Geophysical Research Abstracts, Vol. 9, 07037, 2007
SRef-ID: 1607-7962/gra/EGU2007-A-07037

Comparing machine learning approaches in estimating model uncertainty of hydrological conceptual models

D.L. Shrestha, D.P. Solomatine
UNESCO-IHE Institute for Water Education ({d.shrestha, d.solomatine}@unesco-ihe.org)

This study presents a methodology for assessing total model uncertainty of hydrological conceptual model using machine learning techniques. Historical model errors are assumed to be indicator of total model uncertainty. The model uncertainty is measured in the form of the model errors quantiles or prediction intervals and such expression of uncertainty comprises all sources of uncertainty (e.g. model structure, model parameters, input data and output data etc.) without attempting to separate the contribution given by the different sources of uncertainties.

The idea of estimating model prediction uncertainty using data-driven models was presented by Shrestha and Solomatine earlier (published in 2006 in the Neural Networks Journal, at the Int. Conference on Hydroinformatics, and at EGU-2006). This presentation focuses on a wider framework of using machine learning to estimate model uncertainty.

The main idea is to partition the model input data into different clusters where the data belonging to the same cluster have similar values of model errors (or at least mean and variance). This is done by building a data matrix by combining (some of the) historical model inputs and corresponding model errors; partitioning this calibration data using clustering techniques such as crisp cluster or fuzzy clustering. Prediction interval is constructed for each cluster by constructing empirical distribution of the model errors. The estimation of prediction intervals for input data can be done by i) “eager” supervised classification, ii) instance-based (prototype) learning, and iii) supervised regression method.

In classification method classifiers are built from the cluster labels and input data matrix and this classifier classifies the unseen input data. Estimation of prediction intervals for the given input data consists of query of lookup table between cluster

labels and prediction intervals. In instance-based learning instead of building classifier, distance function is used to identify the cluster for the given validation input data, and represent it by its prototype (typically, its center). In regression method, prediction intervals to each input in calibration data set are computed. Two regression models are trained from the input data matrix and computed prediction intervals. The trained regression models are applied to estimate prediction intervals in the unseen validation data set.

The study compares the different learning methods to compute prediction intervals. The methodology was employed to estimate uncertainty of simulated river flows by a conceptual hydrological model to the case study of Brue catchment in United Kingdom.