INTERNATIONAL JOURNAL OF CIRCUITS, SYSTEMS AND SIGNAL PROCESSING
DOI: 10.46300/9106.2020.14.65 Volume 14, 2020

A Multi Fusion Data Mining Algorithm for Solar Energy Efficiency

Yue Lin[1], Zhan Shuo[1], Bai Jing[1], Kanae Shunshoku[2]

[1]College of Electrical and Information Engineering, Beihua University, Jilin, China

[2]College of Engineering, Fukui University of Technology, Fukui(910-0004), Japan

Received: June 6, 2020. Revised: August 13, 2020. Accepted: August 17, 2020. Published: August 17, 2020.

***Abstract*—The output power of renewable energy has the characteristics of random fluctuation, which have the harmful effect on stability of renewable power grid and causes the problem of low utilization ratio on renewable energy output power. Thus, this paper proposed a method to predict the output power of renewable energy based on data mining technology. Data mining is performed using linear regression algorithm, decision tree, and random forest. The simulation experiment results show the variation of solar radiation size and inclination angle, which improves solar panel position control accuracy and solar energy utilization in solar photovoltaic power generation systems. And this provides the scientific basis for theory and application of the efficiency of utilizing solar energy.**

***Keywords*—**Solar Energy, Python, Data Mining, Algorithm.

I. INTRODUCTION

With the continuous development of new energy generation in the power system, the proportion of renewable energy generation in the power system has gradually increased. However, the random fluctuation of renewable power generation will cause power grid unstable. In addition, the new energy generation system represented by solar power generation is relatively random due to the change of sunlight, making the utilization rate of new energy resources at a relatively low level. In photovoltaic power generation systems, the relationship between solar panels and sunlight will directly affect the efficiency of their power generation. When the radiation level of sunlight on the local surface is at a high level, the power generation effect of the panel will be at a relatively high level.

In recent decades, many efforts have been devoted to renewable energy resources and generation power prediction. These methods can be divided into three categories: physical model, statistical model and hybrid model. In the physical model, a specialized model is established for a specialized scenario and they do not need lots of historical data to train the model. The statistical model is a kind of time-series prediction method. Compared with physical techniques, it is more widely implemented in practice. These models are based on using historical data to develop the relationship between several variables. The conventional statistical model includes autoregressive model, moving average model, autoregressive moving average (ARMA) model and other variants of similar models.

Apart from the mentioned forecasting methods above, artificial intelligence has been adopted for time-series based forecasting due to its abilities such as self-learning, easy implementation and establishing non-linear relationships between inputs and outputs. Artificial neural network (ANN) such as radial basis function neural network, back propagation (BP) neural network, Elman neural network and extreme learning machines (ELM) have been implemented for forecasting renewable energy sources and loads. Another machine learning tool called support vector machine (SVM) has also been used as forecasting engine. However, there are some drawbacks such as falling into local minima and over-fitting. Meanwhile, SVM are sensitive to parameter selection and time consuming.

Python is an object-oriented interpreted computer programming language invented by the Dutch Guido van Rossum in 1989. The first public release was released in 1991 [1-3]. The biggest advantages of the language are open source and free, its use will not be subject to any legal or policy restrictions, making the current many new algorithms and tools as its preferred development environment. At the same time, the current mainstream artificial intelligence also regards this language as the primary support language. The compiler platform used by the Python language for this article is Anaconda3.

Data mining technology is the process of extracting potential, valuable knowledge models or rules from massive data [4-6]. Data mining technology is used to predict the solar radiation value and inclination, so that the solar panel can track the change of the sun's inclination, which can effectively improve the photoelectric conversion rate of solar panels.

In this study, mathematical statistics, meteorological model and others basic numerical calculation method is feasible. But the advantages of the proposed method is that these artificial intelligence data mining approach methods has

been adopted for time-series based forecasting due to its abilities such as self-learning, easy implementation and establishing non-linear relationships between inputs and outputs.

II. Realization of Data Mining Technology

A. Data Mining Technology

Data mining technology includes many subject technologies, including database technology, statistics, machine learning, pattern recognition, artificial intelligence, neural networks, and random forests [7,8]. At present, this technology has been used as a discipline and has been studied by major research institutions both at home and abroad. Related research results and related books have been widely published. The research institutions of famous universities in various countries and the research departments of major companies have invested a lot of energy in their research and have obtained many theoretical systems to achieve massive data processing; rough set and fuzzy set theory are integrated for knowledge discovery; fuzzy system identification method is constructed for knowledge achievement of fuzzy system. In recent years, China has also carried out relevant research and development work closely following the international trend. The national research fund has funded corresponding research topics. The research focus is shifting from discovery method to system application and focuses on the integration of multiple discovery strategies and technologies and the interpenetration of multiple disciplines. However, it is mainly based on academic research and practical application is still in its infancy. The relatively new developments at present are: the research of classification technology, trying to establish its collection type; constructing the intelligent expert system; researching the theoretical model and implementation technology of Chinese text mining; using the concept of text mining.

B. Import of Data

Importing data is the first step in data mining. There are many types of data sources in the data source. This paper uses the solar radiation collected by Hawaii Island and its meteorological information as an example to perform data mining. After running through the corresponding code, the imported raw data are shown in **Table 1**.

Table 1. The result returned after executing the import data code

UNIXTime	Data	Time	Radiation	Temperature	.....
1475229326	2016-09-29	23:55:26	1.21	48	.....
1475229023	2016-09-29	23:50:23	1.21	48	.....
1475228726	2016-09-29	23:45:26	1.23	48	.....
1475228421	2016-09-29	23:40:21	1.21	48	.....
1475228124	2016-09-29	23:35:24	1.17	48	.....
.....	.....	.....	.....	.....	.....

After the import is successful, "df" represents the variable name of the source data. When calling this data, type the variable name "df". Under normal circumstances, in order to facilitate the follow-up work, there is usually a need to first look at the overall structure of the imported data. At this time, using the corresponding code, Python will automatically calculate the overall structure of the data, such as the total number, average value, maximum value, standard deviation and other commonly used statistical parameters and return the result value. By compiling the corresponding code, the result is shown in Table 2.

Table 2. The result of the overall analysis of the data

	Radiation	Temperature	Pressure	Humidity	Speed
count	32686.00	32686.00	32686.00	32686.00	32686.00
mean	207.12	51.10	30.42	75.01	6.24
std	315.92	6.20	0.05	25.99	3.49
min	1.11	34.00	30.19	8.00	0.00
25%	1.23	46.00	30.40	56.00	3.37
50%	2.66	50.00	30.43	85.00	5.62
75%	354.23	55.00	30.46	97.00	7.87
max	1601.26	71.00	30.56	103.00	40.50

C. Preprocessing of Data

Due to the data collection process, it is inevitable that some abnormal situations will occur, or the recorded data types are text types that cannot be analyzed by mathematical operations. This situation will have a great impact on the later excavation work, seriously affecting the accuracy of the calculation. So you need to deal with this kind of value.

There are three main situations when the data needs to be preprocessed: missing data, data record type text values, and some features in the extracted data. For the missed value, the general method is to delete the data corresponding to the value, or to use statistical parameters such as the mean to fill. For text values, it is generally artificially defined numbers to replace all of them (for example, "0" instead of "sunny"). For feature extraction, according to the extracted features, the corresponding language rules are written and the data is processed.

D. Visual Display of Data

The visual display of data can be in the form of a statistical graph, which visually depicts the characteristics of the data, or show the relationship between the various features in the data.

Before selecting the appropriate algorithm for data mining, it usually visualizes the overall situation of the data or the relationship between variables, providing a reference for the choice of the next algorithm. By compiling the corresponding code, the degree of correlation between data variables is obtained, as shown in **Fig 1**.

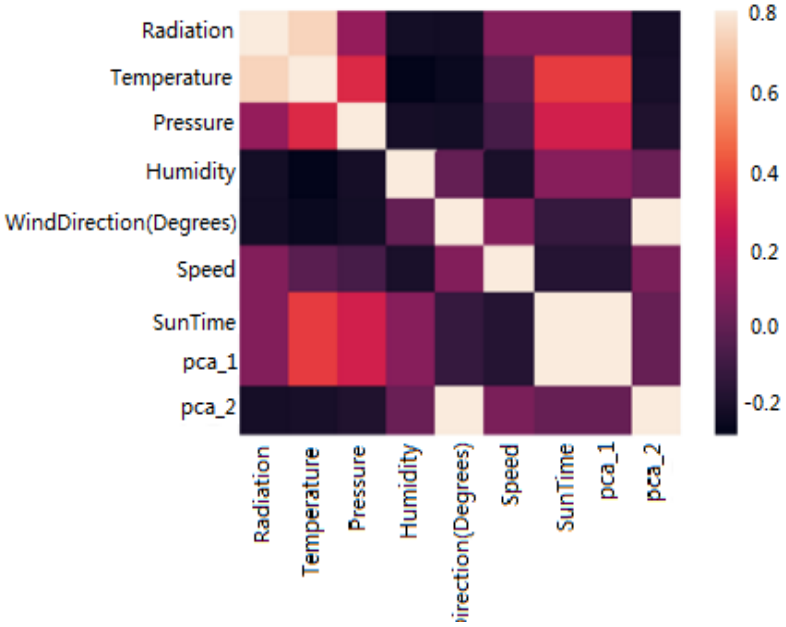

Fig. 1. Analyze the Results Returned by Relationships between Data Variables

From this figure, it can be directly seen that the relationship between the variable "Radiation" and "Temperature" is relatively large, so the correlation analysis is performed on the two variables. The results obtained are shown in **Fig 2**.

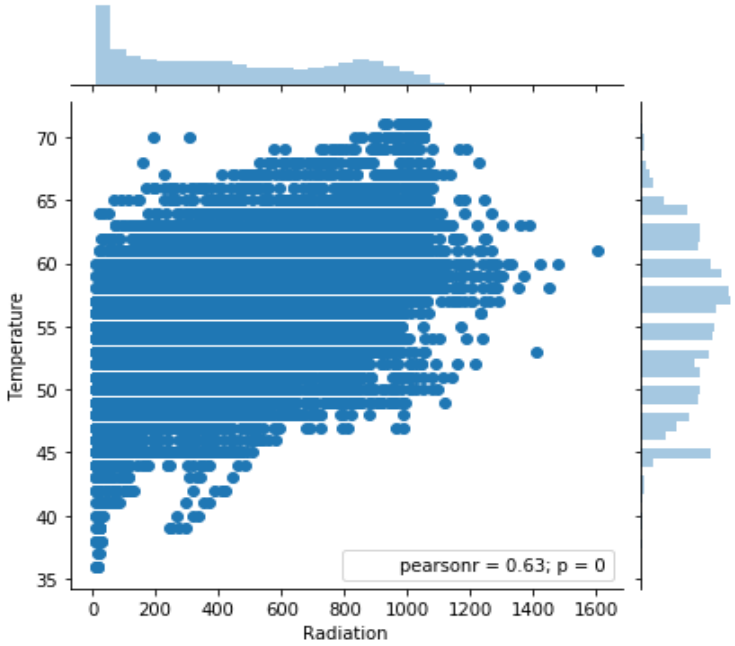

Fig.2. Analysis of the Relationship between Temperature and Radiation

Previous work has laid the foundation for data mining. The core of data mining work is the algorithm. Different algorithms dig out different results. As for which algorithm to choose, there is currently no uniform law. It is often based on the experience of the workers and the overall structure of the data, and with the expected results, an optimal judgment and selection are made, and the most effective algorithm is finally selected. In the following section, based on the data analyzed above, combined with different algorithms, the data mining work and compare the effects of different algorithms.

III. Solar Radiation Prediction

The greatest feature of solar energy is its volatility and instability. The value of solar radiation is directly related to the efficiency of solar energy use. This paper first analyzes the radiation intensity data of solar energy and predicts the variation law of solar radiation value. Based on this, the best working time of solar panels can be determined. Based on this time, predict the change law of the sun's inclination to increase the utilization of solar energy. Although the value of solar radiation seems to be random, it is related to many factors, such as time, season and weather. Through the processing and excavation of these data, find out its internal laws and provide important theoretical reference for the operation of new energy.

A. Linear Regression Algorithm

Linear regression algorithm is one of the most basic but important algorithms in data mining. Linear regression is a statistical analysis method that uses the regression analysis in mathematical statistics to determine the quantitative relationship between two or more variables. It is widely used. For the linear regression algorithm, the analysis flow is generally given for each sample and its correct answer in the given data set. Select a model function h and find the optimal solution for the function h (not necessarily global). The parameters of h under the optimal solution [9-11]. Here, the given data set is named Training Set. Not all data can be used for training. A part of them are used to verify the accuracy of the model. This part is called the Test Set. Available formula (1) means:

$$h_\theta(x) = \sum_{i=0}^{n} \theta_0 x_0 + \theta_1 x_1 + \ldots + \theta_n x_n \quad (1)$$

The results of the linear regression analysis can be expressed as (2)

$$R^2 = 1 - \frac{\sum_{i=1}^{m} (\hat{y}_i - y_i)^2}{\sum_{i=1}^{m} (y_i - \bar{y})^2} \quad (2)$$

In the formula, $\hat{y}_i$ is the regression prediction results, and y_i is the result value of the original data, and $\bar{y}$ is the average value.

The division of training sets and test sets in data also does not have a uniform standard. In this data, because it contains a time series, the division of the training set and the test set in this data will be intercepted in chronological order. The data is a time series data, so the data divide amount ratio of the test and training set is 4:1 on the basic of ascending date sequence in order to obtain an accurate model. **Fig.3** shows the divide of the test and training set of the data. By compiling the corresponding code, the division of the training set (Training Set) and the test set (Test Set) in the source data are completed. After the excavation work can directly call the above divided data for processing.

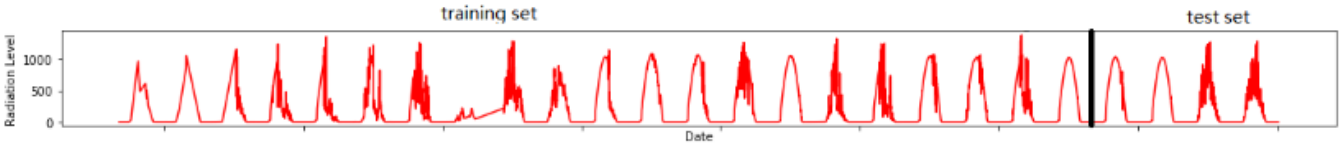

Fig.3 The divide of training data set and test data set

The following is the use of linear regression algorithm for data mining work, and the specific code is as follows:

The code completes the mining work using the linear regression algorithm. The returned result is shown in **Fig 4** below.

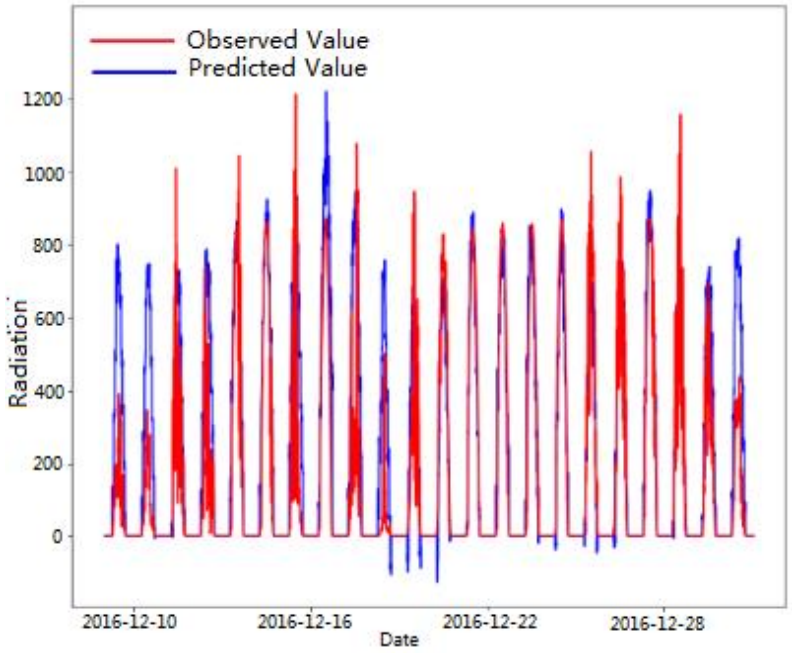

Fig. 4. Linear regression prediction results returned by running code

The prediction accuracy using the linear regression algorithm is 0.611, which is approximately 61%. As can be seen from the figure, the blue forecast has a negative number. However, the actual radius value may not be negative, indicating that the algorithm is not particularly suitable for this data, and the prediction accuracy is not high enough. There is a need to further use other algorithms for analysis and prediction.

B. Decision Tree

A decision tree is a tree structure similar to a flowchart in which each internal node (non-leaf node) represents a test on an attribute, and each branch represents one output of the test, and each Tree leaf (or end node) stores a classification label. The topmost node of the tree is the root node. A typical decision tree is shown in **Fig 5**. It can be thought of as a set of if-then rules, or as a conditional probability distribution defined in feature space and class space. The main advantages of the decision tree are that the model has readability and the classification is fast. During learning, using the training data, a decision tree model was established based on the principle of minimizing the loss function. When forecasting, classify new data using decision tree models[12-15].

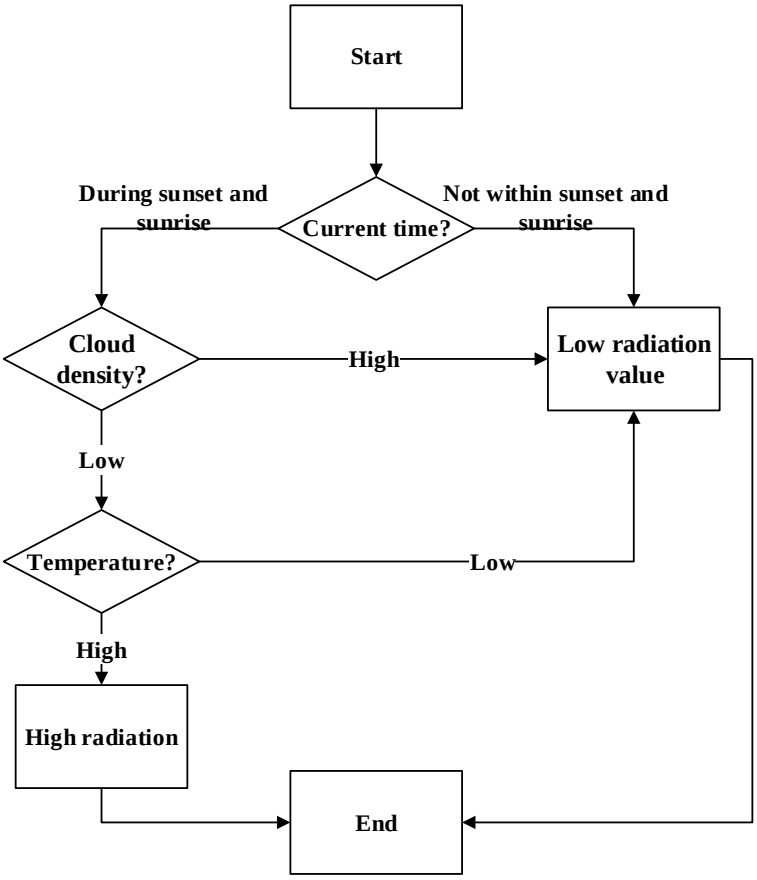

Fig.5. Decision Tree Block Diagram

Decision tree learning usually includes three steps: feature selection, decision tree generation, and decision tree pruning [16]. Feature selection is to extract the features needed in the data as a basis for classification. In some massive data occasions, the data might contain hundreds and thousands of features. This will greatly influence and decrease the speed of decision tree generation. Therefore, if the number of features is large, the features are selected at the beginning of the decision tree learning, leaving only features that are sufficiently categorized for the training data and then starting to generate a decision tree. The purpose of pruning the decision tree is to make the tree simpler and more generalized.

The following section will use the decision tree algorithm to compile the corresponding code and re-analyze the previous data. The results obtained are shown in **Fig 6**.

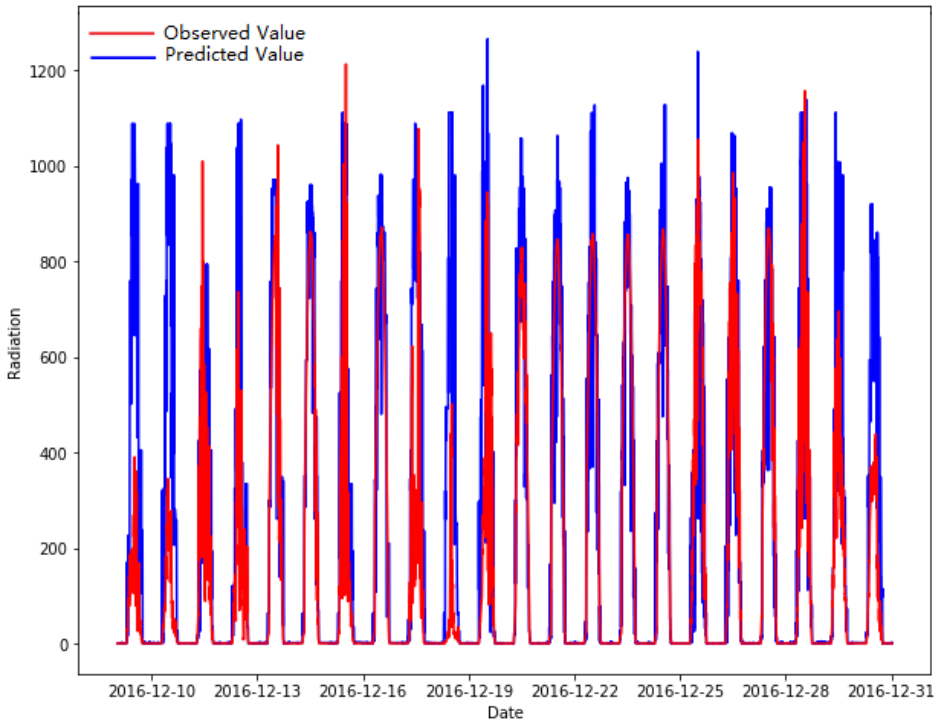

Fig 6. The result returned by the decision tree algorithm

As can be seen from the above figure, there is no negative value for the foretasted value, indicating that the predicted value is close to the actual situation. However, it can be seen from the figure that some of the prediction results have a large deviation from the actual observations, and the accuracy of the return is only 0.392, which is about 39%, indicating that the prediction results do not have reference value and the algorithm still needs to be optimized.

C. Random Forest

A random forest is a classifier that contains multiple decision trees, and its output category is determined by the mode of the output of individual trees, hence the name random forest. Random forest is a statistical learning theory. Random forest algorithm uses the bootstrap re-sampling method to extract multiple samples from the original sample, performs decision tree modeling for each bootstrap sample, and then combines the predictions of multiple decision trees to obtain the final result through voting forecast result. Therefore, random forest combined multiple prediction results from each decision tree, and the final result is obtained by voting. Random forest algorithm has high prediction accuracy, good tolerance to outliers and noise, and is not easily over-fitted. It has a wide range of applications in medicine, bioinformatics, and management. [17-19]

The schematic diagram of the random forest is shown in **Fig.7**. The basic idea is: First, use k bootstrap sampling to extract k samples from the original training set, and the sample size of each sample is the same as the original training set; and secondly, separate the k samples. Set up k decision tree models to obtain k classification results; Finally, each record is voted according to k classification results and the final classification is determined.

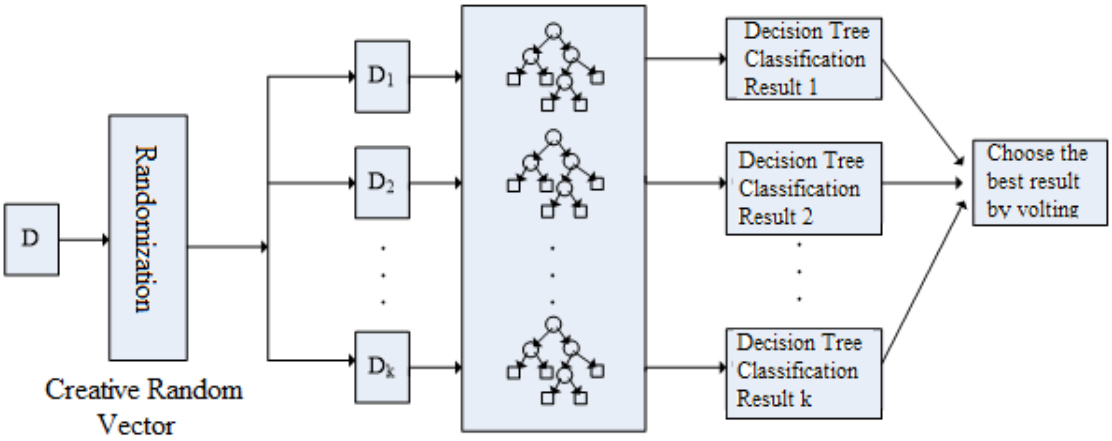

Fig. 7. Schematic Diagram of Random Forest

Random forests increase the difference between classification models by constructing different training sets, thereby increasing the extrapolation prediction ability of the combined classification model. Through k-round training, a classification model sequence $\{h_1(X), h_2(X), \cdots, h_k(X)\}$ is obtained. Then use them to form a multi-class model system, the system's final classification results using a simple majority vote method. The final classification decision can be expressed by the following formula (3).

$$H(x) = \arg\max_{Y} \sum_{i=1}^{k} I(h_i(x) = Y) \qquad (3)$$

In the formula, $H(X)$refers to combined classification model, h_i is the single decision tree classification model, Y refers to output variable (or target variable), and $I(\circ)$ is the representation function. Equation (3) illustrates the use of majority voting decisions to determine the final classification.

This paper will use the random forest algorithm to re-analyze the data again. By compiling the corresponding code, the returned results are shown in **Fig 8**.

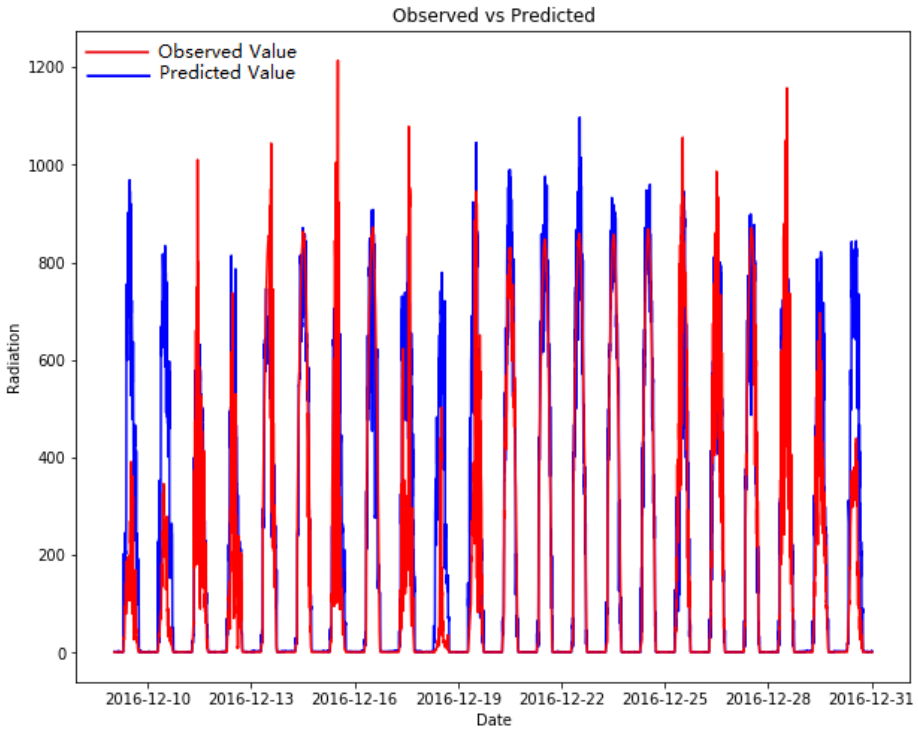

Fig 8. The results returned by the random forest algorithm

As can be seen from the figure, the performance of the random forest algorithm is better, the matching degree between the predicted value and the observed value is very high, and the returned accuracy is 0.662, which is about 66%.

IV. Solar Dip Prediction

Based on the premise that solar radiation values are determined, based on such a premise, the solar energy utilization rate can be further improved by predicting the change law of the solar dip angle.

Angstrom proposed the most popular theoretical model for estimating global solar radiation based on sunshine duration, where the daily extraterrestrial solar radiation on a horizontal surface, $H_0(MJ/m^2/day)$, is calculated from the following Equation:

$$H_0 = \frac{24*3.6*10^{-3}*I_{SC}}{\pi}*(1+0.033\cos(360*\frac{d}{365}))* (\cos\phi\cos\delta\sin\omega_s + \frac{2\pi}{360}\omega_S\sin\phi\sin\delta) \quad (4)$$

Where d is the Julian day number; ISC is the solar constant with a value if 1367 W/m^2; ϕ is the latitude of the location; δ is the declination angle:

$$\delta = 23.45\sin\frac{360(284+d)}{365} \quad (5)$$

and ω_s is sunset hour angle in degree which equals:

$$\omega_s = \arccos(-\tan\phi\tan\delta) \quad (6)$$

They relate monthly average daily global radiation (H) to the average daily sunshine hours S, by the following first order regression equation:

$$\frac{H}{H_0} = a + b\frac{s}{s_0} \quad (7)$$

where maximum sunshine hours or day length(S_0) is:

$$S_0 = \frac{2}{15}\arccos(-\tan\phi\tan\delta) \quad (8)$$

Solar panels, an important tool for solar energy, have a direct relationship between the efficiency of the output power and the incident azimuth angle of sunlight on the surface of the panel. Theoretical analysis shows that the solar energy receiving rate differs by 37% between tracking and non-tracking [20-21]. Solar panels can achieve maximum output efficiency at this time when sunlight is incident vertically. Therefore, ensuring the vertical relationship between the solar panel and the incident light is one of the important conditions for improving the solar energy utilization rate. Therefore, by digging the solar dip angle data, predicting the change law of solar dip angle provides an important theoretical basis for the tracking of solar panels. In the following text, the solar dip angle data is used for mining and forecasting.

The raw data of a region's solar dip (zenith angle) is used as the standard, and the original data is imported. In **Table 3**, the UNIXtime is the local time stamp, GHI is Global Horizontal Irradiance value, and DNI is Direct Normal Irradiance value, as shown in **Table 3**.

Table 3. Original Data Content (first five rows)

	Unixtime	GHI	DNI	Temperature	Sloar Zenith Angle
0	915121800	0	0	-8.046881	159.487386
1	915125400	0	0	-7.776343	152.576279
2	915129000	0	0	-7.505621	142.762115
3	915132600	0	0	-7.426703	131.938871
4	915136200	0	0	-7.617773	120.860264
.....	.....	.....	.....	.....	.....

A. Linear Regression

Using the linear regression algorithm code above and through the operation, the result is shown in **Fig 9**.

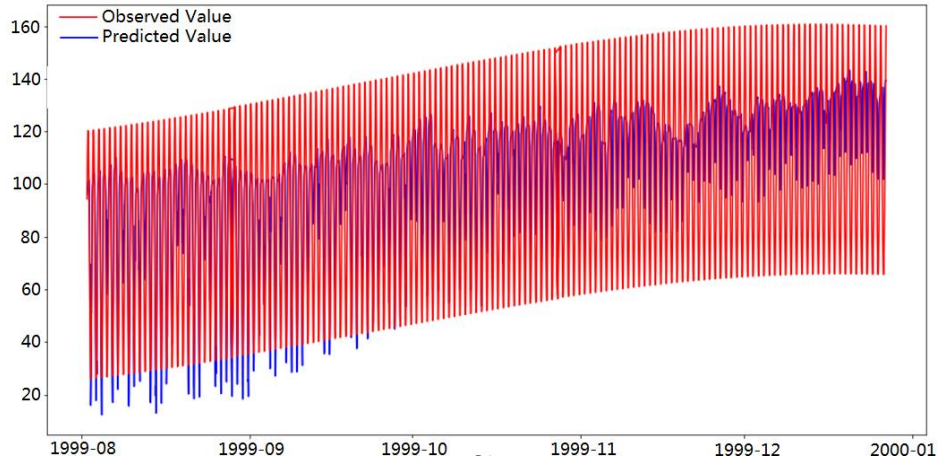

Fig. 9. The prediction graph returned by the linear regression algorithm

Using the linear regression algorithm, it can be seen from the figure that the overall tracking situation is closer to the actual trend, but the accuracy returned is only 48%, so further optimization of the algorithm is needed.

B. *Decision Tree*

Using the decision tree algorithm code above, the result is shown in **Fig 10**.

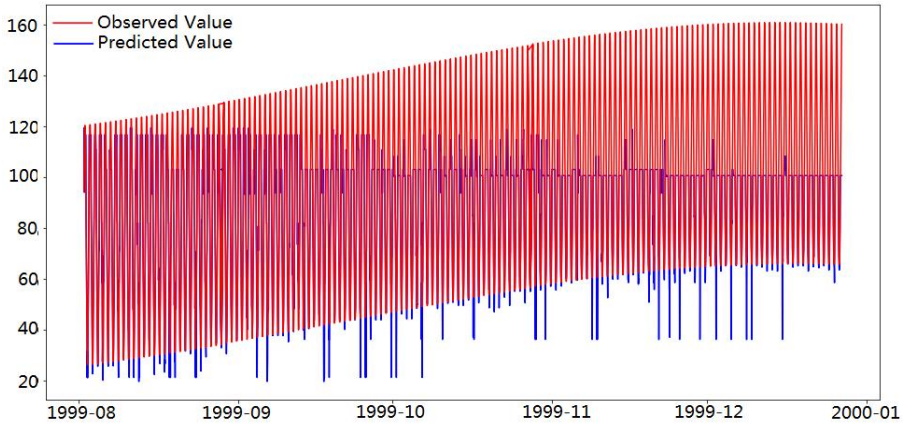

Fig. 10. The prediction graph returned by the decision tree algorithm

As can be seen from the figure, the overall tracking effect is closer to the real observation value than the linear regression, and the return accuracy is 55%.

C. *Random Forest*

The results returned by using the random forest algorithm are shown in **Fig 11**.

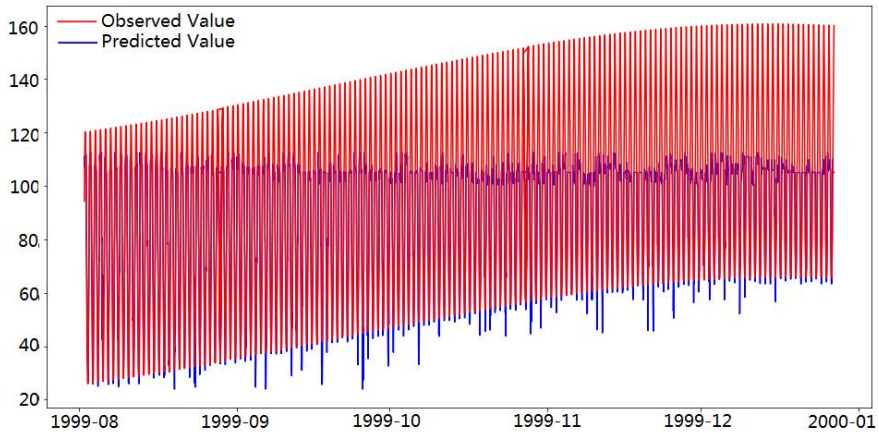

Fig. 11. The forecast graph returned by the random forest algorithm

As can be seen from the figure, the accuracy of the random forest prediction is the highest, reaching 67%. It can be seen from the above figure that among these parameters, the local unixtime has the greatest influence on the final result. However, there is still a large part of the deviation between the predicted value and the observed value.

V. Conclusion

In this paper, three different data approach methods and prediction accuracy result is proposed. First, the data related to the renewable power generation is prepared for data mining. Then using various approaches to mine and obtain the renewable power generation value within a period. The prediction accuracy of linear regression algorithm on the data test set is 61%, and the accuracy of the decision tree is 39%, the accuracy of random forest reached 66%. The accuracy of the linear regression algorithm in solar dip prediction is 48%, the decision tree is 55%, and the random forest is 67%.

It can be seen that for the same data, using different analysis algorithms, the results are not the same; while the same algorithm is applied to different data, the results are not the same. Since the accuracy of data mining can only approach 100% indefinitely, multiple data mining analyses are required. The effect of mining is comprehensively compared and selected. There is no uniform algorithm selection criterion and multiple attempts are required.

The prediction model has obvious advantages in short-term or long-term renewable power forecasting. The prediction results can be used as reference for the renewable dispatch

system. By processing the dataset related to the renewable power generation, the problem of the randomness of renewable power generation and the complexity of model structure are overcome. By using intelligent algorithm approach method, the prediction result can be more stable and reliable. This will have theoretical and practical reference for the future work about renewable energy in data mining.

ACKNOWLEDGEMENT

This work is supported by the science and technology project in Jilin Province (20170312031ZG).

REFERENCES

[1] Xing Cai,Hans Petter Langtangen,Halvard Moe. On the Performance of the Python Programming Language for Serial and Parallel Scientific Computations. Scientific Programming, v13, n1, p31-56, 2005.

[2] Stancin I, Jovic A. An overview and comparison of free Python libraries for data mining and big data analysis. 2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO 2019 - Proceedings, p 977-982, May 2019

[3] Agnihotri Lalitha, Mojarad Shirin, Essa Alfred. Educational Data mining with python and apache spark. ACM International Conference Proceeding Series, v 25-29-April-2016, p 507-508.

[4] Buda, Micha&lstrok. Data mining algorithms in the analysis of security logs from a honeypot system. Advances in Intelligent Systems and Computing, v470, p 63-73, 2016.

[5] Alcalá-Fdez, J., Robles I., Herrera F. Introduction to the experimental design in the data mining tool KEEL. Intelligent Soft Computation and Evolving Data Mining: Integrating Advanced Technologies, p1-25, 2010.

[6] Embrechts, Mark J. Szymanski, Boleslaw. Introduction to Scientific Data Mining: Direct Kernel Methods and Applications. Computationally Intelligent Hybrid Systems: The Fusion of Soft Computing and Hard Computing, p317-362, June 08, 2012.

[7] Besimi Nuhi,Çiço Betim, Besimi Adrian.Overview of data mining classification techniques: Traditional vs. parallel/distributed programming models.2017 6th Mediterranean Conference on Embedded Computing, MECO 2017 - Including ECYPS 2017, Proceedings, July 12, 2017

[8] Zhang Xiaoren,Chen Xiangdong,Ding Ling.Self-service product innovation based on data mining technology.International Journal of Database Theory and Application, v 6, n 5, p 105-118, 2013.

[9] J.M. Łęski, N. Henzel. Generalized ordered linear regression with regularization. Bulletin of the Polish Academy of Sciences: Technical Sciences, v60, n3, p 481-489, September 2012.

[10] Xue Tao, Li Ting-Ting. Research on parallelization of KNN locally weighted linear regression algorithm based on mapReduce.Journal of Communications, v10, n11, p 864-869, 2015.

[11] Zhao Weihua, Zhang, Riquan, . Robust and efficient variable selection for semiparametric partially linear varying coefficient model based on modal regression. Annals of the Institute of Statistical Mathematics, v66, n1, p 165-191, February 2014.

[12] Aloufi Asma, Hu Peizhao,Wong Harry W. H. Blindfolded Evaluation of Random Forests with Multi-Key Homomorphic Encryption. IEEE Transactions on Dependable and Secure Computing, 2019.

[13] Aich Satyabrata, Choi Kiwon, .Prediction of Parkinson disease using nonlinear classifiers with decision tree using gait dynamics. ACM International Conference Proceeding Series, p52-57, November 12, 2017.

[14] Tayefi Maryam, Esmaeili Habibollah, Saberi Karimian Maryam. The application of a decision tree to establish the parameters associated with hypertension. Computer Methods and Programs in Biomedicine, v139, p83-91, February 1, 2017.

[15] Ziegler Andreas, König Inke R. Mining data with random forests: Current options for real-world applications. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, v4, n1, p 55-63, January/February 2014.

[16] Zhuang Xu,Zhu Yan,Chang Chin-Chen.Feature bundling in decision tree algorithm. Intelligent Data Analysis, v 21, n 2, p 371-383, 2017

[17] Chong Su, Shenggen Ju. Improving Random Forest and Rotation Forest for highly imbalanced datasets. Intelligent Data Analysis, v 19, n 6, p 1409-1432, November 3, 2015.

[18] Junfeng Zhu,William P. Pierskalla. Applying a weighted random forests method to extract karst sinkholes from LiDAR data. Journal of Hydrology, v533, p343-352, February 01, 2016.

[19] Hassan Fathabadi. Novel high accurate sensorless dual-axis solar tracking system controlled by maximum power point tracking unit of photovoltaic systems. Applied Energy, v173, p448-459, July 1, 2016.

[20] Zhe Mi, Jikun Chen, Nuofu Chen. Open-loop solar tracking strategy for high concentrating photovoltaic systems using variable tracking frequency. Energy Conversion and Management, v117, p142-149, June 1, 2016.

[21] Sabran Rexel U,Fajardo Arnel C.Sunflower inspired solar tracking strategy: A sensorless approach for maximizing photovoltaic panel energy generation.2018 IEEE 10th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment and Management, HNICEM 2018, p251-254, July 2, 2018.