Geophysical Research Abstracts
Vol. 20, EGU2018-13856, 2018
EGU General Assembly 2018

IDmining: An R Package for Mining Large Datasets with the Morisita Estimator of Intrinsic Dimension

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 have caused the volume of data to increase dramatically over the past few decades. Consequently, the size of datasets has been increasing rapidly across many scientific disciplines, which poses great challenges in terms of information and knowledge extraction. Traditional issues, such as the multi-scale variability of data or the presence of extremes, noise and outliers, become harder to handle. Besides, new pitfalls must be addressed. They mainly follow from the empty space phenomenon and the increase in computational efficiency requirements.

IDmining is an R package that deals with the above-mentioned problems. It contains algorithms that were developed according to the following idea: data points often reside on a non-linear manifold of much lower dimension than that of the space in which they are embedded. Traditionally, the dimension of such a manifold is called Intrinsic Dimension (ID). However, the proposed package considers the more general case where the data ID can be a non-integer value, and it relies on the Morisita estimator (a fractal-based estimator) for the ID computation. Thus, IDmining uses the (possibly non-integer) ID of data to perform data mining tasks, such as spatial autocorrelation detection and quantification as well as supervised and unsupervised feature selection.

In the presented work, the main functions of IDmining are explained and applied to real-world datasets (hyperspectral images, environmental pollution, renewable energy). The goal is to show the ease of use and the efficiency of the functions. Finally, future developments are discussed with regard to the use of ID in challenging data mining tasks.

References

J. Golay, M. Laib, IDmining: Intrinsic Dimension for Data Mining, R package, 2016 (available from the CRAN repository: https://CRAN.R-project.org/package=IDmining).

J. Golay, M. Kanevski, C. D. Vega Orozco, M. Leuenberger, The multipoint Morisita index for the analysis of spatial patterns, Physica A: Statistical Mechanics and its Applications 406:191-202, 2014.

J. Golay, M. Kanevski, A new estimator of intrinsic dimension based on the multipoint Morisita index, Pattern Recognition 48(12):4070-4081, 2015.

J. Golay, M. Kanevski, Unsupervised feature selection based on the Morisita estimator of intrinsic dimension, Knowledge-Based Systems 135:125-134, 2017.

J. Golay, M. Leuenberger, M. Kanevski, Feature selection for regression problems based on the Morisita estimator of intrinsic dimension, Pattern Recognition 70:126-138, 2017.