Geophysical Research Abstracts
Vol. 20, EGU2018-2379-2, 2018
EGU General Assembly 2018

Information-theoretic-based input variable selection for hydrology and water resources

John Quilty (1), Jan Adamowski (1), Bahaa Khalil (1), and Maheswaran Rathinasamy (2)

(1) Department of Bioresource Engineering, McGill University, Montreal, Canada, (2) MVGR College of Engineering, Vizianagaram, India

Nonlinear input variable selection (IVS) has become a growing interest within the hydrological and water resources domains. Nonlinear IVS methods (e.g. Partial Mutual Information Selection (PMIS)) have been shown to outperform linear methods (e.g. Pearson's Partial Linear Correlation Input Selection (PCIS)) when specifying inputs (explanatory variables) for data-driven hydrological and water resources models. Many of these nonlinear methods (e.g. PMIS) rely on information-theoretic concepts such as entropy, mutual information (MI), and conditional MI (CMI) to identify relevant, redundant, and extraneous inputs given a target process and a set of explanatory variables.

Our discussion begins by comparing several information-theoretic-based nonlinear IVS methods that utilize CMI: PMIS, kernel density estimation (KDE), and k nearest-neighbours (KNN), including our new method based on Edgeworth Approximations (EA), with a focus on their assumptions and parametric requirements. An important conclusion is drawn that many of the earlier CMI-based approaches (PMIS, KDE, and KNN) are more computationally expensive and require careful parametric optimization that greatly affects their performance, while the new EA method provides a simple computationally inexpensive and parameter-free formulation. We further develop the EA method to include an assessment of the input variable selection uncertainty using the bootstrap and rank statistics, resulting in the bootstrap rank-ordered CMI (broCMI) approach. Afterwards, the EA and broCMI approaches are benchmarked against the existing (PCIS, PMIS, KDE, and KNN) IVS approaches for different modeling scenarios: 1) synthetic linear and nonlinear datasets; 2) a partially-synthetic rainfall-runoff example; and 3) a real-world urban water demand forecasting problem. Selection accuracy is used for evaluating the IVS performance for the synthetic problems while root mean square error and Nash-Sutcliffe Efficiency Index are used to judge the forecasting performance for the real-world example.

The results indicate the general superiority of the nonlinear IVS methods, especially for the nonlinear scenarios. The EA method is shown to provide similar performance to PMIS, KDE, and KNN at a reduced computational cost while broCMI is shown to provide the best overall performance against all competitors. We conclude our discussion with a set of future research directions that includes using information-theoretic IVS methods for the exploratory analysis of time series properties (e.g., trend, periodicities, etc.) and as an integral component of a stochastic data-driven forecasting framework, tools that may be useful for a wide number of hydrological and water resources problems.