Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Hydrology 301 (2005) 75–92
Journal of Hydrology
www.elsevier.com/locate/jhydrol

Input determination for neural network models in water resources applications. Part 1—background and methodology

Gavin J. Bowden[a], Graeme C. Dandy[b], Holger R. Maier[b,*]

[a]*Division of Engineering and Applied Sciences, Harvard University, 29 Oxford Street, Pierce Halt 110J, Cambridge, MA 02138, USA*
[b]*Centre for Applied Modelling in Water Engineering, School of Civil and Environmental Engineering, The University of Adelaide, Adelaide 5005, Australia*

Received 29 April 2003; revised 28 May 2004; accepted 15 June 2004

Abstract

The use of artificial neural network (ANN) models in water resources applications has grown considerably over the last decade. However, an important step in the ANN modelling methodology that has received little attention is the selection of appropriate model inputs. This article is the first in a two-part series published in this issue and addresses the lack of a suitable input determination methodology for ANN models in water resources applications. The current state of input determination is reviewed and two input determination methodologies are presented. The first method is a *model-free* approach, which utilises a measure of the mutual information criterion to characterise the dependence between a potential model input and the output variable. To facilitate the calculation of dependence in the case of multiple inputs, a partial measure of the mutual information criterion is used. In the second method, a self-organizing map (SOM) is used to reduce the dimensionality of the input space and obtain independent inputs. To determine which inputs have a significant relationship with the output (dependent) variable, a hybrid genetic algorithm and general regression neural network (GAGRNN) is used. Both input determination techniques are tested on a number of synthetic data sets, where the dependence attributes were known a priori. In the second paper of the series, the input determination methodology is applied to a real-world case study in order to determine suitable model inputs for forecasting salinity in the River Murray, South Australia, 14 days in advance.

Keywords: Artificial neural networks; Input determination; Self-organizing map; Genetic algorithm; Mutual information; General regression neural network

1. Introduction

Artificial neural networks (ANNs) have been used in a wide variety of hydrologic contexts (ASCE Task Committee on Application of Artificial Neural Networks in Hydrology, 2000b; Dawson and Wilby, 2001; Maier and Dandy, 2000), ranging from rainfall-runoff models (e.g. Minns and Hall, 1996; Tokar and Johnson, 1999) to the development of ANNs for temporal rainfall disaggregation (Burian et al., 2001). One of the most important steps in the ANN development process is the determination of significant input variables. Usually, not all of the potential

* Corresponding author.
E-mail addresses: gbowden@deas.harvard.edu (G.J. Bowden), gdandy@civeng.adelaide.edu.au (G.C. Dandy), hmaier@civeng.adelaide.edu.au (H.R. Maier).

0022-1694/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.jhydrol.2004.06.021

input variables will be equally informative since some may be correlated, noisy or have no significant relationship with the output variable being modelled. Despite this, in most water resources ANN applications, very little attention is given to the task of selecting appropriate model inputs (Maier and Dandy, 2000). This is primarily because ANNs belong to the class of data driven approaches, whereas conventional statistical methods are model driven (Chakraborty et al., 1992). In the latter, the model's structure is determined a priori by using empirical or analytical approaches, before estimating the unknown model parameters, whereas data driven approaches are usually assumed to be able to determine which model inputs are critical. This has meant that ANN practitioners often present a large number of inputs to the ANN and rely on the network to identify the critical model inputs. There are a number of shortcomings associated with this approach, including (Back and Trappenberg, 1999; Maier and Dandy, 1997; Zheng and Billings, 1996):

- As input dimensionality increases, the computational complexity and memory requirements of the model increase.
- Learning becomes more difficult with irrelevant inputs.
- Misconvergence and poor model accuracy may result from the inclusion of irrelevant inputs due to an increase in the number of local minima present in the error surface.
- Understanding complex models is more difficult than understanding simple models that give comparable results.
- Due to the curse of dimensionality, some types of ANN models with many irrelevant inputs behave poorly since the network uses almost all its resources to represent irrelevant portions of the input-output mapping. Other types of networks that can efficiently concentrate on important regions of the input space require more data to efficiently estimate the connection weights when irrelevant inputs are included.

Consequently, there are obvious advantages in using analytical procedures to select an appropriate set of inputs for ANN models. The challenge of input determination is to select a subset of inputs from all potential inputs that will lead to a superior model as measured by some optimality criterion. For d potential inputs, there are 2^d-1 input subsets, hence, it is possible to test all subset combinations for small values of d, but for large values of d, as is often the case for complex problems, efficient algorithms are required. The problem is further exacerbated in time series studies, where appropriate lags must also be chosen. The difficulty lies in determining how many lagged values to include from each input time series. In general, if there are n input time series ($x_{j,t-1}$, $x_{j,t-2},\ldots,x_{j,t-N}$, $j=1,2,\ldots,n$), then the problem is finding the maximum lag for each input time series (k_j: $k_{j,\max}<N, j=1,2,\ldots,n$) beyond which, values of the input time series have no significant effect on the output time series. The maximum lag is also called the memory length. As the memory length increases, so too does the number of inputs and the complexity of the ANN model.

The objective of this article is to present a methodology that can be used for selecting the significant inputs to an ANN model. In developing this methodology, a review of the current input determination techniques used in hydrologic applications of ANN models has been conducted. The advantages and limitations of the techniques used in the past are addressed and used to formulate two input determination methods applicable to all water resources case studies. The efficacy of the proposed methods is determined by applying them to a number of test problems. In the second paper of this two-part series, the proposed methods are applied to a hydrologic case study involving forecasting salinity in a river environment.

2. Review of input determination in water resources ANN applications

Maier and Dandy (2000) reviewed 43 journal papers (published up until the end of 1998) on the application of ANNs for modelling water resources variables and found that, in many cases, the lack of a methodology for determining input variables raised doubt about the optimality of the inputs obtained. In some instances, inputs were chosen arbitrarily. In other cases, a priori knowledge was used for input selection and, when different approaches such as