

Developing Artificial Neural Networks for Water Quality Modelling and Prediction

Robert James May

PhD Thesis Abstract

School of Civil, Environmental & Mining Engineering, The University of Adelaide

October 2009

Modelling water quality within complex, man-made and natural environmental systems can represent a challenge to practitioners. Many conventional modelling tools are not capable of representing the complexities of physical and chemical processes often observed in these systems. Consequently there has been a great deal of interest in the application of computational intelligence techniques, such as artificial neural networks (ANNs). However, "black-box" approaches, such as ANN modelling, are often criticised due to a perceived lack of transparency in the model development methodology. This research has therefore focussed on improving the tools and techniques that are used in the development of ANN models for water quality prediction and forecasting.

The body of research presented in this thesis is described by several peer reviewed articles. These articles describe the theoretical basis and practical context for the ANN model development techniques that have been proposed and applied as a part of this research. Specifically, the ANN development framework has been further enhanced by this research through the development of novel approaches to perform two key tasks: input variable selection (IVS) and data splitting.

The IVS problem is to select variables as ANN inputs from a number of potential candidates, so as to minimise the number of inputs, but maximise the predictive performance of the model. A forward-selection approach for IVS has been examined that is based on partial mutual information (PMI), which can identify an optimal set of variables to use as inputs to ANN models, given a set of candidate variables. Of particular concern is that the use of MI in place of the more traditionally used correlation, provides a more appropriate basis for the selection of inputs based on non-linear relevance. Moreover, the accuracy of MI estimates is necessary to determine critical values of MI, since this forms the basis for of the termination criterion that stops the forward selection process.

Novel termination criteria were developed that alternatively determine the optimum number of candidate input variables. In comparison to the existing approach, which is based on a computationally expensive, yet potentially inaccurate bootstrap approach, the alternative criteria were found to both reduce the computational requirements and increase selection accuracy of the PMI-based IVS approach, resulting in a much improved algorithm.

Data splitting is an essential part of ANN model development, as the available modelling data must be partitioned into subsets for training, testing and validation. Depending on the data splitting method employed, the data split can have a significant effect on model performance, or reduce confidence in performance assessment. A popular method based on clustering of the self-organizing map (SOM) was examined. The approach was found to be sensitive to SOM size and the manner in which samples are drawn from within the SOM units. However, despite an optimal number of partitions, the SOM can generate partitions that are non-uniformly distributed, and which differ in size and shape. Although conventional

rules to increase the sampling rate within larger clusters can reduce variance, the remaining variance can still be significant.

A hybrid algorithm called SOMPLEX was developed, which combines clustering on the soM, and the DUPLEX algorithm used to perform intra-cluster sampling. DUPLEX is a fully deterministic algorithm that generates a representative sample, regardless of the size or distribution of data within a SOM cluster. For several example applications to predicting water quality, SOMPLEX was found to generate representative data for training, testing and validation, with no variation. The hybrid SOMPLEX approach combines the strengths of the two individual data splitting algorithms, in that the clustering on the SOM reduces the operational complexity, and the DUPLEX sampling improves on random sampling of SOM units to reduce sample variability and increase the representativeness of datasets generated.

In terms of the overall ANN development framework, the outcomes of this research have been an increased understanding of how to best implement ANN techniques, and an appreciation for their place within the context of a water quality modelling toolkit, which comprises both conventional and non-conventional modelling approaches. It was also observed that although the ANN modelling paradigm is quite powerful, it is not without limitations. Many of the limitations and problems encountered with ANN model development are more indicative of the application, rather than the modelling approach itself.