Available online at www.sciencedirect.com
ScienceDirect
Neural Networks 20 (2007) 537–549
Neural Networks
www.elsevier.com/locate/neunet

2007 Special Issue

Predictive uncertainty in environmental modelling

Gavin C. Cawley[a,*], Gareth J. Janacek[a], Malcolm R. Haylock[b], Stephen R. Dorling[c]

[a] School of Computing Sciences, University of East Anglia, Norwich NR4 7TJ, United Kingdom
[b] Climatic Research Unit, University of East Anglia, Norwich NR4 7TJ, United Kingdom
[c] School of Environmental Sciences, University of East Anglia, Norwich NR4 7TJ, United Kingdom

Abstract

Artificial neural networks have proved an attractive approach to non-linear regression problems arising in environmental modelling, such as statistical downscaling, short-term forecasting of atmospheric pollutant concentrations and rainfall run-off modelling. However, environmental datasets are frequently very noisy and characterized by a noise process that may be heteroscedastic (having input dependent variance) and/or non-Gaussian. The aim of this paper is to review existing methodologies for estimating predictive uncertainty in such situations and, more importantly, to illustrate how a model of the predictive distribution may be exploited in assessing the possible impacts of climate change and to improve current decision making processes. The results of the WCCI-2006 predictive uncertainty in environmental modelling challenge are also reviewed, suggesting a number of areas where further research may provide significant benefits.

Keywords: Predictive uncertainty; Environmental modelling; Multilayer perceptron; Statistics

1. Introduction

Neural networks have been shown to provide a simple and flexible approach to a wide variety of non-linear regression problems arising in the environmental sciences. Some recent applications include statistical downscaling (Harpham & Wilby, 2005), water level-discharge modelling (Bhattacharya & Solomatine, 2005), river stage forecasting (Dawson et al., 2005) and air quality forecasting (Schlink et al., 2003). The presence of special sessions devoted to environmental sciences and climate modelling at IJCNN-2005 and IJCNN-2006 provides further evidence of the importance of this field of research. Environmental modelling problems are typically very noisy and often characterized by a noise process that is heteroscedastic (i.e. the variance of the noise process is input-dependent) and may also be non-Gaussian, for example the target data may be strictly non-negative or highly skewed. Conventional neural network regression techniques aim to estimate the conditional mean of the target data, via minimization of a sum-of-squares error function. The aim of this paper is to demonstrate that practical benefits can be accrued by attempting to model the entire conditional distribution of the noise contaminating the data in addition to the conditional mean. For example, we may estimate the conditional variance of a Gaussian noise process, which may be achieved by training a second regression network to predict the squared residuals of the first (e.g. Nix and Weigend (1994)). The combined model provides a Gaussian *predictive distribution* indicating the relative plausibility of different values for the target function. The provision of a predictive distribution, instead of only the conditional mean, can be exploited in a number of ways:

- The predictive distribution implies a plausible interval (a.k.a. "error bars") on all predictions, which in turn provide a valuable indicator of the reliability of the model.
- An estimate of the predictive distribution allows the estimation of the true *risk*, i.e. we may integrate the loss associated with all plausible outcomes, weighted by the probability of their occurrence.
- Where a neural network is used as one component within a much larger model, the uncertainties associated with the inputs and outputs of each component, may be propagated through the model (e.g. via a Monte-Carlo simulation) so that all sources of uncertainty can be integrated over to obtain a moderated prediction.

* Corresponding author.
E-mail addresses: gcc@cmp.uea.ac.uk (G.C. Cawley), M.Haylock@uea.ac.uk (M.R. Haylock), S.Dorling@uea.ac.uk (S.R. Dorling).

0893-6080/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.neunet.2007.04.024

- Often we are interested in predicting extreme events, especially the exceedance of some arbitrary threshold, for instance predicting episodes of poor air quality. By their very nature, extreme events are not modelled well by an estimate of conditional mean of the data, and so a conventional sum-of-squares model will consistently under-predict extreme events. However, given a full predictive distribution, we may at least estimate the *probability* of an extreme event by integrating the upper tail of the predictive distribution, even if the estimate of the conditional mean never exceeds the threshold.

Modelling predictive uncertainty in environmental data is also interesting from a machine learning perspective as the noise processes involved are often non-Gaussian and/or heteroscedastic, and so "off-the-shelf" solutions may not be entirely satisfactory; thus there is significant scope for further research.

The remainder of this paper is structured as follows: Section 2 describes the four benchmark datasets used in the WCCI-2006 predictive uncertainty in environmental modelling challenge. Section 3 presents a variety of conventional statistical approaches, and discusses the deviations from the usual modelling assumptions often encountered in environmental modelling. Section 4 describes a simple methodology for estimating the predictive distribution based on methods developed by Peter Williams (Williams, 1991, 1995, 1996, 1998). Section 5 demonstrates that an estimate of the predictive distribution can be exploited to provide practical benefits for the end-user, via an illustrative (if a little contrived) example based on the estimation of insurance losses associated with flood hazards. The results of the WCCI-2006 Predictive Uncertainty in Environmental Modelling Competition, which aimed to stimulate research in this area, are presented in Section 6. Section 7 discusses some areas where further research may provide significant benefits. Finally, the work is summarized and conclusions drawn in Section 8.

2. Datasets

In this section, we describe the four benchmark datasets that were used in the WCCI-2006 predictive uncertainty in environmental modelling challenge. These benchmarks are also used in this paper to demonstrate the importance of estimating predictive uncertainty, especially for datasets characterized by a non-Gaussian or heteroscedastic variance structure. These datasets are freely available from the challenge website (http://theoval.cmp.uea.ac.uk/~gcc/competition/).

2.1. The SYNTHETIC benchmark

A synthetic heteroscedastic regression problem, taken from Williams (1996), provides a relatively small dataset that can be easily visualized for the purposes of model development and for illustrating the importance of predictive uncertainty. As the true conditional mean and variance functions are known, it is straightforward to assess the quality of the model without direct access to the test data. The univariate input patterns, x, are drawn from a uniform distribution on the interval $(0, \pi)$, the corresponding targets, y, are drawn from a univariate Normal distribution with mean and variance that vary smoothly with x:

$$x_i \sim \mathscr{U}(0, \pi),$$
$$y_i \sim \mathscr{N}\left(\sin\left[\frac{5x_i}{2}\right]\sin\left[\frac{3x_i}{2}\right], \frac{1}{100} + \frac{1}{4}\left[1 - \sin\left[\frac{5x_i}{2}\right]\right]^2\right).$$

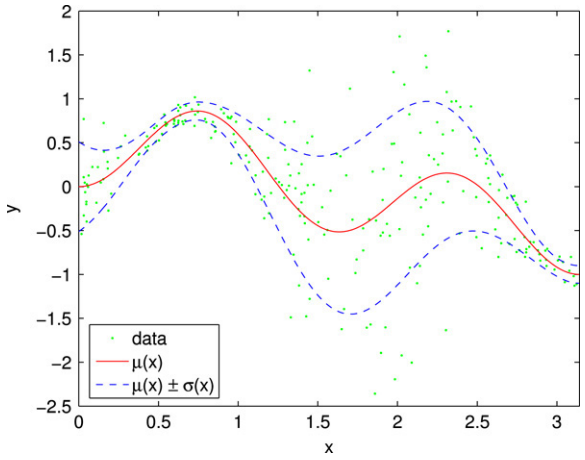

Fig. 1. Plot of the training data for the SYNTHETIC benchmark dataset, along with an indication of the true conditional mean, $\mu(x)$ and conditional standard deviation, $\sigma(x)$.

Fig. 1 shows a plot of the synthetic benchmark dataset, along with indications of the true conditional mean and standard deviation. The heteroscedastic (input-dependent variance) nature of the data is clearly evident.

2.2. The SO2 benchmark

The SO2 benchmark represents an atmospheric pollution forecasting problem, where the aim is to predict 24 hours in advance the SO_2 concentration in urban Belfast, based on meteorological conditions and current SO_2 levels (see Nunnari (2004) for further details). The meteorological conditions are important in this case as the air pollution problem in urban Belfast is largely due to domestic (commonly coal-fired) heating, and so is at its worst during periods of cold weather. Also high atmospheric pressure and temperature inversions tend to cause stagnant conditions and consequently poor dispersion of atmospheric pollutants.

2.3. The PRECIP benchmark

The PRECIP benchmark models a realistic statistical downscaling exercise, the aim of which is to predict the (scaled) precipitation for Newton Rigg, a relatively wet station in the North-West of the United Kingdom, using inputs representing large scale circulation features (see Cawley, Dorling, Jones, and Goodess (2003); Haylock, Cawley, Harpham, Wilby, and Goodess (2006) for further details). Fig. 2 shows a histogram of the target data for the training set of the PRECIP benchmark, highlighting a number of unusual features of this dataset. First, the data are non-negative (it would make little sense to talk