Energy Conversion and Management 138 (2017) 413–425

Contents lists available at ScienceDirect

Energy Conversion and Management

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

A novel hybrid model for hourly global solar radiation prediction using random forests technique and firefly algorithm

Ibrahim Anwar Ibrahim [a,*], Tamer Khatib [b]

[a] Department of Electrical, Electronic and Systems Engineering, Universiti Kebangsaan Malaysia, 43600 Bangi, Selangor, Malaysia
[b] Energy Engineering and Environment Department, An-Najah National University, 97300 Nablus, Palestine

ARTICLE INFO

Article history:
Received 14 December 2016
Received in revised form 13 January 2017
Accepted 4 February 2017

Keywords:
Solar radiation
Prediction
Random forests technique
ANN
Optimization
Firefly algorithm

ABSTRACT

Reliable knowledge of solar radiation is an essential requirement for designing and planning solar energy systems. Thus, this paper presents a novel hybrid model for predicting hourly global solar radiation using random forests technique and firefly algorithm. Hourly meteorological data are used to develop the proposed model. The firefly algorithm is utilized to optimize the random forests technique by finding the best number of trees and leaves per tree in the forest. According to the results, the best number of trees and leaves per tree is 493 trees and one leaf per tree in the forest. Three statistical error values, namely, root mean square error, mean bias error, and mean absolute percentage error are used to evaluate the proposed model for the internal and external validation. Moreover, the results of the proposed model are compared with conventional random forests model, conventional artificial neural network and optimized artificial neural network model by firefly algorithm to show the superiority of the proposed hybrid model. Results show that the root mean square error, mean absolute percentage error, and mean bias error values of the proposed model are 18.98%, 6.38% and 2.86%, respectively. Moreover, the proposed random forests model shows better performance as compared to the aforementioned models in terms of prediction accuracy and prediction speed.

1. Introduction

The availability of solar radiation data in a small scale becomes extremely important for modeling, sizing and controlling of solar energy systems and other applications such as agricultural productions, hydrological and ecological studies. Solar radiation is usually measured using many types of instruments at a particular location. These instruments are considered as the most accurate to meaure the solar radiation data. However, the cost of these instruments including calibration and maintenance cost is slightly high. Therefore, the availability of these data for all sites is limited in many meteorological stations around the world. Moreover, most of the recorded datasets of solar radiation suffer of missing data during outages. Thus, there is a need for prediction more so as to restore these datasets and to generate solar radiation data series at location where measurement devices are missing.

Based on that, many research works have been done before to model the global solar radiation. Several models including the empirical models [1,2], satellite-derived model [3], regression models [4] and stochastic algorithm model (Markov chain) [5] have been developed for modeling global solar radiation.

In general, empirical models are widely used to model global solar radiation by correlating the solar radiation with different measured meteorological data and geographical coordinates. However, the accuracy of such models is questionable, especially when dealing with a highly uncertain data. In the meanwhile, the application of satellite-derived models is promising for predicting global solar radiation into a large region using satellite images. This method is costly and suffers of the lack of historical meteorological datasets because it is relatively new. On the other hand, regression models are good in predicting monthly averages of solar radiation. However, recent challenges to solar energy systems show a dire need for modeling solar energy systems on hourly basis and consequently prediction of hourly solar radiation becomes a must for development in recent research work in the field of solar energy [6]. These models have a low performance when modeling global solar radiation on a long term basis, and they are not recommended when there are missing data.

Due the necessity of accurate and reliable modeling of hourly global solar radiation, machine learning and artificial intelligence (AI) techniques have been broadly utilized due to their powerful techniques in predicting hourly solar radiation considering the

* Corresponding author.
E-mail addresses: ibrahim.a.ibrahim@hotmail.com, ibrahim.a.ibrahim@ieee.org (I.A. Ibrahim), t.khatib@najah.edu (T. Khatib).

http://dx.doi.org/10.1016/j.enconman.2017.02.006
0196-8904/

Nomenclature

Abbreviations

RFs	random forests
ANN	artificial neural network
FFA	firefly algorithm
OOB	Out-Of-Bag
VI	variable importance
RMSE	root mean squared error
MAPE	mean absolute percentage error
MBE	mean bias error

Symbols

$\beta^{c(tr)}$	OBB samples for a specific tree in the RFs
tr	tree number (1, 2,...,T) in the RFs
T	the total number of trees in the RFs
$c_a^{(tr)}$, $c_{a,\pi z}^{(tr)}$	the predicted classes for each sample in the tree before and after altering the variable in the RFs
x_a	the sample value in the training stage of the RFs
L_b	the true label in the training stage of the RFs
I	importance function that got based on the values of L_b
a	the number of samples per leave in tree in the RFs
b	the number of samples per tree in the RFs
r_{ij}	the distance between any two fireflies i and j at positions x_i and x_j
γ	affixed light absorption coefficient
x_k^i, x_k^j	the kth components of the ith and jth firefly positions within the search-space
d	the dimensionality of the problem
α	is a step size scaling factor
β	the variation of attractiveness
β_0	the attractiveness at $r = 0$
ϵ_i	a randomization parameter
I_{P_m}	the predicted value
I_m	the target value
n	the number of observations

uncertainty of such data. Artificial neural networks (ANNs) have been utilized many times in the literature for this purpose [6–10]. A detailed review on solar radiation prediction using ANN models can be found in [11]. In [12], six ANN-based models to predict daily global solar radiation at a location in Saudi Arabia are presented. A different combination of input variables considering sunshine duration, ambient temperature, relative humidity and the number of day in the year is utilized for these models. The results show that the sunshine duration and ambient temperature play very important role to gain high accurate results. In [13], a hybrid model to predict monthly mean daily global solar radiation in Saudi Arabia is developed as well. Here, the number of hidden neurons in the ANN model is optimized using particle swarm optimization (PSO) technique in order to improve model's accuracy. The model uses sunshine duration, the number of month in the year, latitude and longitude as inputs of the developed hybrid model. From the results presented in this research, the developed hybrid (ANN-PSO) model showed better performance than the conventional ANN model. In [14], an adaptive neuro-fuzzy inference system (ANFIS) model is used to predict daily global solar radiation based on the day of the year at a specific location in South-Khorasan. The ANFIS is considered as a hybrid model which merges the ANNs with the knowledge representation of the fuzzy logic. The number of day is the only input used in this model. In general, sunshine ratio, ambient temperature as well as relative humidity are the most correlated meteorological variables to solar radiation. Moreover, ANNs are the most powerful tools for predicting hourly solar radiation. These results were supported in [15,16] as well.

Nevertheless, other machine learning techniques are also implemented for this purpose. Recently, support vector machine (SVM) technique has been widely used to predict solar radiation [17]. In [18], a SVM model is employed to predict daily and monthly global solar radiation in Isfahan, Iran. The radial basis function (SVR-rbf), and polynomial basis function (SVR-poly) are utilized for this purpose. Two inputs are used in this research. These inputs are; sunshine hours as well as maximum possible sunshine hours. The superiority of the proposed SVM model has been proved as compared to the empirical and the PSO-based models. In [19], a new hybrid model using SVM with Wavelet Transform (WT) algorithm is presented to predict daily and monthly global solar radiation in Iranian coastal city. Measured data that contain daily global solar radiation, sunshine hours, average, maximum and minimum ambient temperature, relative humidity and water vapor pressure are utilized. The capability of the hybrid (SVM–WT) is verified by compering its performance with data that are obtained by ANN, genetic programming (GP) autoregressive and moving average (ARMA) models.

Though, ANNs and SVM are powerful prediction models for such a task, there are some novel techniques such as random forests technique that can predict global solar radiation in a more accurate way. Random Forests (RFs) technique is a good option for predicting hourly global solar radiation. RFs algorithm is a novel and ensemble machine learning technique which incorporates the feature selection. Moreover, RFs have several merits over other modeling approaches. In which, RFs can handle with both continuous and discrete variables, RFs algorithm does not over fit as a predictor, and run fast and efficiently when handling large datasets [20]. In addition, RFs algorithm has only two hyper-parameters which are; the number of trees in the forest, and the number of splits in the subset at each node (number of leaves per tree) [21]. In the meanwhile, some studies are presented for predicting hourly and daily solar radiation based on RFs. These studies do not mention any use of the optimization techniques to optimize the parameters of the RFs (number of trees in the forest and number of leaves per tree). In [22], the authors used daily meteorological data, solar radiation and air pollution index for three sites in China to develop a RFs model for predicting daily solar radiation. Here, the default number of trees and leaves per tree (500 trees and 5 leaves per tree) are used. The results show that the performance of the RFs model is better than the empirical methodologies (linear, exponential and logarithmic models) in predicting the daily solar radiation. In addition, in [23], the authors have used a RFs model for predicting hourly and daily solar radiation. In this research, the values of the number of trees and leaves per tree are not mentioned. Moreover, the authors in [24] employed different machine learning techniques to predict hourly solar radiation. In this study, the authors assumed different values of the number of trees and leaves per tree (10, 50, 100 and 300 trees, and 5 and 20 leaves per tree) in the proposed RFs model. Then, the best values of the number of trees and leaves are selected based on the performance of the model.

In general, optimal number of trees and number of leaves per tree should be found so as to assure accurate prediction of solar