Environmental Modelling & Software 71 (2015) 78–96

Contents lists available at ScienceDirect

Environmental Modelling & Software

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

Improving partial mutual information-based input variable selection by consideration of boundary issues associated with bandwidth estimation

Xuyuan Li, Aaron C. Zecchin*, Holger R. Maier

School of Civil, Environmental and Mining Engineering, The University of Adelaide, Adelaide, South Australia, 5005, Australia

ARTICLE INFO

Article history:
Received 4 December 2014
Received in revised form
23 May 2015
Accepted 24 May 2015
Available online 19 June 2015

Keywords:
Artificial neural networks
Data-driven models
Partial mutual information
Kernel density estimation
Kernel bandwidth
Boundary issues
Hydrology and water resources
Input variable selection

ABSTRACT

Input variable selection (IVS) is vital in the development of data-driven models. Among different IVS methods, partial mutual information (PMI) has shown significant promise, although its performance has been found to deteriorate for non-Gaussian and non-linear data. In this paper, the effectiveness of different approaches to improving PMI performance is investigated, focussing on boundary issues associated with bandwidth estimation. Boundary issues, associated with kernel-based density and residual computations within PMI, arise from the extension of symmetrical kernels beyond the feasible bounds of potential inputs, and result in an underestimation of kernel-based marginal and joint probability distribution functions in the PMI. In total, the effectiveness of 16 different approaches is tested on synthetically generated data and the results are used to develop preliminary guidelines for PMI IVS. By using the proposed guidelines, the correct inputs can be identified in 100% of trials, even if the data are highly non-linear or non-Gaussian.

Software availability

Software name: IVS_PMI_2014
Developers: Xuyuan Li, Postgraduate Student, the University of Adelaide, School of Civil, Environmental & Mining Engineering, Adelaide, SA 5005, Australia, xliadelaide@gmail.com
Hardware requirements: 64-bit AMD64, 64-bit Intel 64 or 32-bit x86 processor-based workstation or server with one or more single core or multi-core microprocessors; 256 MB RAM
Software requirements: All versions of Visual Studio 2012, 2010 and 2008 are supported except Visual Studio Express; PGI Visual Fortran 2003 or later version; Windows or Linux 2.6.32.2 operating system
Language: English
Size: 4.55 MB
Availability: Free to download for research purposes from the following website: https://github.com/xuyuanli/IVS_PMI_2014

1. Introduction

Input variable selection (IVS) plays a vital role in the development of data driven environmental models, such as artificial neural networks (ANNs), as the performance of such models can be compromised significantly if either too few or too many inputs are selected (Galelli et al., 2014; Maier et al., 2010; Wu et al., 2014a,b). Although the task of IVS is not unique to environmental modelling, its application in an environmental modelling context is complicated by a lack of understanding of the underlying physical processes, the presence of significant temporal and spatial variation in potential input variables, the non-Gaussian, correlated and collinear nature of potential input variables, and the non-linearity and inherent complexity associated with environmental systems themselves, as emphasised in Galelli et al. (2014). Given the importance and challenges associated with the IVS problem, a large number of approaches, categorised as either model free (utilising a statistical measure of significance between the candidate inputs

* Corresponding author. Tel.: +61 8 8303 3027; fax: +61 8 8303 4359.
E-mail addresses: xliadelaide@gmail.com (X. Li), aaron.zecchin@adelaide.edu.au (A.C. Zecchin), holger.maier@adelaide.edu.au (H.R. Maier).

http://dx.doi.org/10.1016/j.envsoft.2015.05.013
1364-8152/

and the output) or model based (utilising an optimization algorithm for determining the combination of input variables that maximizes the performance of a pre-selected data-driven model), have been developed and refined for the purpose of more accurate IVS (e.g. Galelli and Castelletti, 2013; Galelli et al., 2014; Li et al., 2015; May et al., 2011; May et al., 2008b; Sharma, 2000), with the specific aim to determine the number of inputs that best characterise the input–output relationship with the least amount of variable irrelevance or redundancy (Galelli et al., 2014; Guyon and Elisseeff, 2003). Among existing IVS techniques, partial mutual information (PMI) based approaches are among the most promising model free techniques, as they account for both the significance and independence of potential inputs and have been successfully and extensively implemented in environmental modelling (e.g. Bowden et al., 2005a,b; Fernando et al., 2009; Galelli et al., 2014; Gibbs et al., 2006; He et al., 2011; Li et al., 2015; May et al., 2008a,b; Wu et al., 2014b; Wu et al., 2013).

The PMI IVS approach was introduced by Sharma (2000) and is based on Shannon's entropy (Shannon, 1948), which measures the Mutual Information (MI) between a random input variable X and a random output variable Y as the reduction in uncertainty of Y due to observation of X. As part of the PMI algorithm, inputs are chosen as part of a forward selection approach, during which one input variable is selected at each iteration of the algorithm (starting with an empty set), based on the amount of information a potential input provides (in addition to inputs selected at previous iterations), until certain stopping criteria are met. The amount of information provided by a potential input is given as a function of mutual information (MI) and the contribution of already selected inputs is accounted for by calculating the MI between potential inputs and the residuals of models between the already selected inputs and the desired output, referred to as PMI. Consequently, the performance of different implementations of the PMI algorithm, in terms of input variable selection accuracy and computational efficiency, is a function of the methods used for mutual information (MI) and residual estimation (RE), as highlighted in Li et al. (2015) and May et al. (2008b).

In previous studies on the use of PMI for IVS for data-driven environmental models, the requisite MI and RE are a function of marginal and joint PDFs estimated by kernel density and kernel regression (for the estimation of kernel density based weights) based methods (e.g. Bowden et al., 2005a,b; Gibbs et al., 2006; He et al., 2011; Li et al., 2015; May et al., 2008a,b). Kernel methods are the approaches to constructing input/output (I/O) models from input and output data. The resulting I/O model is an ensemble of kernel functions, each centred about a data point in the input space, and returns a weighted average of the influence of all data points. The weight associated with each data point is dependent on the proximity of the input to that data point (i.e. closer points have more influence). Kernel methods are primarily controlled by a bandwidth parameter, which determines the extent to which a single kernel is spread throughout the input space (e.g. a small bandwidth means that data points will only have a localised influence). As such, the performance of PMI IVS is heavily influenced by the accuracy of the kernel density estimates required for MI and RE, which are a function of bandwidth (used interchangeably with smoothing parameter) selection and how well any boundary issues are addressed (Santhosh and Srinivas, 2013; Scott, 1992; Wand and Jones, 1995), as discussed below.

Determination of the optimal bandwidth (the bandwidth that provides the most accurate estimation of the density function) is not trivial, as there is no clear consensus as to which bandwidth estimator performs best for general cases. Over-estimating the bandwidth can lead to an over-smoothing of the probability density function (PDF) or residual predictions, so that detailed local information will not be effectively captured. On the contrary, under-estimating the bandwidth can make the general trend become more vulnerable to localised features, or even noise (Li et al., 2014). Although many methods for bandwidth estimation exist in other disciplines (e.g. mathematics and statistics (e.g. Hall et al., 1992; Park and Marron, 1990; Rudemo, 1982; Scott, 1992; Scott and Terrell, 1987)), in almost all existing PMI IVS studies in environmental modelling (e.g. Bowden et al., 2005a,b; He et al., 2011; May et al., 2008a,b) the Gaussian reference rule (GRR) has been used predominately for bandwidth estimation due to its simplicity. However, as highlighted by Harrold et al. (2001) and Galelli et al. (2014), use of the GRR can result in less accurate estimation of MI and PMI for data that are highly non-Gaussian, which is generally the case in environmental and water resources modelling problems. In addition, Li et al. (2015) showed that PMI IVS performance can be improved if alternative bandwidth estimation methods are used for MI and RE for data that are non-Gaussian.

Another potential problem with kernel based methods is the so called 'boundary issue', which is associated with the inaccuracies in density estimation arising from the extension of symmetrical kernels beyond the feasible bounds of potential input variable values (e.g. densities associated with negative values of flow obtained using symmetrical kernels) (Wand and Jones, 1995) and generally results in an underestimation of MI or residuals near the boundary. This is commonly encountered in environmental and water resources modelling by the fact that data can be bounded due to their physical feasibility (e.g. rainfall-runoff data are bounded at zero). Although a number of potential methods have been proposed within the statistical literature for addressing this issue (e.g. Cowling and Hall, 1996; Dai and Sperlich, 2010; Fan, 1992; Fan and Gijbels, 1996; Gasser and Müller, 1979; Hall and Park, 2002; Marron and Ruppert, 1994; Schuster, 1985; Zhang and Karunamuni, 1998), their effectiveness has not yet been tested in the context of PMI IVS for data-driven environmental modelling. However, this is likely to be a significant problem, as environmental data can be highly skewed near variable boundaries. Consequently, there is a need to establish to what degree the performance of PMI IVS is influenced by the boundary issue, and which methods are the most effective in addressing such issue.

In order to address the aforementioned research need, the objectives of the current study are: (i) to assess if, and to what degree, the performance of PMI IVS can be improved by various approaches to addressing boundary issues for data with different properties (i.e. degree of linearity and degree of normality); and (ii) to develop and test a set of preliminary empirical guidelines for the selection of the most appropriate methods for bandwidth estimation and addressing boundary issues for data with different properties. The remainder of this paper is organised as follows. An explanation of PMI IVS and boundary issues is provided in Section 2, followed by the methodology for fulfilling the outlined objectives in Section 3. The results are presented and analysed in Section 4. The proposed guidelines are validated on the semi-real studies in Section 5, before a summary and conclusions given in Section 6.

2. Background on PMI IVS and boundary issues

2.1. PMI IVS

Although details of the PMI IVS approach are provided in a number of papers (e.g. Sharma, 2000; Bowden et al., 2005a; May et al., 2008b; He et al., 2011; May et al., 2011; Li et al., 2015), a brief outline of the main steps in the process are given below for the sake of completeness: