Journal of Hydrology 542 (2016) 18–34

Contents lists available at ScienceDirect

Journal of Hydrology

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

Research papers

An information theoretic approach to select alternate subsets of predictors for data-driven hydrological models

R. Taormina [a,*], S. Galelli [a], G. Karakaya [b], S.D. Ahipasaoglu [a]

[a] Pillar of Engineering Systems and Design, Singapore University of Technology and Design, Singapore
[b] Department of Business Administration, Middle East Technical University, Ankara, Turkey

ARTICLE INFO

Article history:
Received 8 May 2016
Received in revised form 27 July 2016
Accepted 29 July 2016
Available online 30 July 2016
This manuscript was handled by Geoff Syme, Editor-in-Chief

Keywords:
Input variable selection
Information theory
Data-driven models
Extreme learning machines
Neural networks

ABSTRACT

This work investigates the uncertainty associated to the presence of multiple subsets of predictors yielding data-driven models with the same, or similar, predictive accuracy. To handle this uncertainty effectively, we introduce a novel input variable selection algorithm, called Wrapper for Quasi Equally Informative Subset Selection (W-QEISS), specifically conceived to identify all alternate subsets of predictors in a given dataset. The search process is based on a four-objective optimization problem that minimizes the number of selected predictors, maximizes the predictive accuracy of a data-driven model and optimizes two information theoretic metrics of relevance and redundancy, which guarantee that the selected subsets are highly informative and with little intra-subset similarity. The algorithm is first tested on two synthetic test problems and then demonstrated on a real-world streamflow prediction problem in the Yampa River catchment (US). Results show that complex hydro-meteorological datasets are characterized by a large number of alternate subsets of predictors, which provides useful insights on the underlying physical processes. Furthermore, the presence of multiple subsets of predictors—and associated models—helps find a better trade-off between different measures of predictive accuracy commonly adopted for hydrological modelling problems.

1. Introduction

Data-driven hydrological models are the product of an 'inductive' process, in which a functional relationship between a subset of predictors, or input variables, and an output variable is inferred from observational data. This process involves four main stages (Young, 2001): identification of the possible predictors, specification of an appropriate model structure, estimation of the parameters that best characterize this structure (calibration), and predictive validation on a dataset different from the one used in calibration. An additional step is the quantification of the uncertainty associated to the model, which generally depends on multiple sources—e.g., measurement errors in the observational data or uncertainty in the model structure, parameters and selected predictors. Uncertainty in the model structure and parameters has been extensively addressed during the past decade; available methods include Bayesian techniques (Khan and Coulibaly, 2006; Kingston et al., 2008; Zhang et al., 2009, 2011), bootstrapping (Srivastav et al., 2007; Sharma and Tiwari, 2009; Tiwari and Chatterjee, 2010), fuzzy theory (Shrestha and Solomatine, 2006; Alvisi and Franchini, 2011; Taormina and Chau, 2015a) and ensemble modelling (Dawson et al., 2002; Parasuraman and Elshorbagy, 2008), for instance. All these methods study the effect of different structures and parameterizations to issue an uncertainty estimation of the output in the form of confidence intervals. Alternatively, one can directly use the time series of residuals to create a model of the predictive uncertainty (see Pianosi and Raso (2012), and references therein).

The uncertainty in the selected predictors is often addressed by means of input variable selection (Guyon and Elisseeff, 2003), whose aim is to identify a combination of predictors that best characterizes the input–output relationship being modelled. In particular, the objective of variable selection is to simplify the model identification process by avoiding the interference of redundant and irrelevant predictors, leading to parsimonious, cost-effective and easy to interpret models. Variable selection algorithms can be categorized in *wrapper* and *filter* approaches, depending on the way the selection process is carried out (Maier et al., 2010). Wrappers use global optimization techniques to select the combination of predictors that maximizes the accuracy of a given model structure. Filters conduct a local search usually informed by a statistical measure of significance (e.g., linear correlation, mutual information (MacKay, 2003)) that quantifies the relevance of each

* Corresponding author.
E-mail address: riccardo_taormina@sutd.edu.sg (R. Taormina).

http://dx.doi.org/10.1016/j.jhydrol.2016.07.045
0022-1694/

predictor with respect to the output. Filters are computationally less expensive than wrappers. On the other hand, wrappers explore the search space in a more exhaustive way and are thus likely to identify more accurate models. Variable selection algorithms are used in a variety of water resources modelling problems, such as flood regionalization (Wan Jaafar et al., 2011), statistical downscaling (Phatak et al., 2011), streamflow and water quality modelling (Bardsley et al., 2015; Li et al., 2015; Creaco et al., 2016), medium-term hydro-climatic forecasts (Sharma, 2000; Noori et al., 2011) and forecast of urban water demand (Quilty et al., 2016) (see Galelli et al. (2014) for a review). Whilst all these modelling problems are characterized by the presence of alternate subsets of predictors, the majority of variable selection algorithms yields one subset of predictors, thereby providing a narrow view of this source of uncertainty. Such issue may sharpen as larger observational datasets become available, owing to advances in remote sensing techniques (Ceola et al., 2015), ubiquitous mobile technology (Overeem et al., 2013) and crowdsourcing (Fraternali et al., 2012; Mount et al., 2016). Only recently, some techniques designed to identify a few subsets of predictors with the same (or similar) information content have been proposed (Sharma and Chowdhury, 2011; Taormina and Chau, 2015b; Creaco et al., 2016).

To get around the uncertainty in the selected predictors, we introduce a novel variable selection algorithm, called Wrapper for Quasi Equally Informative Subset Selection (W-QEISS), which is specifically conceived to identify all equally informative subsets of predictors in a given hydrological dataset—in the context of this study, we say that two subsets are quasi equally informative if they have (almost) the same predictive accuracy with respect to a given model class (Karakaya et al., 2016). W-QEISS builds on the formulation of a four-objective optimization problem that maximizes the predictive accuracy of a pre-selected model, minimizes the number of predictors, and optimizes two information theoretic metrics of relevance and redundancy. The optimization of the predictive accuracy ensures that the interaction between the model structure and data is maximized, while the minimization of the cardinality is aimed at simplifying the final models. The adoption of relevance and redundancy metrics guarantees that the selected subsets are highly informative and with low intra-subset similarity. These metrics provide a generic measure of dependence between variables, so they can characterize any functional relationship (Sharma and Mehrotra, 2014). Extreme Learning Machines (Huang et al., 2012)—a learning algorithm for single-layer feedforward neural networks—are used as model structure, since they have high predictive accuracy and limited computational burden as compared to other non-linear data-driven techniques. This aspect is particularly relevant for wrappers, which repeat the calibration and validation process several times to identify the best subset of predictors. The global optimization problem is solved by means of Borg MOEA (Hadka and Reed, 2013), a Multi-Objective Evolutionary Algorithm designed to handle multimodal problems and expensive objective function evaluations. We show that the identification of quasi equally informative subsets of predictors yields two main advantages. First, it determines the relative importance of each predictor, thus assisting in the interpretation of the underlying physical processes. Second, it helps find a better trade-off between multiple measures of predictive accuracy commonly used to assess the performance of hydrological models.

In the remainder of the paper, we first formulate the quasi equivalence of subsets, give some background on information theoretic criteria and describe W-QEISS algorithm. We then assess its performance on two synthetic (linear and non-linear) test problems and demonstrate it on a real-world case study of streamflow prediction in the Yampa River catchment (US), whose discharge is contributed by both rainfall and snowmelt processes. A comparison between the subsets of predictors selected by W-QEISS and four popular variable selection algorithms is included as benchmarking exercise. Concluding remarks and future research directions are outlined in Section 6.

2. Methods and tools

2.1. Quasi equivalence of subsets

In this study, we assume that for a given dataset there can be multiple subset of predictors resulting in models having the same (or similar) predictive accuracy (Karakaya et al., 2016). We define the *quasi equivalence of subsets* as follows.

Definition 1. Let $\bar{f}(\cdot)$ be a metric of predictive accuracy taking values between 0 and 1, where 0 indicates that a model does not have any predictive skill while 1 denotes perfect predictive accuracy. Given a subset of predictors S_j, we say that another subset S_i is δ-quasi equally informative to S_j if the two subsets have almost equal predictive accuracy with respect to a given model class, i.e., subset S_i is δ-quasi equally informative to subset S_j if $\bar{f}(S_i) \geqslant (1-\delta)\bar{f}(S_j)$ for $0 \leqslant \delta \leqslant 1$.

The presence of quasi equivalent subsets depends on the specific interactions between the variables in the observational dataset at hand. As discussed in Yu and Liu (2004), predictors can be classified depending on their relevance with respect to the output variable as strongly relevant, weakly relevant and irrelevant. A predictor is strongly relevant if it cannot be removed from a subset without affecting its predictive accuracy, while it is irrelevant if it is not necessary with respect to the modelling task. A predictor is weakly relevant if there exists a combination of other relevant predictors that subsume the same information about the output variable and the rest of the predictors—this combination is named *Markov blanket*. A weakly relevant predictor can be either redundant or necessary in the modelling process. It is redundant if its Markov blanket is included in the selected subset, while it is necessary (i.e., weakly relevant but non redundant) in the opposite case. A set of predictors can thus be divided into four disjoint parts on the basis of their relevance and redundancy (Fig. 1): strongly relevant, weakly relevant but non redundant, weakly relevant and redundant, irrelevant predictors. Since different partitions of weakly relevant predictors can result during the selection process, different quasi equally informative subsets can potentially be identified (Liu et al., 2015). The identification of these subsets requires specific metrics of relevance and redundancy and a search algorithm, especially when the dimension of the observational dataset precludes an exhaustive search in the predictors space.

2.2. Background on information theoretic criteria

Information theoretic criteria (e.g., Mutual Information, Partial Mutual Information) are widely used for quantifying the dependence between two or multiple variables. Their main advantage is that they do not assume any hypothesis on the underlying functional relationship (e.g., linear dependence), so they are adaptable to different modelling contexts (MacKay, 2003). These criteria are based on the concept of Shannon entropy (Shannon, 1948), which is a measure of the uncertainty associated with a random variable. For a continuous random variable X, the entropy $H(X)$ is defined as

$$H(X) = -\int p(x)\log p(x)dx, \tag{1}$$