Geophysical Research Abstracts
Vol. 18, EGU2016-5753, 2016
EGU General Assembly 2016

Improving permafrost distribution modelling using feature selection algorithms

Nicola Deluigi, Christophe Lambiel, and Mikhail Kanevski
Institute of Earth Surface Dynamics (IDYST), University of Lausanne, Lausanne, Switzerland (nicola.deluigi@unil.ch)

The availability of an increasing number of spatial data on the occurrence of mountain permafrost allows the employment of machine learning (ML) classification algorithms for modelling the distribution of the phenomenon. One of the major problems when dealing with high-dimensional dataset is the number of input features (variables) involved. Application of ML classification algorithms to this large number of variables leads to the risk of overfitting, with the consequence of a poor generalization/prediction. For this reason, applying feature selection (FS) techniques helps simplifying the amount of factors required and improves the knowledge on adopted features and their relation with the studied phenomenon. Moreover, taking away irrelevant or redundant variables from the dataset effectively improves the quality of the ML prediction.

This research deals with a comparative analysis of permafrost distribution models supported by FS variable importance assessment. The input dataset (dimension = 20-25, 10 m spatial resolution) was constructed using landcover maps, climate data and DEM derived variables (altitude, aspect, slope, terrain curvature, solar radiation, etc.). It was completed with permafrost evidences (geophysical and thermal data and rock glacier inventories) that serve as training permafrost data. Used FS algorithms informed about variables that appeared less statistically important for permafrost presence/absence. Three different algorithms were compared: Information Gain (IG), Correlation-based Feature Selection (CFS) and Random Forest (RF). IG is a filter technique that evaluates the worth of a predictor by measuring the information gain with respect to the permafrost presence/absence. Conversely, CFS is a wrapper technique that evaluates the worth of a subset of predictors by considering the individual predictive ability of each variable along with the degree of redundancy between them. Finally, RF is a ML algorithm that performs FS as part of its overall operation. It operates by constructing a large collection of decorrelated classification trees, and then predicts the permafrost occurrence through a majority vote. With the so-called out-of-bag (OOB) error estimate, the classification of permafrost data can be validated as well as the contribution of each predictor can be assessed.

The performances of compared permafrost distribution models (computed on independent testing sets) increased with the application of FS algorithms on the original dataset and irrelevant or redundant variables were removed. As a consequence, the process provided faster and more cost-effective predictors and a better understanding of the underlying structures residing in permafrost data. Our work demonstrates the usefulness of a feature selection step prior to applying a machine learning algorithm. In fact, permafrost predictors could be ranked not only based on their heuristic and subjective importance (expert knowledge), but also based on their statistical relevance in relation of the permafrost distribution.