Applied Soft Computing xxx (2016) xxx–xxx

Contents lists available at ScienceDirect

Applied Soft Computing

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

A pruning approach to optimize synaptic connections and select relevant input parameters for neural network modelling of solar radiation

Sajid Hussain, Ali Al Alili *

Department of Mechanical Engineering, The Petroleum Institute, P.O. Box 2533, Abu Dhabi, United Arab Emirates

ARTICLE INFO

Article history:
Received 30 June 2015
Received in revised form 2 June 2016
Accepted 26 September 2016
Available online xxx

Keywords:
Solar radiation modelling
Neural network
Optimal brain surgeon
Optimization

ABSTRACT

The accurate modelling of solar radiation is important for many applications including agriculture and energy management. Previous research has widely focused on the development of artificial neural network (ANN) models for this task. Since the inputs to these models include different meteorological parameters, it is usually time consuming and costly to acquire these parameters. Therefore, selection of most relevant input parameters is a key step in the construction of these models. The overall goal of this research is to develop ANN models with less number of parameters and at the same time, high modelling accuracy. To this aim, an algorithm for network pruning, the optimal brain surgeon (OBS), is proposed to achieve two main objectives, selection of most relevant input parameters and optimization of the network structure at synaptic level. Four meteorological parameters, namely: temperature (T), relative humidity (RH), wind speed (WS), and sunshine duration (SSD) are used to model solar radiation, the global horizontal irradiation (GHI), over Abu Dhabi, the United Arab Emirates. The results show that the least relevant input parameter is RH with a contribution of 15.2% in the modelling process as compared to 24.8% for WS, 47.8% for T, and 54.7% for SSD. The parameter selection results coincide with recently used techniques in solar radiation research, J48 technique in Waikato Environment for Knowledge Analysis (WEKA) software and Automatic Relevance Determination (ARD) method. The modelling performance is compared with an all-connected ANN using all the four input parameters and ten hidden layer neurons with a total of 61 synaptic connections including biases. On the one hand, the proposed technique has successfully achieved an ANN with three inputs (T, WS, SSD), seven active hidden layer neuron, and 17 synaptic connections including biases. On the other hand, there are improvements observed in statistical evaluation metrics including mean absolute biased error (MABE), adjusted coefficient of determination ($\bar{R}^2$), Akaike information criterion (AIC), Akaike final prediction error (FPE), Rissanen's minimum description length (MDL), and cross entropy (CE).

1. Introduction

There have been increasing concerns over global warming, greenhouse gas emissions and climate change for the last couple of decades. One of the major culprits in global warming is the coal and oil fired conventional electricity generation. This encourages the use of renewable energy sources as alternatives. These sources include wind, solar, hydro, wave, tidal, biomass, geothermal, and others. Due to its strategic position in the solar belt, the United Arab Emirates (UAE) has abundance of sunshine hours throughout the year with the yearly average global horizontal irradiance (GHI) of about 6 kWh/day [1]. This gives the UAE an opportunity to develop renewable energy technologies in the region and to promote clean and green environment. As solar radiation is measured with different ground based sensors including pyranometers, pyroheliometers, and radiometers, etc. combined with data acquisition hardware and software, it is time consuming, cumbersome, and expensive to install these sensors everywhere. This gives more importance to solar radiation models. The accuracy of the solar radiation models directly relates to the accuracy of modelling power output from solar plants and that consequently affects the management and planning of future renewable energy systems.

There are many soft computing approaches to solar radiation modelling that include genetic programming (GP) [2], particle swarm optimization (PSO) [3], hybrid coral-reef algorithms [4], neuro-fuzzy [5], fire-fly [6] and more. Many researchers have used

* Corresponding author.
E-mail address: alialalili@pi.ac.ae (A.A. Alili).

http://dx.doi.org/10.1016/j.asoc.2016.09.036
1568-4946/

the artificial neural network (ANN) techniques to model solar energy. Multi-layer perceptron (MLP), with one hidden layer, is the most widely used ANN for solar radiation applications [7–11]. Recently, a very comprehensive review on the use of ANN for time series modelling in solar radiation applications is presented by Yadev & Chandel [12]. The authors have pointed out a need to identify the most relevant parameters to be used for ANN models and omit the non-relevant ones. One of the reasons is the capital costs of the required sensors that are directly related with the number of parameters measured. Another reason is the expensive data acquisition systems and the amount of hardware and software needed to process the data.

Different authors have addressed this problem in different ways. Yadav et al. [13] found that the modelling accuracy of the ANN varied with change in number of input parameters. They used Waikato Environment for Knowledge Analysis (WEKA) software for selection of most relevant input parameters to be used for solar radiation applications in different states of India. WEKA is a collection of machine learning programs together offering one single user interface. WEKA uses J48 algorithm (an implementation of c4.5 [14] algorithm in WEKA) to construct decision trees for classification. WEKA determines a rank for each input variable and ones with the lowest rank are discarded. The authors studied the selection of relevant input parameters from a set of candidates that included, T, minimum temperature ($T_{\min}$), maximum temperature ($T_{\max}$), altitude (Alt), SSD, latitude (Lat), and longitude (Lon). The authors found that the geographical parameters, Lat and Lon had the lowest rank for each month. The ranking of the parameters as determined by WEKA was T, $T_{\min}$, $T_{\max}$, Alt, SSD, Lat, and Lon from high to low priority. The authors simulated three different structures of ANN. ANN-1 with 7 inputs (T, $T_{\min}$, T_{max}, Alt, SSD, Lat, and Lon), ANN-2 with 5 inputs (T, $T_{\min}$, $T_{\max}$, Alt, and SSD) and ANN-3 with 4 inputs (T, $T_{\min}$, $T_{\max}$, and Alt). The output layer dimension for all the ANN architectures was 1. The authors also varied the number of hidden layers from 9 to 19 for ANN-1, from 8 to 18 for ANN-2 and from 7 to 18 for ANN-3. The authors found out that ANN-1 with the structure of 7-9-1 gave better performance with mean absolute percent error (MAPE) of 20.12%. The ANN-2 with the structure of 5-10-1 gave better performance with MAPE of 6.89% and the ANN-3 with the structure of 4-10-1 gave better performance with MAPE of 9.09%. It was therefore concluded that ANN-2 with Lat, Lon excluded gave better performance and ANN-1 with Lat, Lon included degraded the modelling performance. An interesting finding was the exclusion of SSD that degraded the performance of ANN-3 as compared to ANN-2 with SSD included. A second study was performed by the same authors using WEKA [15] for the same dataset. Two additional variables, extraterrestrial radiations (ER) and clearness index (CI) were added. The authors concluded lower ranks to clearness index, extraterrestrial radiation, Lat, and Lon. Five different ANN architectures were simulated and it was found that the ANN with the same input parameters as mentioned previously (ANN-2) gave better performance.

López et al. [16] used automatic relevance determination (ARD) method to select relevant input parameters in modelling direct normal solar irradiance for Desert Rock (USA). The ARD is a Bayesian approach for the ANN [17–19] that considers a priori probability distribution over the network weights and determines the most relevant input parameters by introducing a hyperparameter for each input unit of the ANN. The prior distribution of the network weights are controlled by the hyerparameters. The ANN is presented with the training data and the posterior distribution of weights and hyperparameters are calculated using Bayes rule. After the training is finished, the relevancy of the inputs is determined by the value of hyperparameter associated with that particular input. A comparatively large value of hyperparameter indicates less importance of the corresponding parameter. The authors used T, RH, pressure (p), WS, dew-point temperature (T_d), CI, precipitable water (w), and relative optical air mass (m_r) as candidate parameters for ANN with 2 and 10 hidden layer nodes. The ARD method found that the CI and m_r were the most relevant parameters to be used in their case. Bosch et al. [20] also used ARD and reported the estimation errors for different combinations of the input parameters over a mountainous area in Spain. The input parameters used were Lat, Lon, Alt, slope of ground station, azimuth, day of year (DOY), daily extraterrestrial solar radiation, and daily CI. The ANN models daily global irradiation. The authors found that Alt was a most relevant geographical parameter along with the DOY and daily CI. After the determination of important parameters, the authors used different ANN architectures with varying number of hidden units from 3 to 30. They trained every architecture 10 times to cater for random nature of weights initialization and to increase reliability of the results. The authors observed that the ANN with 14 hidden neurons gives better performance in this case and remains almost constant for higher number of hidden units. For the final ANN, the value of root-mean-squared error (RMSE) was recorded as 6.0% and mean biased error (MBE) as 0.2%. Yacef et al. [21] also used ARD for selecting an optimum combination of input parameters for ANN to estimate solar radiation in Al-Madinah, Saudi Arabia. The authors used T, RH, SSD and extraterrestrial irradiation (EI) as input parameters. The ARD determined EI and RH as comparatively less relevant parameters because of high values of hyperparameters associated with these variables. Therefore, the authors used T and SSD as two inputs to different ANN architectures with hidden neurons varying from 1 to 30 and found that the ANN with one hidden layer and two hidden neurons was the best architecture with RMSE of 8.4137%, MBE of 3.0658% and mean absolute error (MAE) of 5.9092%. Will et al. [22] used a niching genetic algorithm (NGA) for selection of input parameters to estimate solar radiations in El Colmenar (Tucuman, Argentina). The NGA is a modified form of the normal GA where all realizations of global and local optimums are found and for a same estimation error, the algorithm produces several combinations of input parameters. The NGA performs categorization of the problem, assigns limits to each input parameter, and calculates the distance in the search space. The authors used data from 14 different stations in Argentina and found out that T, RH, p, and SSD were the relevant parameters. The values of RMSE were reported to be 2.36 MJ/m^2 for 70 individuals/85 generations and 2.34 MJ/m^2 for 200 individuals/150 generations. Many authors on the other hand, used trial and error methods to determine the most relevant set of input parameters used for ANN to estimate solar radiations. Alam et al. [23] developed sixteen different ANN models to estimate solar radiation for different Indian stations. Each model used a different set of input parameters. A total of ten input parameters including Lat, Lon, Alt, time (t), the month of the year (MOY), T, RH, rainfall (RF), WS, and a net long wavelength (LW) were considered to develop these sixteen models. They also used ANN models with a varying number of hidden layers and trained for different seasons of the year. They found Lat, Lon, Alt, t, MOY, AT, RH, WS, and LW parameters were suitable for the summer season with ANN architecture containing 13 hidden neurons and coefficient of determination (R^2) value of 0.931. For the raining season, the best parametric combination was Lat, Lon, Alt, t, MOY, AT, RH, RF, WS, LW and ANN architecture with 11 hidden neurons and R^2 value of 0.907. For the winter season, the best parametric combination was Lat, Lon, Alt, t, MOY, AT, RH, LW and ANN architecture with 9 hidden neurons and R^2 value of 0.923.

Ahmet et al. [24] reported the effects of the number of input parameters on the estimation error. They changed the input layer parameters for ANN from 2 to 6 and recorded the relevant statistical error indicators for five different cities in Turkey. They concluded that training the ANN with all the six parameters (Lat, Lon, Alt, month, average cloudiness, and SSD) as inputs and 8 hidden