EGU22-6482
https://doi.org/10.5194/egusphere-egu22-6482
EGU General Assembly 2022
© Author(s) 2022. This work is distributed under
the Creative Commons Attribution 4.0 License.

On Machine Learning from Environmental Data

Mikhail Kanevski
University of Lausanne, Institute of Earth Surface Dynamics, Lausanne, Switzerland (mikhail.kanevski@unil.ch)

Geo- and environmental sciences produce a wide variety and numerous data which are extensively used both in fundamental research on Earth processes and in important real-life decision-making. Most natural phenomena are non-linear, multivariate, highly variable and correlated at many spatio-temporal scales. Analysis and treatment of such complex data and their integration/assimilation with science-based models is a difficult problem. Contemporary machine learning (ML) proposes an important set of effective approaches to address this problem at all phases of the study.

Nowadays, Geosciences are one of the major customers of ML ideas and technologies. To a large degree, it is connected to the local and global challenges facing humanity: sustainable development, biodiversity, social and natural hazards and risks, meteo- and climate forecasting, remote sensing Earth observation, etc. Despite being theoretically a universal modelling tool, the success of ML applications significantly depends on the problem formulation, quantity and quality of data and objectives of the study. Therefore, an efficient application of ML demands a good knowledge of the phenomena under study and a profound understanding of learning algorithms which can be achieved in close collaboration between experts in the corresponding domains.

In the current presentation, the study of geo- and environmental data using different machine learning algorithms is reviewed. A problem-oriented approach, which follows a generic data-driven methodology, is applied. The methodology consists of several important steps, in particular, optimization of monitoring and data collection, comprehensive exploratory data analysis and visualization, feature engineering and relevant variables selection, modelling with careful validation and testing, explanation and communication of the results. Advanced experimentation with data by using different supervised and unsupervised ML algorithms helps in better understanding of original data and constructed input feature space, obtaining more reliable and robust results and making intelligent decisions. The presentation is accompanied by simulated and real data case studies from natural hazards (avalanches, forest fires, landslides), environmental risks (pollution) and renewable energy assessment. In conclusion, some general remarks and future perspectives are discussed.