remote sensing

MDPI

Article

A Survey of Active Learning for Quantifying Vegetation Traits from Terrestrial Earth Observation Data

Katja Berger [1,*], Juan Pablo Rivera Caicedo [2], Luca Martino [3], Matthias Wocher [1], Tobias Hank [1] and Jochem Verrelst [4]

[1] Department of Geography, Ludwig-Maximilians-Universität München (LMU), Luisenstr. 37, 80333 Munich, Germany; m.wocher@lmu.de (M.W.); tobias.hank@lmu.de (T.H.)
[2] Secretary of Research and Graduate Studies, CONACYT-UAN, 63155 Tepic, Nayarit, Mexico; jprivera@conacyt.mx
[3] Department of Signal Processing, Universidad Rey Juan Carlos (URJC), Mostoles, 28933 Madrid, Spain; luca.martino@urjc.es
[4] Image Processing Laboratory (IPL), Parc Científic, Universitat de València, Paterna, 46980 València, Spain; jochem.verrelst@uv.es
* Correspondence: katja.berger@lmu.de

Abstract: The current exponential increase of spatiotemporally explicit data streams from satellite-based Earth observation missions offers promising opportunities for global vegetation monitoring. Intelligent sampling through active learning (AL) heuristics provides a pathway for fast inference of essential vegetation variables by means of hybrid retrieval approaches, i.e., machine learning regression algorithms trained by radiative transfer model (RTM) simulations. In this study we summarize AL theory and perform a brief systematic literature survey about AL heuristics used in the context of Earth observation regression problems over terrestrial targets. Across all relevant studies it appeared that: (i) retrieval accuracy of AL-optimized training data sets outperformed models trained over large randomly sampled data sets, and (ii) Euclidean distance-based (EBD) diversity method tends to be the most efficient AL technique in terms of accuracy and computational demand. Additionally, a case study is presented based on experimental data employing both uncertainty and diversity AL criteria. Hereby, a a simulated training data base by the PROSAIL-PRO canopy RTM is used to demonstrate the benefit of AL techniques for the estimation of total leaf carotenoid content (C_{xc}) and leaf water content (C_w). Gaussian process regression (GPR) was incorporated to minimize and optimize the training data set with AL. Training the GPR algorithm on optimally AL-based sampled data sets led to improved variable retrievals compared to training on full data pools, which is further demonstrated on a mapping example. From these findings we can recommend the use of AL-based sub-sampling procedures to select the most informative samples out of large training data pools. This will not only optimize regression accuracy due to exclusion of redundant information, but also speed up processing time and reduce final model size of kernel-based machine learning regression algorithms, such as GPR. With this study we want to encourage further testing and implementation of AL sampling methods for hybrid retrieval workflows. AL can contribute to the solution of regression problems within the framework of operational vegetation monitoring using satellite imaging spectroscopy data, and may strongly facilitate data processing for cloud-computing platforms.

Keywords: Gaussian process regression; EnMAP; hyperspectral; query strategies; optimal experimental design

Citation: Berger, K.; Rivera Caicedo, J.P.; Martino, L.; Wocher, M.; Hank, T.; Verrelst, J. A Survey of Active Learning for Quantifying Vegetation Traits from Terrestrial Earth Observation Data. *Remote Sens.* **2021**, *13*, 287. https://doi.org/10.3390/rs13020287

Received: 6 December 2020
Accepted: 12 January 2021
Published: 15 January 2021

1. Introduction

In view of the unprecedented data availability delivered by recently launched and planned optical satellite missions, agricultural and other ecosystem applications will benefit largely from the provided up-to-date information regarding vegetation status and