Journal of Hydrology 367 (2009) 165–176

Contents lists available at ScienceDirect

Journal of Hydrology

journal homepage: www.elsevier.com/locate/jhydrol

Selection of input variables for data driven models: An average shifted histogram partial mutual information estimator approach

T.M.K.G. Fernando *, H.R. Maier, G.C. Dandy

School of Civil, Environmental and Mining Engineering, The University of Adelaide, Adelaide, Australia

ARTICLE INFO

Article history:
Received 26 March 2008
Received in revised form 29 August 2008
Accepted 20 October 2008

This manuscript was handled by P. Baveege, Editor-in-Chief, with the assistance of Chong-yu Xu, Associate Editor.

Keywords:
Artificial neural networks
Input selection
Average shifted histograms
Mutual information

SUMMARY

The use of artificial neural networks (ANNs) for the modelling of water resources variables has increased rapidly in recent years. This paper addresses one of the important issues associated with artificial neural network model development; input variable selection. In this study, the partial mutual information (PMI) input selection algorithm is modified to increase its computational efficiency, while maintaining its accuracy. As part of the modification, use of average shifted histograms (ASHs) is introduced as an alternative to kernel based methods for the estimation of mutual information (MI). Empirical guidelines are developed to estimate the key ASH parameters as a function of sample size. The stopping criterion used with the original PMI algorithm is replaced with a more computationally efficient outlier detection technique based on the Hampel distance. The performance of the proposed PMI algorithm, in terms of computational efficiency and input selection accuracy, is first investigated by using it to identify significant variables for data series where dependencies of attributes are known *a priori*. The proposed ASH PMI input variable selection algorithm with the Hampel distance stopping criterion consistently selects the correct inputs, while being computationally efficient. The modified PMI algorithm is then applied to identify suitable inputs to forecast salinity in the River Murray at Murray Bridge, South Australia, with a lead time of 14 days using an ANN approach. The ANN models developed with the inputs selected with the modified PMI algorithm perform very well when compared with results obtained using ANN models with different input sets developed in previous studies. Furthermore, the proposed input variable selection algorithm results in more parsimonious ANN models.

Introduction

There has been a rapid increase in the use of artificial neural networks (ANNs) for hydrological modelling (Maier and Dandy, 2000; Dawson and Wilby, 2001) due to their ease of development, decreased reliance on expert knowledge of the system under investigation and non-linear modelling capabilities. One of the significant steps in ANN model development is the selection of an appropriate set of input variables from the available candidates (e.g. Maier and Dandy, 2001; Bowden et al., 2005a; May et al., 2008). This is because the performance of data driven techniques, such as ANNs, is highly sensitive to the selected input variables. If relevant inputs are omitted (i.e. the model is under specified), the model is unable to capture the desired input–output relationships. On the other hand, if the model includes superfluous inputs (i.e. the model is over-specified), the following effects might incur: (i) the size, computational complexity and memory requirements of the model increase, (ii) model calibration becomes more difficult due to an increase in the size of the search space and the increased presence of local optima, (iii) the extraction of physical meaning from calibrated models is more difficult and (iv) more data are needed to efficiently estimate the optimal values of the connection weights (Back and Trappenberg, 1999; Maier and Dandy, 1997; Zheng and Billings, 1996).

Due to the negative consequences of over- and under-specification outlined above, there are distinct advantages in using analytical procedures for selecting an optimal model input vector from a set of candidates. When choosing an appropriate input selection algorithm, the following four factors need to be considered.

1. Input selection algorithms need to be able to determine the strength of the relationship between potential model inputs and outputs. The approaches used to achieve this generally fall into two categories: model-based and model-free. Model-based approaches use the performance of calibrated models with different inputs as the basis for choosing the most appropriate input vector. This has the advantage that non-linear relationships in the data can be taken into account. However, as ANNs have to be trained before the strength of the relationship between potential model inputs and output(s) can be determined, near-optimal model structures and values of the

* Corresponding author. Tel.: +61 08 8303 6139.
E-mail address: gfernand@civeng.adelaide.edu.au (T.M.K.G. Fernando).

0022-1694/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.jhydrol.2008.10.019

connections weights have to be obtained for each of the models developed. If this is not the case, the model inputs obtained based on the strength of the relationships extracted from the trained model are likely to be sub-optimal or even misleading. Model-free approaches overcome this shortcoming by using statistical measures of dependence to determine the strength of the relationship between candidate model inputs and the model output prior to model specification and calibration. However, care needs to be taken that non-linear dependence measures, such as mutual information, are used, rather than linear measures, such as correlation.

2. Input selection algorithms should cater for redundancy in candidate model inputs. Although a candidate model input might have a strong relationship with the model output, this information might be redundant if the same information is already provided by another input. In model-based approaches, redundancy is generally taken into account implicitly by use of a stepwise model-building process (e.g. forward selection or backward elimination). In model-free approaches, stepwise partial model building approaches (e.g. partial correlation, partial mutual information) can be used to eliminate redundant inputs.
3. Both model-based and model-free approaches generally use a stepwise process in the identification of appropriate model inputs. Consequently, there is a need for a stopping criterion that helps determine when to stop adding or removing candidate inputs. In model-based approaches, information criteria, such as Akaike's and Bayes' information criterion, provide viable alternatives, as they balance prediction error with model complexity. When using model-free approaches, the use of significance measures is well established for correlation-based methods. However, the same does not apply to non-linear dependence measures, such as mutual information, although Sharma (2000) and Bowden et al. (2005a) showed that the use of bootstrap methods can provide a suitable means of determining appropriate stopping criteria in such instances.
4. The computational efficiency of the input selection approach used is paramount, particularly when dealing with large datasets, as is the case for many hydrological modelling applications. Model-based approaches are generally very computationally inefficient, as they require the development (e.g. determination of optimal model structure, model calibration) of a large number of models. The efficiency of model-free approaches is a function of the dependence measure used. Use of correlation as the dependence measure is computationally efficient, but does not cater for non-linear dependencies, as discussed above. Use of mutual information as the dependence measure is generally computationally inefficient, although this depends on the method used to obtain estimates of mutual information, as well as the stopping criterion used.

In a review of approaches used to select the inputs to ANN models, Bowden et al. (2005a) concluded that the partial mutual information (PMI) algorithm of Sharma (2000) was superior to methods commonly used to determine the inputs to ANN models, as it is model-free, uses a non-linear measure of dependence (mutual information), is able to cater for input redundancy and has a well-defined stopping criterion. They also demonstrated the utility of the algorithm for selecting appropriate inputs to ANN models for hydrological applications (Bowden et al., 2005b). However, the algorithm is relatively computationally inefficient, particularly when dealing with large datasets.

The PMI algorithm's computational efficiency is a function of the method used to estimate mutual information (MI), as well as the number of times estimates of MI have to be obtained. In relation to the calculation of mutual information, this requires the estimation of marginal and joint probability densities. Commonly used density estimation techniques in the context of calculating MI include histograms and kernel density methods. Kernel based MI estimators are generally considered to be more reliable than histogram based approaches (Moon et al., 1995) and have therefore been used as part of the PMI algorithm (Bowden et al., 2005b; May et al., 2008; Sharma et al., 2000). However, the computational requirements associated with kernel density estimates are much greater than those associated with histogram based methods, especially when dealing with large data sets, such as those commonly used in hydrological modelling. In relation to the number of times estimates of MI have to be obtained, this is a function of the stopping criterion used. Both Sharma (2000) and Bowden et al. (2005a,b) used bootstrap methods to estimate the 95th percentile confidence limit of MI in order to determine when to stop adding candidate inputs. Use of this stopping criterion requires repeated estimation of MI values, further reducing the computational efficiency of the algorithm. Sharma (2000) and Bowden et al. (2005a) used 100 bootstraps for each iteration, generally requiring 1000s of estimates of MI, depending on the sample size and number of significant inputs. While this bootstrap size was found to produce accurate results for the test cases considered, bootstrap sizes between 5000 and 10,000 may be needed to obtain reliable estimates of the required confidence intervals, depending on the complexity of the problem (Chernick, 1999). This could make use of the algorithm computationally intractable for many hydrological case studies.

In this paper, a more computationally efficient version of the PMI algorithm is developed, tested and applied. As part of the proposed algorithm, computationally efficient estimates of MI are obtained by using average shifted histograms (ASHs) for density estimation. In addition, the number of times MI estimates have to be obtained is reduced significantly by using a stopping criterion that does not rely on the bootstrap method. As the accuracy of ASHs is a function of two user-defined parameters, guidelines for choosing these parameters are developed. In order to test the utility of the proposed algorithm, its performance is compared with that of the traditional PMI algorithm for a number of benchmark test cases in terms of accuracy and computational efficiency. Finally, the proposed algorithm is applied to the case study of forecasting salinity in the River Murray at Murray Bridge, South Australia, 14 days in advance, and the results compared with those obtained in previous studies in terms of the inputs identified as significant and the accuracy of the resulting forecasts.

Background

Mutual information

For a set of N bivariate measurements, $z_i = (x_i, y_i)$, $i = 1, \ldots, N$, which are assumed to be independent, identically distributed realizations of a random variable $Z = (X, Y)$, mutual information is defined as

$$I(X,Y) = \int\int f_{x,y}(x,y) \log_e \frac{f_{x,y}(x,y)}{f_x(x) f_y(y)} \mathrm{d}x\mathrm{d}y \quad (1)$$

where $f_x(x)$ and $f_y(y)$ are the marginal probability density functions of X and Y, respectively, and $f_{x,y}(x,y)$ is the joint probability density function of X and Y.

The mutual information score in (1) can be approximated as

$$MI = \frac{1}{N} \sum_{i=1}^{N} \log_e \left[\frac{f_{x,y}(x_i, y_i)}{f_x(x_i) f_y(y_i)} \right] \quad (2)$$

where $f_x(x_i)$, $f_y(y_i)$ and $f_{x,y}(x_i, y_i)$ are the respective marginal and joint densities estimated at the sample data point (Bonnlander, 1996;