Solar radiation prediction based on random forest of feature-extraction

J Liu¹, M Y Cao², D Bai¹ and R Zhang¹
¹College of Information and Control Engineering, Shenyang Jianzhu University, Shenyang 110168, China
²Qianju Technology Development Co., LTD, Beijing 100094, China

E-mail: jeanliu10@163.com; 1582729192@qq.com

Abstract: In order to use solar energy efficiently, the accurate prediction of solar radiation is extremely important. The aim of this paper is to discuss and build the solar radiation prediction model, and an efficient method based on improving random forest is proposed in here. The random forest regression model is created for predicting solar radiation. The nine variables are taken as input of the original model, and the solar radiation as the output. The original data is normalized using Min-Max method and is mapped to the values in the range of zero to one. The large amount of radiation data has been dealt with univariate feature extraction. The classifier of the random forest tree is constructed by characteristic factors that can influence solar radiation, and the optimal parameters of the model are selected by the OOB error analysis. The relationship between M-try and OOB can be obtained by computing of the R platform. Hence, a prediction model of improved random forest solar radiation based on feature extraction has been described. Under the specific area of solar radiation forecast, the studies have been done, and the solar radiation is analyzed and compared with the different prediction methods. The simulation results prove the validity of the improved method; the root mean square error is reduced and the precision of prediction is improved. It is significant for the prediction of the amount of solar radiation under complex environment and the effective use of photovoltaic power generation.

1. Introduction
With the increase of environmental pollution, in recent years, solar energy as optimal clean new energy resources have attracted more and more attention. So, how to forecast accurately the solar radiation is particularly critical for the solar energy utilization [1-8].

In the first instance, Angstrom proposed the climatology method which calculates monthly total solar radiation through utilization of solar radiation under clear day and the sunshine percentage [9]. Besides, Zhou Jin and some others fitted the coefficient of the correction formula according to the solar radiation data from 69 meteorological stations and the sunshine material [10]. In recent years, more and more intelligent prediction methods have been applied in the solar radiation prediction. For example, Zhang Jingjing [11] established a BP neural network simulation model to predict the solar radiation, which was trained by using the actual data and realized the prediction of future solar radiation. Many researchers have proposed solar radiation prediction based on SVM and SVR. Such as Zen Jianwu and Qiao Weihuo built a Least Squares Support Vector Machine (LS-SVM) model to predict the amount of solar radiation in the short term [12]. As for the strong non-linearity and stationarity of the daily solar radiation
energy time series, Wang Jianping applied neural network combining with the Fourier transform and wavelet multi-scale decomposition to equivalently establish daily solar radiation energy wavelet process neural network real-time online prediction model [13], which has more degrees of freedom and a stronger nonlinear approximation and random estimation ability, especially influenced by many factors such as atmospheric condition, solar elevation, Earth-Sun distance, the sunshine time, the altitude, terrain and landforms, the obstacles and so on. This paper mainly studies the solar radiation prediction problem which is based on the random forest algorithm, because the accurate and effective solar radiation prediction method has vital actual significance to develop solar energy resources and improve the efficiency of photovoltaic power generation effectively [14, 15].

The contents of the paper consist of five sections. In section 2, the random forest algorithm is introduced. The solar radiation prediction model is established based on random forest algorithm. In section 3, we analyze the parameters with the amount of solar radiation and then propose the solar radiation prediction model based on feature extraction of random forest. In section 4, the improved method based on the characteristics processing has been compared with the original one, in case of the monthly average of solar radiation value. In section 5, the research conclusions are presented.

2. Random forest algorithm
Random forest (referred to as RF) [16-18] is composed of many decision tree model \( h(X; \theta_k) \), \( k = 1,2,\ldots,K \) which is a combined regression model. Where \( \theta_k \) are independent and identically distributed random vectors. Let \( X \) be the input and \( Y \) be the output variables, then \( (X, Y) \) constitute the original data set. The construction of each decision tree includes two processes of sampling and complete split. The first step constitutes of two random sampling processes, in which the input data of random forests are separately made for row sampling and column sampling. In row sampling, taking back method is used, and namely repeated samples may be contained in the sampled data sets. Assume that the input sample is \( N \), then sampling samples is \( n (n< N) \). So, in training, the input samples of each tree are not all the samples. On the column sample, we randomly select \( m \) \((m \ll M) \) from the \( M \) attributes. Then using the split of the data sampling method, we establish decision tree.

Since the two random samples ensure the randomness of the aforementioned process, then there is no need for pruning, and it will not appear over-fitting [19]. Random forest method does not generate over-fitting as the decision tree increases. The generalization error tends to be upper bound. We give the theoretical basis as the following.

Definition 1: Let \( I(.) \) and \( aV_i(.) \) are two functions for expressing average value. \( mg(x, y) \) is the margin function of the average values. \( Y \) is correctly classified for the vector \( X \). The last term represents maximum value of votes that vector \( X \) is classified by mistake. The marginal function reflects the degree of difference between the two. The higher the value, the combination classifier performs better. It is represented as follows:

\[
mg(x, y) = aV_i I(h(x; \theta_k) = y) - \max_{j \neq y} aV_i I(h(x; \theta_k) = j)
\] (1)

Definition 2: Assume that generalization error reflects the classification ability of samples, which does not appear in the training set. The smaller the value is, the better the performance of the algorithm gets. The generalization error \( PE^* \) is defined as follows:

\[
PE^* = P_{x,y}(mg(x, y) < 0)
\] (2)

Theorem 1: When the random number is large in forest tree, the tree structure and the law of large numbers to ensure that the random forest does not produce over-fitting:

\[
\lim_{k \to \infty} PE^* = P_{x,y}(\theta_k h(x; \theta_k) = y) - \max_{j \neq y} P_{x,y}(\theta_k h(x; \theta_k) = j)
\] (3)

Theorem 2: The generalization error that exists on the boundary is given by the following equation:
Where, $\rho$ reflects the correlation between decision trees. $S$ reflects the ability to classify single decision tree. Thus, from equation (4), we can know that the smaller the correlation between decision tree, the better the single trees classification ability of decision tree, and the less generalization error in random forests.

While, Gini Index is adopted in the splitting attribute of decision tree, assuming that $T$ contains $N$ categories of records, hence the index is:

$$Gini(T) = 1 - \sum_{j=1}^{N} P_j^2$$  \hspace{1cm} (5)

Where $i$ is category, $j$ is the frequency of occurrence. If set the $T$ is divided into the $m$ parts $N_1, N_2, \ldots, N_m$, then the partition Gini is:

$$Gini_{\text{split}}(T) = \frac{N_1}{N} Gini(T_1) + \cdots + \frac{N_m}{N} Gini(T_m)$$ \hspace{1cm} (6)

By the $k$ rounds training, we get a regression Prediction model list $\{h(X; \theta_1), h(X; \theta_2), \ldots, h(X; \theta_k)\}$. Then a combined forecast model is built by these lists. The final result is:

$$\tilde{h}(X) = (1/K) \sum_{k=1}^{K} h(X; \theta_k)$$ \hspace{1cm} (7)

The training set of the Random forest has been generated by means of bagging method. The Out-Of-Bag (referred as OOB) generated with each sampling data is used to evaluate the accuracy of random forest regression prediction.

3. Prediction model based on feature extraction of random forest

3.1. Parameters with the amount of solar radiation

Solar radiation is the main energy of the earth's atmospheric motion. The total radiation intensity includes direct and the scattered radiation intensity. In the case of atmospheric attenuation, if the direct solar radiation at some point is $E_b$ and the direct solar radiation on a horizontal plane is $E_{bh}$, then the formula of $E_b$ and the direct radiation intensity $E_{bh}$ which consider solar elevation angle were:

$$E_b = E_0 \times \tau_b$$ \hspace{1cm} (8)

$$E_{bh} = E_0 \times \tau_b \times \sinh$$ \hspace{1cm} (9)

There, $\tau_b$ is a direct radiation atmospheric transparency coefficient; $h$ is the height of the sun angle.

As sun scattering part is more complex, it has some connection with cloud amount, atmospheric, cloud-shaped distribution of the sky. Let $E_d$ be the diffused solar irradiance to arrive on horizontal plane, then:

$$E_{dh} = \frac{1}{2} \times \sinh \left[ \frac{1 - \tau_d}{1 - 1.4 \ln[\tau_d / m(z,b)]} \right] \times k$$ \hspace{1cm} (10)

Where, $\tau_d$ is the scattered radiation atmospheric transparency, $k$ is selected according to the quality of air in between 0.60 and 0.90.

In general, the intensity of solar radiation reaching the ground can be $E_{sh}$:

$$E_{sh} = E_{bh} + E_{dh}$$ \hspace{1cm} (11)
3. Random forest forecast model based on feature-extraction

Feature extraction is a process for the related amount of solar radiation data component decomposition, recombination and selection. It is a key part of random forests, which not only determines the quality of subsequent processing, efficiency, system complexity and robustness, but also to be able to dig what knowledge and the results of the physical meaning of interpretability. We use the single variable feature extraction algorithms to calculate the relationship between a feature and regression variables, so that to calculate a score for each feature, excluding those important indicators. Here, a chi-square method to test features score is applied.

Chi-square test is a common distribution hypothesis testing method based on $\chi^2$. The null hypothesis $H_0$ is that there is no difference observed in frequency and expected frequency. The basic idea of the test is that we assume $H_0$ is correct, the value $\chi^2$ is calculated based on this premise. That indicates that the degree of deviation from the observed values with the theoretical value. According to the Chi-square distribution and degrees of freedom we can correctly determine the current statistics in the case of $H_0$ and the probability under more extreme circumstances. If the $P$ value is small, the degree observation value with the theoretical value is too big, hence we should reject the null hypothesis.

Considering the difference between the order of magnitude of these variables is larger, in order to make the model predictions more accurate, we have normalized the original variables, which have been transformed into dimensionless data. Using the Min-Max normalization method, all the data is mapped to the original values in the range of zero to one. The formula of the normalized is as follows:

$$x_i^* = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}}$$ (12)

Where, $x_i^*$ is the input variable after been normalized, $x_{\min}$ and $x_{\max}$ are minimum and maximum values of the original data, respectively. The normalized vectors are used as input of the RF algorithm, corresponding to the amount of solar radiation for the output of the RF algorithm.

Here, the data is the amount of solar radiation of a four-year 48 sets. The data formats Random Forests are analyzed in the database $n \times m$. Each row is used as a sample of the data and $m-1$ columns is used as variables. Then we create an improved random forest regression prediction model of solar radiation with nine inputs and one output. The random forest regression models are structured thought MATLAB.

4. Simulation results and analysis

4.1. Prediction model simulation and comparison

To verify the feasibility of study, using a common solar radiation monitoring database data from laboratory of Shenyang Institute of Engineering is used. Taking the Shenyang (41°50′N, 123°28′E) 2004 - 2014 solar radiation data for a sample, which includes sunshine hours, average atmospheric pressure, average wind speed, the average air temperature, solar elevation angle, the average humidity. Table 1 lists the predicted result of the three kinds of model of 2014. Due to the randomness of daily solar radiation at different times, discrete data can not completely describe the solar radiation principle, monthly average solar radiation value can better reflect the trend of solar radiation and the affected factors. Therefore, the monthly average of solar radiation value has been taken as in our research. Then we regard all data as monthly average value in data processing.

Table 1. Predicted result of the three kinds of model of 2014 ($\text{MJ} \cdot \text{m}^{-2} / \text{month}$).

| Month | MV  | IRF | RF  | SVM |
|-------|-----|-----|-----|-----|
| 1     | 114.32 | 96.99 | 88.18 | 83.65 |
| 2     | 135.74 | 160.52 | 108.39 | 157.47 |
In order to further verify the validity of this model, the real data was predicted through the experiment. Here, we predicted the monthly total solar radiation in April, July, October, January ranging from the 2004 to 2013 with the same three methods, with four month forecast data as four seasons. The solar month radiation predicted results with three kinds of methods can be seen in (a), (b), (c), (d) of the figure 1.

In addition, under the same training samples from 2004 to 2013, the three kinds of solar radiation average radiation prediction model are established, respectively. The predicted results of the solar year radiation are as shown in figure 2. From the fig.2, the prediction effect of the improved random forest method is better than others.

|   | 2004 | 2005 | 2006 | 2007 | 2008 | 2009 | 2010 | 2011 | 2012 | 2013 |
|---|------|------|------|------|------|------|------|------|------|------|
| 3 | 201.38 | 223.04 | 218.16 | 172.00 |
| 4 | 316.04 | 297.83 | 288.00 | 287.93 |
| 5 | 298.55 | 321.64 | 315.90 | 343.22 |
| 6 | 374.63 | 368.28 | 349.99 | 320.91 |
| 7 | 403.58 | 383.87 | 422.65 | 371.55 |
| 8 | 365.91 | 338.05 | 340.34 | 332.88 |
| 9 | 355.38 | 339.34 | 324.71 | 326.05 |
| 10 | 217.67 | 203.31 | 231.75 | 237.83 |
| 11 | 128.51 | 111.33 | 100.35 | 97.10 |

Figure 1. Solar month radiation predicted results in different months.
4.2. Model evaluation and inspection

It is a core that the evaluation and test of model in the entire modeling process, through model test to determine the fitting effect of the model. The model is evaluated based on the model decision coefficient ($R^2$) and root mean square error (RMSE).

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$  \hspace{1cm} (13)

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n}}$$  \hspace{1cm} (14)

Where, $y_i$ is the measured value; $\hat{y}_i$ is the model forecast; $\bar{y}$ is the average value of the sample; $n$ is the sample.

Through testing of the improved random forest regression model of the samples, the deviation statistics and evaluation of the predictive ability of model, have been compared through some statistics which include the general relative error ($R_s$), the average relative error ($E_1$), the overall relative error absolute value ($E_2$) and precision ($P$) are as follows:

$$R_s = \frac{\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \hat{y}_i}{\sum_{i=1}^{n} y_i}$$  \hspace{1cm} (15)

$$E_1 = \frac{1}{n} \sum_{i=1}^{n} \left( \frac{y_i - \hat{y}_i}{\hat{y}_i} \right) \times 100\%$$  \hspace{1cm} (16)

$$E_2 = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{\hat{y}_i} \right| \times 100\%$$  \hspace{1cm} (17)

In the formula, $y_i$ is the observed values; $\hat{y}_i$ is the model forecast; $n$ is the predicted sample.

The sample mean and sample can be calculated based on the sun's total radiation measured values in 2014, taking the results both of the calculation and the prediction into the formula (13) and (14). The accuracy test result as shown in table 2. Now, the decision coefficient of random forest regression model $R^2 = 0.9665$, the root mean square error $\text{RMSE} = 19.6085$, the fitting effect of this model is ideal and better than the other.

**Figure 2.** Solar year radiation predicted results.
Table 2. Accuracy test results of the three kinds of model.

| Month | R²   | RMSE  |
|-------|------|-------|
| IRF   | 0.9665 | 19.6085 |
| RF    | 0.9448 | 23.9800 |
| SVM   | 0.9329 | 27.724  |

The improved random forest regression model has the highest forecast precision, it can be seen that the improved random forest model forecast ability is better.

5. Conclusions
In this paper, an improved random forest algorithm is presented. The decision trees have been constructed and the variables of importance ranking have been optimized by evaluation function of random forest function. It is assumed that under the same condition, the three kinds of method have been compared. Studies and trials have led to the conclusion that under supposed conditions of too many cloudy days, the simulation conditions, using the improved random forest model predicted better than the results of other two algorithms. The new method is significant for the prediction of the amount of solar radiation under complex environment and for the use of photovoltaic power generation. There are a number of factors not taken into consideration, such as topography, altitude factor and so on, this segment still requires to be further researched.

Acknowledgments
This work is supported by the scientific research projects in Liaoning Province Natural Science Foundation of China (201602616), Liaoning Province Department of Education (L2015443) and Ministry of Housing and Urban-Rural Development of China (2015-K2-015).

Author contributions
For the research articles, the four authors distributed the responsibilities as Liu Jian carried out the theoretical algorithm research; Cao Meiyan carried out the experimental research and performed the analysis; Bai Di performed the experiments; Zhang Rui conducted the experimental data acquisition and analysis; Liu Jian and Cao Meiyan wrote the paper.

References
[1] Yadav AK and Chandel SS 2014 J. Renewable and Sustainable Energy Reviews 33 pp 772 – 781
[2] Yang MQ and Wang XY 2014 J. Applied Mechanics and Materials 571 pp 217 – 222
[3] Duz F, Montero G and Escobar J M 2014 J. Main Thematic Areas 69
[4] Alonso-Montesinos J and Batlles F J 2015 J. Energy 83 pp 387 – 393
[5] Feudo T L, Avolio E and Gulli D 2015 A preliminary study, in a coastal site of South Italy (Lamezia Terme), 17 p 11774
[6] Voyant C, Darras C and Muselli M 2014 J. Applied Energy 114 pp 218 – 226
[7] Troccoli A and Morcrette JJ 2014 J. Journal of Applied Meteorology and Climatology 53 pp 2571 – 2588
[8] Peng H W, Yang X F and Liu F R 2010 J. Power System and Clean Energy pp 82 – 84
[9] Angstrom A 1924 J. Journal of the Royal Meteorological Society 50 pp 121 – 126
[10] Zhou J, Wu Y Z and Yan G 2006 J. Journal of Harbin Institute of Technology 38 pp 925 – 927
[11] Zhang JJ 2014 J. Industrial control computer pp 39 – 40
[12] Zeng J and Qiao W 2013 J. Renewable Energy 52 pp 118 – 127
[13] Wang J, Jiang H and Wu Y 2015 J. Energy 81 pp 627 – 644
[14] Hosenuzzaman M, Rahim N A and Selvaraj J 2015 J. Renewable and Sustainable Energy Reviews 41 pp 284 – 297
[15] Kaushika N D, Tomar R K and Kaushik S C 2014 J. Solar Energy pp 327-342
[16] Gu W, Vieira A R and Hoekstra R M 2015 *J. Epidemiology and Infection* pp 1-9
[17] Ramedani Z, Omid M and Keyhani A 2014 *J. Renewable and Sustainable Energy Reviews* 39 pp 1005—1011
[18] Liaw A and Wiener M 2002 *J. R news* 2 pp 18－22
[19] Criminisi A, Shotton J and Konukoglu E 2012 *Foundations and Trends in Computer Graphics and Vision* 7 pp 81－227