Ocean Engineering 117 (2016) 292–301

Contents lists available at ScienceDirect

Ocean Engineering

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

Massive missing data reconstruction in ocean buoys with evolutionary product unit neural networks

A.M. Durán-Rosal [a,*], C. Hervás-Martínez [a], A.J. Tallón-Ballesteros [b], A.C. Martínez-Estudillo [c], S. Salcedo-Sanz [d]

[a] Department of Computer Science and Numerical Analysis, Universidad de Córdoba, Rabanales Campus, 14071 Córdoba, Spain
[b] Department of Languages and Computer Systems, Universidad de Sevilla, 41012 Seville, Spain
[c] Department of Management and Quantitative Methods, Universidad Loyola Andalucía, 41014 Seville, Spain
[d] Department of Signal Processing and Communications, Universidad de Alcalá, 28805 Alcalá de Henares, Madrid, Spain

ARTICLE INFO

Article history:
Received 15 May 2015
Received in revised form
23 February 2016
Accepted 22 March 2016
Available online 1 April 2016

Keywords:
Significant wave height
Missing values reconstruction
Product unit neural networks
Evolutionary algorithm

ABSTRACT

In this paper we tackle the problem of massive missing data reconstruction in ocean buoys, with an evolutionary product unit neural network (EPUNN). When considering a large number of buoys to reconstruct missing data, it is sometimes difficult to find a common period of completeness (without missing data on it) in the data to form a proper training and test set. In this paper we solve this issue by using partial reconstruction, which are then used as inputs of the EPUNN, with linear models. Missing data reconstruction in several phases or steps is then proposed. In this work we also show the potential of EPUNN to obtain simple, interpretable models in spite of the non-linear characteristic of the neural network, much simpler than the commonly used sigmoid-based neural systems. In the experimental section of the paper we show the performance of the proposed approach in a real case of massive missing data reconstruction in 6 wave-rider buoys at the Gulf of Alaska.

1. Introduction

Oceanographic buoys are measuring instruments used, among other purposes, to characterize wind-generated wave properties. The availability and accuracy of buoys data is crucial in very different problems and applications such as the design and maintenance of marine/coastal structures, wave height forecasting for safe ship navigation, or the design and operation of wave energy converters, etc. (López et al., 2013). There are different agencies such as the National Data Buoy Centre (NDBC) of the USA that maintain a large network of OBs to collect wave data on a regular basis (Londhe, 2008). A number of unexpected events can make buoys break down (such as storms, Rao and Mandal, 2005, navigation accidents, maintenance periods, etc.), causing data gaps, and therefore discontinuities in the buoys data time series, lasting from the causing event until the buoy is repaired/maintained. Some data analysis methods may allow gappy data (Liu and Wisberg, 2005), while most statistical methods require data to be gaps free (Thomson and Emery, 2014). Due to this point, the reconstruction of missing wave values has become a key topic in oceanic research.

Very different techniques for recovering of lost/missing OB data have been proposed in the last three decades, with prevalence of techniques focused on reconstruction of wave height data. The first approaches were quite naive, such as random sampling of data points suggested in Thompson (1971), or Monte Carlo methods applied to fill up gaps at random in a known time series of monthly mean sea level in Sturges (1983). In more recent works such as Soares and Cunha (2000) and Agrawal and Deo (2002), auto-regressive, auto-regressive moving average (ARMA) or auto-regressive integrated moving average (ARIMA) have been successfully applied to reconstruction of wave heights time series. Other constructive techniques such as cubic splines or fractal methods have been recently tested in wave height reconstruction problems (Liu et al., 2014).

In recent years, the number of works applying data machine learning (ML) techniques has been massive. Among ML techniques, neural networks (NNs) (Haykin, 1998) may have been the most used prediction methods. In Bhattacharya et al. (2003), NNs have been used to compute missing wave data in time series, measured at Europlatform station, in the North Sea. NNs have been found to be specially reliable to reach accurate estimations of missing wave data (Balas et al., 2004). In that work, feedforward multi-layer perceptrons (MLPs) and recurrent neural networks were trained by the steepest descent with momentum algorithm and the conjugate gradient algorithm, and their estimations were

* Corresponding author.
E-mail address: i92duroa@uco.es (A.M. Durán-Rosal).

http://dx.doi.org/10.1016/j.oceaneng.2016.03.053
0029-8018/

compared to those computed by using conventional stochastic models. The recurrent neural network approach trained by the conjugate gradient algorithm was found to predict wave height, period, and direction more accurately than the NN. In Gunaydin (2008), the prediction of significant wave heights using neural networks, trained with the classical back-propagation algorithm in three stations in the Atlantic area has been carried out. In Londhe and Panchang (2007) the problem of reconstructing lost information from buoys by applying neural networks trained with Levernberg–Marquardt algorithm in the north-eastern part of United States was tackled. In Malekmohamadi et al. (2008) a NN coupled to a numerical prediction model was used to obtain wave height prediction/reconstruction values. Results in buoys located at lake Superior and in the Pacific Ocean were used to validate the model. In Zamani et al. (2008) a NN with *k*-nearest neighbor algorithm has been applied to a problem of wave estimation using significant heights in previous hours. Results in two buoys located at Caspian sea have shown a good agreement of the prediction with real data. In Arena and Puca (2004) the use a multivariate neural network for the reconstruction of significant wave height time series was proposed, and in Setiawan et al. (2008) alternative methodologies for training neural networks are introduced, including some concepts from the rough set theory.

NNs and genetic programming (GP) (Koza, 1992) were successfully applied in Londhe (2008) to estimate missing wave heights from neighboring stations height measurements. Six buoys from Gulf of Mexico were used, and results obtained showed a good reconstruction of weight heights using these techniques. GP and NNs have also been applied to wave forecast from wind data in Nitsure et al. (2012), obtaining good results in data from the east coast of the USA and India. Also using the data from adjacent buoys, significant wave height values at a buoy location have been estimated by means of a neuro-fuzzy approach (Özger, 2009) based on Sugeno-type fuzzy inference. In that paper, the antecedent and consequent part parameters of the fuzzy IF–THEN rules have been inferred after training the fuzzy inference system by the adaptive neuro fuzzy inference system (ANFIS) method (Jang, 1993). The experimental work was also carried out by using data measured by buoys located at the Gulf of Mexico. NN and ANFIS methods were tested in a problem of wave parameters estimation in Mahjoobi et al. (2008), over lake Ontario data obtained from a deep-water buoy. In Kalra and Deo (2007), particular GP models based on spatial correlation with neighboring values have been applied to wave height estimation. The results suggest that GP achieves slightly better results than the NNs explored by the authors in previous works. In a similar scientific standpoint, Ustoorikar and Deo (2008) leads to the conclusion that gaps of missing values are slightly better computed by the GP algorithm than those estimated by the NN designed for comparative purposes, in particular if the gaps of missing values are small. Radial basis functions (RBF) have also been applied to similar problems like in Camus et al. (2011), where this class of neural network was applied to a problem of down-scaling waves data to shallow waters. There have been alternative approaches that hybridize physical/numerical methods with statistical or ML approaches, such as Casas-Prat et al. (2014), where different predictive variables from numerical models are hybridized with linear regression for wave height estimation, or Fernández et al. (2015), that used predictive variables from numerical methods hybridized with different ordinal classifiers for discrete wave height estimation in the Gulf of Alaska. Some kernel methods have also been applied to wave height prediction problems such as Mahjoobi and Mosabbeb (2009), where support vector regression algorithms using different kernels were applied to wave height estimation in Lake Michigan.

Other approaches for wave height missing data imputation are based on evolutionary algorithms. One of this works is Nelwamondo et al. (2007), that compares expectation maximization (EM) techniques and auto-associative neural networks and genetic algorithms in the scope of industrial data. Genetic algorithms and neural networks were evaluated to approximate missing data with a variable number of missing values within a register in Abdella and Marwala (2005). The approach in Medina and Serrano-Hidalgo (2004) is based on the use of pruned neural network models optimized by evolutionary strategies and the utility function concept. The method is applied to reconstruct the time series of the significant wave height of a wave buoy at Bilbao on the northern coast of Spain. Evolutionary multi-layer perceptron models trained via genetic algorithms have been recently proposed in Altunkaynak (2013).

Finally, there have been studies of missing values of spatio-temporal wave time series. One of the first works to interconnect data from multiple stations was Tsai et al. (2002). In Ho and Yim (2005) the possible interrelationships between two stations, which are 20 km away from each other, by means of transfer function, was analyzed and more recently, a neuro-fuzzy strategy was proposed in Özger (2009) to recover the data from a buoy with the recorded information in their neighbor buoys.

In all these previous works based on measurements from neighbor buoys, the method is the following: first, complete time periods in all the buoys involved in the reconstruction (the inputs and the objective one) are obtained. Then, the algorithms are trained using part of these data, and the performance of the reconstruction is evaluated in the remainder of the data (test set). Once the algorithms are trained, missing values in the objective buoy can be reconstructed by using the corresponding values in the input buoys. The main issue with this method is observed in cases with massive missing data in some buoys, or cases with a large number of buoys, in which it is very difficult to obtain large periods of common non-missing values. In the case of using ML techniques to missing data reconstruction, there is an additional drawback: the lack of interpretability inherent to these techniques. These methods are usually highly non-linear, with a strong ability in obtaining information from data, but very difficult to interpret in terms of predictive (physical) variables.

With the previous ideas in mind, the objective of this paper is twofold: on one hand, we propose a method for buoy missing data reconstruction in the case of massive amount of data that do not allow obtaining large complete periods for training the algorithms. In order to do this, we apply well-known linear models (transfer functions, Ho and Yim, 2005, and neighbor correlation, Paulhus and Kohler, 1952) combined with product unit neural networks (PUNN), to form a method that allows a complete reconstruction of missing data, using partial reconstruction from linear approaches. On the other hand, PUNN networks have the property that linear models can be obtained from them by taking logarithms in the networks output, so we finally come up with linear interpretable models, but constructed using nonlinear procedures, which partially solves the issue about non-linearity of ML methods. The procedure proposed in this paper have obtained excellent results in terms of missing value reconstruction in a case with a large number of buoys involved in the Gulf of Alaska, with interpretable models that can be discussed in terms of predictive variables.

This paper aims to provide a method of automatic reconstruction of time series with massive missing values. Our purpose is to use EPUNN, where the inputs of the networks are the partial reconstruction obtaining by linear methods previously, to improve others basic functions in terms of accuracy and simplicity. Furthermore, combining these approaches (linear and non-linear methods), we apply the proposed method in the reconstruction of the wave height of 6 wave-rider buoys at the Gulf of Alaska.

The rest of the paper is structured as follows: next section presents the EPUNN method proposed. Linear models serving as