Geophysical Research Abstracts
Vol. 18, EGU2016-18163, 2016
EGU General Assembly 2016

Unsupervised Feature Selection Based on the Morisita Index

Jean Golay and Mikhail Kanevski

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

Recent breakthroughs in technology have radically improved our ability to collect and store data. As a consequence, the size of datasets has been increasing rapidly both in terms of number of variables (or features) and number of instances. Since the mechanism of many phenomena is not well known, too many variables are sampled. A lot of them are redundant and contribute to the emergence of three major challenges in data mining: (1) the complexity of result interpretation, (2) the necessity to develop new methods and tools for data processing, (3) the possible reduction in the accuracy of learning algorithms because of the curse of dimensionality.

This research deals with a new algorithm for selecting the smallest subset of features conveying all the information of a dataset (i.e. an algorithm for removing redundant features). It is a new version of the Fractal Dimensionality Reduction (FDR) algorithm [1] and it relies on two ideas:

(a) In general, data lie on non-linear manifolds of much lower dimension than that of the spaces where they are embedded.
(b) The situation describes in (a) is partly due to redundant variables, since they do not contribute to increasing the dimension of manifolds, called Intrinsic Dimension (ID).

The suggested algorithm implements these ideas by selecting only the variables influencing the data ID. Unlike the FDR algorithm, it resorts to a recently introduced ID estimator [2] based on the Morisita index of clustering and to a sequential forward search strategy. Consequently, in addition to its ability to capture non-linear dependences, it can deal with large datasets and its implementation is straightforward in any programming environment.

Many real world case studies are considered. They are related to environmental pollution and renewable resources.

References

[1] C. Traina Jr., A.J.M. Traina, L. Wu, C. Faloutsos, Fast feature selection using fractal dimension, in: Proceedings of the XV Brazilian Symposium on Databases, SBBD, pp. 158–171, 2000.
[2] J. Golay, M. Kanevski, A new estimator of intrinsic dimension based on the multipoint Morisita index, Pattern Recognition 48(12), pp. 4070-4081, 2015.