Geophysical Research Abstracts
Vol. 21, EGU2019-2760, 2019
EGU General Assembly 2019

Feature Selection Inspired by Geospatial Data Analysis

Jean Golay, Mohamed Laib, and Mikhail Kanevski

University of Lausanne, Institute of Earth Surface Dynamics, Faculty of Geosciences and Environment, Lausanne, Switzerland (jean.golay@unil.ch)

Continuous improvements in technology and software have caused the volume of data to increase dramatically over the past few decades. This unprecedented surge in the amount of data available results in large high-dimensional datasets containing a lot of unnecessary features. Here, “feature” is used as a synonym for “input variable”. Such features can be redundant if they do not carry new information compared to those previously retained. They can also be irrelevant if they are not related to the learning task to be completed (e.g. regression or classification). As a result, these features contribute to the emergence of four major issues that seriously challenge the extraction of new knowledge from data:

1. The increasing difficulty of interpreting the results of data analysis and modelling.
2. The reduction in the accuracy of learning algorithms.
3. The growing complexity of data visualization.
4. The continuous need for better computer performance.

The goal of feature selection is to mitigate these issues by identifying and selecting features that are neither redundant nor irrelevant (i.e. relevant). This can be achieved using supervised or unsupervised techniques. The former consider an output variable that guides the selection process, while the latter do not.

This work presents supervised and unsupervised feature selection techniques based on indices and concepts that have originated from geospatial data analysis. R implementations are available through two packages that can be downloaded from the CRAN repository, namely IDmining [1] and SFtools [2]. Besides, many real-world case studies are considered. They are related to environmental pollution and renewable resources.

References
[1] Golay J. and Laib M. (2018). IDmining: Intrinsic Dimension for Data Mining. R package version 1.0.5. https://CRAN.R-project.org/package=IDmining
[2] Laib M. and Kanevski M. (2017). SFtools: Space-Filling-Based Tools for Data Mining. R package version 0.1.0. https://CRAN.R-project.org/package=SFtools