Geophysical Research Abstracts
Vol. 21, EGU2019-4886, 2019
EGU General Assembly 2019

Feature selection using simple and efficient machine learning models. Case studies and software tools

Federico Amato, Fabian Guignard, and Mikhail Kanevski
University of Lausanne, Institute of Earth Surface Dynamics, Lausanne, Switzerland (federico.amato@unil.ch)

Feature selection (FS) of relevant variables from the original input space is a crucial issue in Machine Learning research, especially in environmental data mining. Besides reducing the dimensionality, removing irrelevant information, increasing learning accuracy and improving the interpretability of the results, feature selection is also often used to optimize data collection, as it identifies which kind of data are more important to gather. There are three fundamental classes of FS – filters, wrappers and embedding [1]. FS can be considered either as a pre-processing step or as a dynamic process integrated into the modelling procedure, which helps to reduce the prediction error and uncertainty.

The present research deals with an experimental study of FS using both simulated data and monthly wind speed data in Switzerland for the year 2008 collected by the MeteoSwiss meteorological network (118 stations). The raw data were embedded into a thirteen-dimensional input feature space generated from the Digital Elevation Model [2]. To identify the relevant features to be used in the prediction of wind speed, two efficient and fast machine learning models, namely Extreme Learning Machine (ELM) [3] and General Regression Neural Network (GRNN) [4], have been implemented. An exhaustive search over the thirteen-dimensional space, giving rise to the 8191 possible models, has been performed with the two algorithms for the twelve monthly datasets. Best models were selected according to the smallest root mean squared error. Subsequently, the obtained results were independently tested by applying a Random Forest model and an Anisotropic General Regression Neural Network (AGRNN). The results obtained on the wind speed dataset confirm the idea that the best subset of features is changing according to the studied month/season, which agrees with a physical understanding of the phenomenon. The future research will consider an extension of the approach to higher dimensional space and forward and backward FS techniques.

The models were implemented in Python. The newly developed AGRNN code is compatible with the most widespread Python libraries, such as Pandas, Numpy and Scipy, and has complete integration with Scikit-learn, the most used Python library in machine learning.

References

[1] I. Guyon and A. Elisseeff. An Introduction to Variable and Feature Selection. Journal of Machine Learning Research 3 (2003) 1157-1182, 2003
[2] S. Robert, L. Foresti, M. Kanevski. Spatial prediction of monthly wind speeds in complex terrain with adaptive general regression neural networks. International Journal of Climatology 33 (7), 1793-1804, 2013
[3] M. Leuenberger, M. Kanevski. Extreme Learning Machines for spatial environmental data. Computers & Geosciences 85, 64-73, 2015
[4] M. Kanevski, A. Pozdnoukhov, V. Timonin. Machine Learning for Spatial Environmental Data: theory, applications and software. EPFL Press, 2009