Geophysical Research Abstracts
Vol. 14, EGU2012-7934-1, 2012
EGU General Assembly 2012
© Author(s) 2012

Information flows in the process of hypothesis testing, insights from Solomonoff inductive inference

S.V. Weijs (1), N. van de Giesen (2), and M.B. Parlange (1)

(1) School of Architecture, Civil and Environmental Engineering, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland (steven.weijs@epfl.ch), (2) Section Water Resources, Delft University of Technology, Delft, The Netherlands

Hydrology is a science mostly dealing with making predictions about complex systems that are only partly observable. As a consequence, hydrology mainly deals with induction, i.e. the finding of general patterns and theories from observations. This is often done by sequentially forming plausible hypotheses, test them, and come up with other and improved hypotheses that might explain the data. In this presentation, we look at the information flows within this process and indicate possible problems with mixing prior and posterior information that could occur.

This is remediated by a second approach, multi-model inference, where multiple hypotheses are simultaneously tested. A slightly complicating factor in hydrology is that all models are wrong. This leads to the zero prior problem, where a purely Bayesian approach is doomed to be forever stuck in wrong models, no matter how much information is gained from observations about models that were not in the prior.

One part of the solution is to only consider hypotheses that make statements in probabilistic terms. These hypotheses are not being proven wrong by the data, but just more or less probable. This means that the error model should be given as part of the hypothesis and the likelihood completely defined by hypothesis and data.

The solution of the zero prior problem lies in a universal prior, that gives a prior probability to all computable hypotheses, before seeing any data. Solomonoff's universal prior can be seen as a quantitative implementation of Occam's razor, based in algorithmic information theory. The prior probability of each model is inversely related to its complexity, as measured by its program length on a reference computer. The final predictions are a Bayesian mixture of outputs from all possible models.

In Solomonoff induction, predictions are central, and hypotheses just a means to achieve good predictions. In the idealized case of infinite computational resources, the concept of hypothesis becomes empty, because anything computable is considered a hypothesis. Although necessarily incomputable in practice, Solomonoff induction is useful to keep in mind as a golden standard; both to guide human efforts of identifying processes through our well-developed pattern recognition capabilities, but also to find approximations, yielding near-optimal computable automated methods for pattern discovery in the increasingly large available hydrological data sets (cf. the 4th paradigm).