University of Malaya

From the SelectedWorks of Faculty of Engineering University of Malaya

March 1, 2015

A new hybrid support vector machine-wavelet transform approach for estimation of horizontal global solar radiation

Available at: https://works.bepress.com/facultyofengineering_universityofmalaya/134/
A new hybrid support vector machine–wavelet transform approach for estimation of horizontal global solar radiation

Kasra Mohammadi a, Shahaboddin Shamshirband b,*, Chong Wen Tong c, Muhammad Arif d, Dalibor Petković e, Sudheer Ch e

a Faculty of Mechanical Engineering, University of Isfahan, Isfahan, Iran
b Department of Computer System and Technology, Faculty of Computer Science and Information Technology, University of Malaya, Kuala Lumpur, Malaysia
c Department of Mechanical Engineering, Faculty of Engineering, University of Malaya, Kuala Lumpur, Malaysia
d University of Nis, Faculty of Mechanical Engineering, Department for Mechatronics and Control, Aleksandra Medvedeva 14, 18000 Nis, Serbia
e Department of Civil and Environmental Engineering, ITM University, Gurgaon 120017, India

ARTICLE INFO

Article history:
Received 11 October 2014
Accepted 16 December 2014
Available online 5 January 2015

Keywords:
Global solar radiation estimation
Support vector machine
Wavelet transform algorithm
Meteorological parameters
Statistical indicators

ABSTRACT

In this paper, a new hybrid approach by combining the Support Vector Machine (SVM) with Wavelet Transform (WT) algorithm is developed to predict horizontal global solar radiation. The predictions are conducted on both daily and monthly mean scales for an Iranian coastal city. The proposed SVM-WT method is compared against other existing techniques to demonstrate its efficiency and viability. Three different sets of parameters are served as inputs to establish three models. The results indicate that the model using relative sunshine duration, difference between air temperatures, relative humidity, average temperature and extraterrestrial solar radiation as inputs shows higher performance than other models. The statistical analysis demonstrates that SVM–WT approach enjoys very good performance and outperforms other approaches. For the best SVM–WT model, the obtained statistical indicators of mean absolute percentage error, mean absolute bias error, root mean square error, relative root mean square error and coefficient of determination for daily estimation are 6.9996%, 0.8405 MJ/m², 1.4245 MJ/m², 7.9467% and 0.9086, respectively. Also, for monthly mean estimation the values are 3.2601%, 0.5104 MJ/m², 0.6618 MJ/m², 3.9935% and 0.9742, respectively. Based upon relative percentage error, for the best SVM–WT model, 88.70% of daily predictions fall within the acceptable range of –10% to +10%.

1. Introduction

Developing solar energy technologies are being substantially increasing to supply the energy demand and provide sustainability in many locations across the globe. Solar energy systems can be employed for variety of purposes such as heating, cooling and providing electricity. One of the greatest applications of solar energy technologies is in the isolated regions where there is no accessibility to accurate and reliable measured solar radiation information. The scarcity of real solar data is chiefly contingent upon several factors such as paucity of solar radiation measurement equipment as well as fiscal issues [1–3]. Nevertheless, despite such unavailability, other meteorological elements including sunshine hours, ambient temperature, relative humidity, pressure, etc. are widely measured in most of sites owing to their significant applications in various fields. Therefore, several models have been developed to estimate horizontal global solar radiation based upon a series of commonly available meteorological and geographical parameters including sunshine duration, ambient temperatures, relative humidity, water vapor and sea level pressures, cloud cover, altitude, latitude and longitude and extraterrestrial radiation [4–20]. In fact, finding an appropriate relationship between horizontal global solar radiation and one or more input variables has been a serious challenge in the realm of solar radiation simulation.

Although a vast number of models have been proposed to estimate the global solar radiation, developing proper algorithms and approaches to achieve further reliability, accuracy and convenience in computational process is still highly demanding. Over the past years, various artificial intelligence and computational intelligence techniques have been applied by researchers as especially efficient approaches for the problem of global solar radiation prediction in many locations across the globe.

Bosch et al. [21] carried out an investigation to estimate daily global solar radiation using artificial neural networks (ANN) over
a mountainous region situated in South East of Spain. Their results demonstrated that ANN can be considered as an effective technique with easy usage for estimation of solar radiation in complex mountain terrains. Rehman and Mohandes [22] used ambient temperature and relative humidity as inputs to predict global solar radiation for Abha city in Saudi Arabia using artificial neural network (ANN) technique. The obtained results illustrated that ANN is favorably capable to estimate global solar radiation based upon these two parameters. Benghanem et al. [23] developed six ANN-based models to estimate horizontal global solar radiation at Al-Madinah in Saudi Arabia. They utilized different combinations of input parameters consisting sunshine hours, ambient temperature, relative humidity and the day of year. Their results showed that a higher accurate model is dependent upon sunshine duration and air temperature. Rahimikhoob [24] applied ANN technique to estimate global solar radiation based on ambient temperature in a semi-arid environment. The ANNs were trained based upon maximum and minimum air temperatures and extraterrestrial radiation. Then the ANNs were compared with a traditional temperature-based empirical model. It was found that estimation of global solar radiation using ANN technique offers higher accuracy. Solmaz and Ozgoren [25] utilized ANN approach for prediction of hourly global solar radiation in six different provinces of Turkey. For this purpose, they developed two ANN-based models using six input parameters. Bezuglyi et al. [26] utilized particle swarm optimization (PSO) technique to develop some new sunshine-based models for estimation of monthly mean global solar radiation in 17 Iranian cities. Their results illustrated that for most of the cities the new models developed based on PSO have higher performance than the existing models. Chen et al. [27] examined the possibility of employing the Support Vector Machines (SVMs) for estimating the monthly mean global solar radiation utilizing maximum and minimum air temperatures at Chongqing station, China. They applied three different functions of SVM and found that the developed SVM model using polynomial kernel function shows superiority over other SVM models. Ozgoren et al. [28] developed an artificial neural network (ANN) model on the basis of multi-nonlinear regression (MNLIR) method for estimation of the monthly global solar radiation over Turkey. They used various variables and then employed the Stepwise MNLIR method to determine the most proper input values. Their results showed that the ANN model can predict the values with acceptable errors compared with the actual data. Yacif et al. [29], in a comparative study, assessed the performance of Bayesian Neural Network (BNN) in comparison with classical Neural Network (NN) and empirical models to estimate daily global solar irradiation at Madinah, Saudi Arabia. They used four different input elements and found that BNN enjoys higher capability for estimation of solar radiation. Rodriguez et al. [30] developed an optimized artificial neural network (ANN) model to calculate daily global solar radiation over Andalusia, Spain. In the developed model, they utilized both clear-sky estimates and satellite images as input elements and also applied genetic optimization to select the optimal inputs. They found that the predicted values by the model are relatively precise. Ramadani et al. [31] employed support vector regression (SVR) technique to develop a model for prediction of global solar radiation in Tehran, Iran. They used two SVR models of radial basis function (SVR-rbf) and polynomial function (SVR-poly). Their results showed the superiority of SVR-rbf technique. Chen et al. [32] evaluated the possibility of applying Support Vector Machine (SVM) for prediction of daily global solar radiation in Liaoning province in China. They developed seven sunshine duration-based SVM models and compared their performance with five empirical models. Their results indicated that all SVM models outperform the empirical models remarkably. Ramadani et al. [33] performed a comparative investigation between fuzzy linear regression (FLR) and support vector regression (SVR) techniques to predict global solar radiation in Tehran, Iran. They found that, owing to substantially lower errors, SVR-rbf approach enjoys superior performance compared to FLR. Chen et al. [34] appraised the transferability of Support Vector Machines (SVM) to estimate global solar radiation using ambient temperature in subtropical zone in China. They found that global solar radiation at one site can be estimated acceptably by SVM model established for another site. Also, the estimation precision is influenced by the distance and temperature difference between two locations and altitude.

Additionally, many authors have aimed at achieving further accuracy in estimating the solar radiation by hybridizing different approaches. Basically, the fundamental objective of combining different approaches is the utilization of specific nature of each technique to obtain further accuracy for estimation of solar radiation.

Wu and Chan [35] combined the Autoregressive and Moving Average (ARMA) model with the controversial Time Delay Neural Network (TDNN) for prediction of hourly solar radiation. The achieved results revealed that the hybrid model has higher capability compared to both ARMA and TDNN. Bharadwaj et al. [36] proposed a hybrid approach which includes hidden Markov models and generalized fuzzy models to estimate solar irradiation in India. They assessed the influence of different meteorological parameters for estimation of solar radiation using the developed model. Their results showed that the predicted values by the proposed model are in a favorable agreement with measured data. Mostafavi et al. [37] developed a hybrid approach by combining Genetic Programming (GP) with simulated annealing (SA) for estimating the global solar radiation. They also performed a sensitivity analysis to assess the influence of the different meteorological parameters on solar radiation estimation. Their results showed that the suggested model provide precise predictions. Hung et al. [38] developed a hybrid Auto Regressive and Dynamical System (CARDS) model to forecast hourly global solar radiation in Mildura, Australia. Their results indicated that the CARDS model can forecast hourly solar radiation favourably. Salcedo-Sanz et al. [39] assessed the capability of a novel Coral Reefs Optimization–Extreme Learning Machine (CRO–ELM) algorithm to predict the global solar radiation at Murcia (southern Spain) using different meteorological data. They concluded that the CRO–ELM approach can predict the daily global radiation accurately with further preciseness than the classical ELM and the Support Vector Regression algorithm. Wu et al. [40] developed a genetic algorithm combing multi-model framework to predict solar radiation. By comparing the prediction performance of the proposed technique with some other algorithms they found higher accuracy and consistency for their approach.

In this paper, the Support Vector Machines (SVMs) and Wavelet Transform (WT) algorithm are combined to propose a new hybrid approach to predict horizontal global solar radiation. The primary aim is achieving further accuracy and reliability in estimations by taking the advantages of both approaches. For this purpose, long-term measured databases consisting horizontal global solar radiation and different meteorological elements for a city situated in south costal part of Iran have been used. The motivation behind this research work is mainly twofold. First, there is a special need to reliable and accurate solar information in various applications such as the design and simulation of solar energy technologies, agricultural production, irrigation management and water resources allocation. Furthermore, despite the substantial solar energy potential, precise long-term measured solar data is unavailable in the vast neighboring region around the considered case study. The merit of proposed hybrid SVM–WT model is assessed thoroughly in terms of both daily and monthly mean daily estimation using various reliable and widely-known statistical indicators to draw more conclusive conclusions. Owing to significance of
proper input parameters, various combinations of meteorological parameters are considered as the required inputs to estimate the horizontal global solar radiation. The accuracy of SVM–WT is verified by comparing to previously established techniques including artificial neural network (ANN), Genetic Programming (GP) and Autoregressive-Moving-Average (ARMA).

The organization of the remaining part of this paper is as follows: Section 2 explains the data bases utilized for the analysis. Section 3, which offers the utilized methodology, is divided into two parts: while in Section 3.1 the support vector machine is described, in Section 3.2 the Wavelet Transform (WT) algorithm is explained. The comparative results and discussion are brought forward in Section 4. Finally, the conclusions are presented in Section 5.

2. Data description and processing

To assess the merit of the proposed hybrid SVM–WT approach, the long-term measured databases for port of Bandar Abbas located in Iran have been used in the simulation. Bandar Abbas, the capital city of the Hormozgan province, is situated in the southern part of Iran at geographical location of 27°13′ N and 56°22′ E, and its elevation is 9.8 m above the sea level. Long warm season and cool short season are the climatic characteristics of the region. Basically the region is a desert zone with extremely low level of atmospheric precipitation [41]. Based upon Köppen classification the climate condition of Bandar Abbas is categorized as BWh which relates to arid desert hot [42]. For this study, long-term measured data including the daily global solar radiation on a horizontal surface ($H$), sunshine duration ($N$), average ambient temperature ($T_{mean}$), maximum ambient temperature ($T_{max}$), minimum ambient temperature ($T_{min}$), relative humidity ($R_h$) and water vapor pressure ($V_p$) provided by Iranian Meteorological Organization for the period of 14 years from January 1992 to December 2005 were utilized.

Generally, precise estimation of global solar radiation models depends upon the acceptability of long-term data at a desired location. It was foreseen that the record lengths of data utilized for this research is adequately long to simulate global solar radiation models in the considered case studies. The precision of the models developed to estimate the solar radiation is also chiefly influenced by the quality of raw data utilized. The data cleaning procedure generally aims at enhancing the data quality by checking and filtering them from any uncertainty or erroneous. In horizontal global solar radiation data used in this study, there were some missing and also unreliable values possibly due to instruments' malfunction. To overcome this issue and enhance the quality of raw data, the following procedure was applied in present study:

1. To identify the incorrect global solar radiation values, the daily clearness index ($k_t$) was computed and the values which were out of range of 0.015 < $k_t$ < 1 were eliminated [43,44].
2. A month containing more than 5 days missing or unreliable global solar radiation values was completely extracted from the databases. Additionally, for a month with less than 5 days the missing or inaccurate values were substituted by proper values obtained using interpolation [43,44].

It is worth mentioning that clearness index ($k_t$) is the ratio between global solar radiation incident on a horizontal surface ($H$) to extraterrestrial solar radiation on a horizontal surface ($H_0$). $H_0$, in any geographical location is a constant value for each specific day, irrespective of change of year. However, solar attenuation occurs as radiation passes through the atmosphere due to some atmospheric phenomena such as aerosol extinction, cloud extinction, Rayleigh scattering and so on. Therefore, in the available solar radiation data all values of $H$ should be smaller than $H_0$, which means $k_t < 1$.

To model the global solar radiation via the developed approach, different combinations of data consisting horizontal global solar radiation ($H$), relative sunshine duration ($N$) defined as the ratio of sunshine duration ($N$) to the maximum possible sunshine duration ($N$), difference between maximum and minimum ambient temperatures ($T_{max}-T_{min}$), relative humidity ($R_h$), water vapor pressure ($V_p$), average ambient temperature ($T_{avg}$) and extraterrestrial global solar radiation on a horizontal surface ($H_0$) are used as inputs. It is worth mentioning that the values of $H_0$ and $N$ were computed by the equations presented in the Appendix A.

The available data for this study are divided into two parts of training and testing data sets. The first set of 10 years from 1992 to 2001 was used for training while the second set of 4 years from 2002 to 2005 was utilized for testing. In fact, on daily basis estimation the system is trained using 3285 days and tested by 1460 days. Also, for monthly mean daily basis calculation 108 months data is used for training and 48 months data is applied for testing.

Some descriptive statistics including mean values, standard deviation, minimum and maximum values as well as the range of the 6 meteorological input parameters used to train the techniques are listed in Tables 1 and 2 for daily and monthly mean daily data sets, respectively.

3. The hybrid SVM–WT model for solar radiation estimation

In this study, the Support Vector Machine (SVM) and Wavelet Transform (WT) Algorithm are combined to propose a new hybrid approach named SVM–WT to estimate the horizontal global solar radiation. This section offers a brief explanation of the support vector machine and Wavelet Transform (WT) algorithm as well as the methodology conducted to predict global solar radiation via the proposed hybrid SVM–WT approach.

3.1. Support Vector Machine (SVM)

Support Vector Machine (SVM) is a well-known machine learning approach which has recently applied in the variety of fields such as computing, hydrology and environmental researches [45–49]. It has mainly utilized in pattern recognition, forecasting, classification and regression analysis. It has been proved that its applications show superior performance compared to prior developed methodologies such as neural network and other conventional statistical models [50–54]. The detail of theory and evolution of SVM developed by Vapnik’s can be found in [55,56].

SVM was developed according to the statistical machine learning development as well as structural risk minimization to reduce the upper bound generalization error compared to local training error, which is known technique in previously used machine learning methodologies. The mentioned technique proved advantages over other machine learning algorithms. Additional advantages provided in this methodology include: (1) applying high

<table>
<thead>
<tr>
<th>Table 1</th>
<th>Descriptive statistics for daily input variables utilized as training data.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$n$</td>
</tr>
<tr>
<td>Min</td>
<td>0.02</td>
</tr>
<tr>
<td>Max</td>
<td>0.99</td>
</tr>
<tr>
<td>Mean</td>
<td>0.71</td>
</tr>
<tr>
<td>St.dev.</td>
<td>0.23</td>
</tr>
<tr>
<td>Range</td>
<td>0.97</td>
</tr>
</tbody>
</table>
Table 2  Descriptive statistics for monthly mean daily input variables utilized as training data.

<table>
<thead>
<tr>
<th>Model</th>
<th>( \mu/N )</th>
<th>( T_{max}-T_{min} (\degree C) )</th>
<th>( R_o )</th>
<th>( V_p/(m/s) )</th>
<th>( T_{avg} (\degree C) )</th>
<th>( H_o (\text{MJ}/\text{m}^2) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min</td>
<td>0.45</td>
<td>5.35</td>
<td>48.0</td>
<td>10.14</td>
<td>14.93</td>
<td>21.63</td>
</tr>
<tr>
<td>Max</td>
<td>0.87</td>
<td>15.98</td>
<td>80.0</td>
<td>40.28</td>
<td>35.77</td>
<td>40.77</td>
</tr>
<tr>
<td>Mean</td>
<td>0.71</td>
<td>10.17</td>
<td>67.12</td>
<td>23.95</td>
<td>26.73</td>
<td>32.32</td>
</tr>
<tr>
<td>St. dev.</td>
<td>0.10</td>
<td>2.12</td>
<td>6.16</td>
<td>3.33</td>
<td>6.07</td>
<td>6.87</td>
</tr>
<tr>
<td>Range</td>
<td>0.42</td>
<td>10.63</td>
<td>32.0</td>
<td>30.14</td>
<td>20.84</td>
<td>19.14</td>
</tr>
</tbody>
</table>

Table 3  The studied models with different input parameters.

<table>
<thead>
<tr>
<th>Model</th>
<th>Input element</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>( \mu/N ), ( T_{max}-T_{min} ), ( R_o ), ( V_p )</td>
</tr>
<tr>
<td>2</td>
<td>( \mu/N ), ( T_{max}-T_{min} ), ( R_o ), ( V_p ), ( T_{avg} )</td>
</tr>
<tr>
<td>3</td>
<td>( \mu/N ), ( T_{max}-T_{min} ), ( R_o ), ( V_p ), ( T_{avg} ), ( H_o )</td>
</tr>
</tbody>
</table>

dimensional spaced set of kernel equations, which discreetly include non-linear transformation; thus, there is no assumption in functional transformation which makes data linearly separable indispensable and (2) unique solution due to the convex nature of the optimal problem.

SVM functions according to Vapnik's theory are represented in Eqs. (1)–(4). \( R = (x_i, d_i) \) is used to assume a set of data points. Where \( x_i \) indicates the input space vector of the data sample and, the target value and data size is defined as \( d_i \) and \( n \), respectively. SVM approximates the function as represented in Eqs. (1) and (2).

\[
f(x) = w \phi(x) + b
\]

\[
R_{SVM}(C) = \frac{1}{2} \|w\|^2 + \frac{1}{n} \sum_{i=1}^{n} \xi_i
\]

In Eq. (1), \( \phi(x) \) indicates high dimensional space characteristic that mapped the input space vector \( x \), \( w \) and \( b \) are normal vector and scalar, respectively. In addition, \( \sum_{i=1}^{n} \xi_i \) stands for the empirical error, risk. Factors \( b \) and \( w \) are measured by minimization of regularized risk equation following by introduction of positive slack variables \( \xi_i \) and \( \xi^*_i \) that indicate upper and lower excess deviation [57]:

\[
\text{Minimize} \quad R_{SVM}(w, \xi) = \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{n} (\xi_i + \xi^*_i)
\]

Subject to \( d_i - w \phi(x_i) + b \leq \xi + \xi^*_i \), \( w \phi(x_i) + b - d_i \leq \xi + \xi^*_i \), \( \xi, \xi^*_i \geq 0, i = 1, \ldots, l \)

where \( \frac{1}{2} \|w\|^2 \) is the regularization term, \( C \) represents the error penalty feature utilized to control the trade-off between the empirical error (risk) and regularization term, \( \epsilon \) represents the loss function associated to approximation accuracy of the trained data point and the number of factors in the training data set defined as the \( I \). Optimality constraints and Lagrange multiplier which can be used to solve Eq. (1) are consequently obtained using a generic function as follow:

\[
f(x, \alpha, \alpha^*) = \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) \phi(x_i) + b
\]

In Eq. (4), \( \phi(x_i) = \phi(x_i) \phi(x_i) \) and the term \( K(x_i, x_j) \) is defined as the kernel function, which is dependent on the two inner vector \( x_i \) and \( x_j \) in the feature space \( \phi(x_i) \) and \( \phi(x_j) \), respectively.

The most significant aim of SVMs is to complete data correlation using non-linear mapping. It could be possible to develop a non-linear learning machine known as a direct calculation method of a kernel equation, denoted by \( K \), if a method of calculation of the inner product in a feature space is available in a straight line as a function to the original input points. The flexibility of SVM can be recognized through using kernel functions by adopting the data to a higher-dimensional feature space. The results in the higher-dimensional feature space stand for the results of the original, lower-dimensional input space.

Sigmoid, linear, polynomial and radial basis functions are the four basic kernel functions which are provided by SVMs. Among all, radial basis function (RBF) has been considered as the best kernel features due to its computationally effectiveness, reliability, simplicity, ease of adaption for optimization as well as its adaptability in handling factors which are more complicated [58–60].

Only the solution of a set of linear functions is required for the training of RBF kernel equation rather than the lengthy and complicated demanding quadratic programming problem [61].

---

**Fig. 1.** Flow chart of the proposed wavelet-based parameter determination approach for the SVM classifier.
Accordingly, the radial basis equation with parameter $\sigma$ is adopted. The non-linear radial basis kernel function is defined as:

$$K(x_i, x_j) = \exp\left(-\gamma ||x_i - x_j||^2\right)$$

(5)

where $x_i$ and $x_j$ are vectors in the input space, i.e., vectors of features computed from training or test samples. In addition, the accuracy of predictions using RBF kernel function depends on the selection of its three factors ($\gamma$, $\zeta$, and $C$). In this study, the optimal values of these factors are established using wavelet transform algorithm, which is described in the following sub-section.

3.2. Discrete wavelet transform

Wavelet Transform (WT) is a signal processing algorithm developed from Fourier transform. It represents a mathematical expression to decompose time series frequency signal into different components. One of its advantages over Fourier transform is the perfect analysis of the resulting decomposed components with well-scaled resolution. This is beneficial in terms of enhancing the capacity of the study model since it captures the requisite information at various levels [62]. It is suitable for analyzing data in frequency and time domain due to its capability of extracting data from non-periodic and transient signal; therefore, it is really useful in time-frequency localization [63]. Wavelet Transform (WT) has many useful basis functions, from which one can be selected depend on the signal been analyzed. Recently, this technique has been applied in a series of engineering applications [64–68]. Continuous wavelet transform (CWT) of a signal $f(t)$ is a time-scale technique of signal processing that can be defined as

![Fig. 2. Scatter plots of the measured values versus predicted daily global solar radiation via: (a) SVM-WT (3), (b) ANN (3), (c) GP (3) and (d) ARMA (3).](http://www.sciencedirect.com/science/article/pii/S0196890414010899)