited prior knowledge. Labeled samples closest to the separating hyperplane with the maximum margin (i.e., the Support Vectors) are identified by learning an initial SVM model. These Support Vectors are then used to generate virtual samples by perturbing the features to which the model should be invariant. Subsequently, the classification model is trained a second time, using the newly created virtual samples in addition to the SVs, to find a new optimal decision boundary. We implement a self-learning procedure to ultimately prune non-informative virtual samples from a potentially arbitrary invariance generation process, allowing for robust and sparse model solutions. The self-learning strategy jointly considers similarity and margin sampling constraints. In contrast to existing approaches, we extend this concept within the context of active learning. By leveraging the self-learning constraints to simultaneously select the most informative unlabeled samples within an active learning framework, we can achieve excellent model accuracies with very little prior knowledge, while simultaneously maintaining sparse models. Experimental results are obtained from two high-resolution multispectral datasets acquired over the city of Cologne, Germany, and the Hagadera Refugee Camp, Kenya. Comparative model accuracy evaluations highlight the favorable performance properties of the proposed methods, particularly in settings with very few labeled samples. Also, new opportunities for quantum machine learning problems evolve from the proposed models.