Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 7
TPCEE 2024

Prediction of the Next Solar Cycle and Solar Maximum Based on Machine Learning

Xuanyi Xiang [#, *], Xinyue Zhang [#], Qingyi Huang, Jingyi Liu

Pittsburgh Institute, Sichuan University, Chengdu, China, 610044

* Corresponding Author Email: 2021141520064@stu.scu.edu.cn

[#]These authors contributed equally.

Abstract. The intensity of solar activity varies with the solar cycle of about 11 years, reaching a peak during the solar maximum. Solar activity can affect space weather in ionospheric states and conditions related to shortwave radio propagation or satellite communications, which brings challenges to space exploration, communications, and weather forecasting. Therefore, accurate prediction of the start time and duration of the solar maximum becomes pivotal. In order to solve this problem, this research first applies the K Nearest Neighbors classification algorithm to predict that the beginning of the next solar cycle will be around June 2030 until about July 2040. Then, by establishing the Extreme Gradient Boosting model, this study forecasts that the solar maximum in the next solar cycle will start in August 2034 and last for 8 months. With this accurate prediction of the next solar maximum, aerospace, communications, meteorology, and other industries can reduce the interference of solar activities.

Keywords: Solar Cycle Time, KNN Regression Algorithm, XGBOOST Regression Model.

1. Introduction

The Sunspots are dark spots on the sun's photosphere, areas of surface temperature reduction caused by magnetic flux concentrations [1]. Zhou and Zhong mentioned that the Sunspots are often found in the active regions of the Sun and tend to be opposite pairs of magnetic polarity [2]. The Sun's cycle of activity is usually around 11 years, but the duration may vary slightly [3]. The peak of sunspot activity is called the solar maximum, and the lowest point of activity is called the solar minimum [4]. This cycle is also related to other solar activity, changes in the sun's magnetic field, and changes in the polarity of the magnetic field. This paper predicts the start and end times of the current and next solar activity by analyzing data from previous solar cycles, and predicts the start time and duration of solar activity using XGBOOST regression models and time series models. According to Amrita, the solar cycle lasts about 11 years. Internationally, the sunspot cycle has been counted as week 1 since 1755 and the current cycle is week 25 [5]. The current solar cycle begins in December 2019. KNN regression is established by observing the change of period number. Based on the algorithm, this paper predicts that the current cycle will end around June 2030 and that the next solar cycle will start around June-July 2030 2040. It is found that XGBOOST regression model has a better fitting effect than the time series model. The XGBOOST regression model uses start time and duration as independent variables and cycle as dependent variables. By collecting observations of past solar cycles and using that data to train the model. The solar maximum of the next solar cycle begins in August 2034 and lasts for eight months. Predicting the number and area of sunspots is a complex and challenging task. A variety of models can be used to predict the number and area of sunspots in the future, all of which are subjected to uncertainty and error caused by the complexity and uncertainty of solar activity [3]. Therefore, when using the forecast results, these factors need to be carefully considered and the forecast results should be adjusted appropriately.

The data quoted in this article comes from: http://solarcyclescience.com/activeregions.html

2. Model Establishment

2.1. KNN Regression Algorithm

KNN algorithm, also known as K nearest neighbor classification algorithm, is one of the simplest methods in data mining classification technology. The nearest K neighbor means that there are K nearest "neighbors", which means that each sample can be represented by its closest K "neighbors" [6].

In a nutshell, the algorithm is to know that some samples in a sample space are divided into several classes, and then, the given data need to be classified, which finds the K samples closest to itself through calculation, and the K samples vote to decide which category the data to be classified belongs to. When the algorithm makes a category decision, it is only related to a very small number of adjacent samples. Since the method KNN mainly relies on the surrounding limited neighboring samples, rather than relying on the method of the discriminant class domain to determine which class it belongs to, the method is more suitable than other methods for the sample set to be divided with more intersection or overlap of class domains [7]. The algorithm selection is as follows.

The basic idea of the regression algorithm is to select the nearest training samples in the Euclidean space according to k, the similarity of the Euclidean distance between the samples, and the regression value k is the mean or weighted value of the nearest neighbor samples. Based on the improvement of gearbox condition monitoring in the literature, the specific steps of the K nearest neighbor regression algorithm are as follows:,

(1) Let's say we have n samples, and let's write them as $X = (X_i, \dots, X_n)$, Each training sample can be represented as $X_t = (x_{i1}, x_{i2}, \dots x_{it}, y_t), i \in n.$ Then the Euclidean distance *D* between the training sample $X_t = (x_{i1}, x_{i2}, \dots x_{it}, y_t), i \in n$ and the test sample *Xi* can be expressed as:

$$D(X_i, X_t) = \sqrt{\sum_{m=1}^{d}(x_{im} - x_{tm})^2 + (y_i - y_t)^2} \tag{1}$$

In this formula, *xim* and *xtm* are the observed values of the MTH auxiliary variable, and *yi* and *yt* serve as the observed values of the monitoring variable.

(2) Calculate the Euclidean distance between all training samples and test samples according to the above formula, and find the X_i previous K nearest neighbor sample

$$X_j' = \left(x_{j1}' + x_{j2}', \dots, x_{jd}', y_j'\right), j \in K \tag{2}$$

(3) Calculate X_t and estimates of the monitored variables of the $\hat{y}_t$ test samples

$$\hat{y}_t = \frac{1}{K}\sum_{j=1}^{K} y_j' \tag{3}$$

2.2. XGBoost

XGBoost is a gradient boosted tree algorithm, which is an ensemble learning algorithm based on decision trees. It interactively trains multiple weak classifiers (decision trees) and combines them into a single strong classifier.

XGBoost is based on the idea of GBDT gradient boosting tree, the GBDT optimization of the model, is an ensemble regression tree model, which can be used to do regression prediction and classification prediction and other problems, and is currently a relatively strong and widely used machine learning

model [8]. The basic principle is the linear weighted sum of a single tree, the complexity of the model is determined by the number of trees, and the prediction error of all tree nodes is the loss of the model.

This article defines an objective function to represent the optimization goal of the model. Typically, a weighted loss function is used, plus a regularization term. Its predictive model is expressed as:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), f_k \in F \tag{4}$$

After the first t iteration, the model's prediction is equal to the previous model t-1 's prediction plus the first tree's prediction t

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_t(x_i) \tag{5}$$

The objective function is:

$$L^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i{}^{(t-1)} + f_t(x_i)\right) + \Omega(f_t), \tag{6}$$

There into $l(y_i, \hat{y}_i)$ for the first prediction error for samples. Use the second-order Taylor to unfold:

$$L^{(t)} \cong \sum_{i=1}^{n} \left[l\left(y_i, \hat{y}_i^{(t-1)}\right) + g_i f_t(x_i) + \frac{l}{2} h_i f_t^{2(x_i)}\right] + \Omega(f_t) \tag{7}$$

Where, g_i are the h_i $l(y_i, \hat{y}_i^{t-1})$ first and second derivatives of pairs, respectively, and let $\hat{y}_i^{(t-1)}$

$$G_j = \sum_{i \epsilon I_j} g_i , H_j = \sum_{i \epsilon I_j} h_i \tag{8}$$

The simplified objective function is:

$$L^* = -\frac{1}{2} \sum_{j=1}^{T} \frac{G_j^2}{H_j + \lambda} + \gamma T \tag{9}$$

Finally, calculate the gain before and after splitting:

$$Gain = \frac{1}{2}\left[\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} \frac{(G_L + G_R)^2}{H_L + H_L + \lambda}\right] - \gamma \tag{10}$$

The higher the value of Gain, the more L decreases after splitting. Therefore, when a leaf node is divided, the Gain corresponding to all candidate features is calculated, and the corresponding feature with the largest Gain is selected for segmentation. By accumulating the prediction results of all decision trees, the final model prediction is obtained [9].

3. Analysis of the Results

3.1. Results of KNN

The Sun's cycle of activity is strongly correlated with changes in its magnetic field and polarity. The Sun's magnetic field reverses over about 11 years, from the South Pole to the North Pole and back again. At the peak of the activity cycle, there are many active sunspots and sunspots on the surface of the sun, which are accompanied by strong solar activity such as solar wind and cosmic rays. These

solar activities have an impact on the Earth's climate and communication systems, among other things. Therefore, it is very important to understand the solar cycle and its associated changes in the magnetic field.

This paper collects historical data including the number and area of sunspots, ensuring that the data includes the data and time stamp, the number of sunspots and the sunspots area in Solar Cycle Science. By collecting the relevant data of the past 24 periods and processing the data from Excel, the curve line chart of the change of the relevant period years is obtained after plotting, as shown in Figure 1. Through graphical observation, it can be seen that the cycle years of the sun's movement show a fluctuating trend as a whole, but there is a certain regularity. Among the 24 cycles, the longest length of the year in a cycle is no more than 15 years, the shortest is no less than 8 years, and the average length of the year is between 10 and 12 years [10]. Locally analyzing the graph, it can be found that the first 8 cycles fluctuate greatly, the length of the years in the 9-15 cycles tends to be stable, and after the 14th cycle begins, the length of the cycle time fluctuates, compared with the first few cycles, the degree of fluctuation change is smaller, the variance is smaller, the change tends to be steady, and there is a certain regularity.

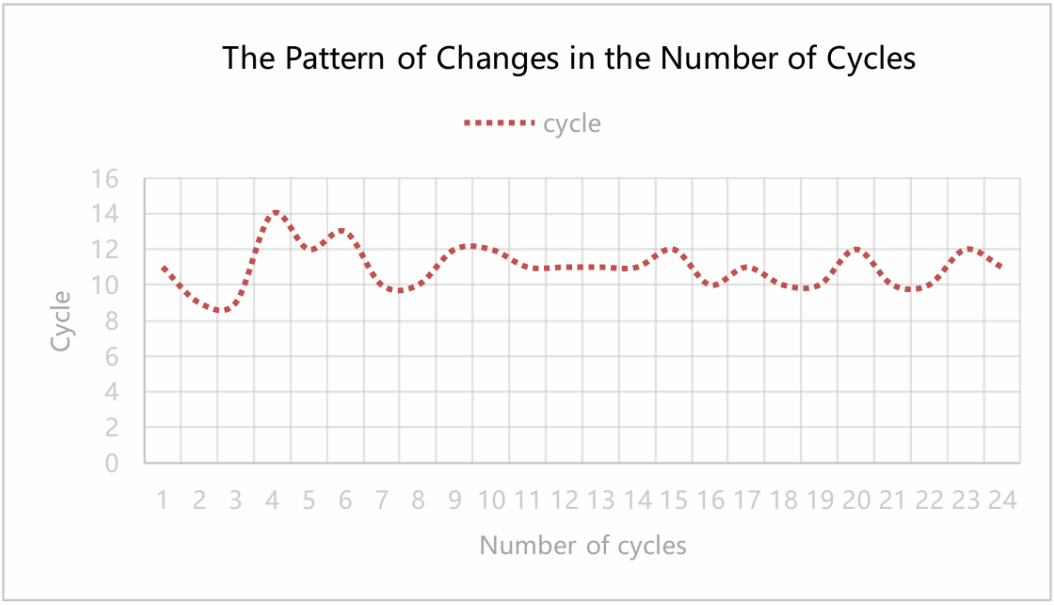

Figure 1. The pattern of changes in the number of cycles.

Firstly, the above model is used to establish K the nearest neighbor KNN regression model through the training set data. After shuffling the data, divide 70% of the data into the training set and 30% of the data into the test set. Use the GBDT-based learner. The L1 regular term is 0 and the L2 regular term is 1. And use 10-fold cross-validation, and continuously train it.

Table 1. Model evaluation results of KNN.

	MSE	RMSE	MAE	MAPE	R^2
Training set	1.25	1.118	0.925	8.475	-0.927
Test set	2.01	1.418	1.05	9.611	-0.937

Table 1 shows the prediction evaluation indicators of the cross-validation MSE set, the training set, and the test set, RMSE and quantifies the MAE prediction effect of MAPE neighbors (R^2) by calculating, sum, and K quantifying these five parameters KNN. The value of the K nearest neighbor (KNN) regression model is 0.925, indicating that the training effect is the best, and the MAE value of the training set is -0.927, indicating R^2 that the model is more reasonable and the fitting effect is better. Therefore, regression models are used to predict the beginning and end of the current and next solar cycles.

Figure 2 shows the prediction of the test data by K the nearest neighbor, the blue line represents the true value, and the green line represents the predicted value KNN, which shows that the deviation between the predicted value and the true value is small, the model error is small, and the fitting effect performs well.

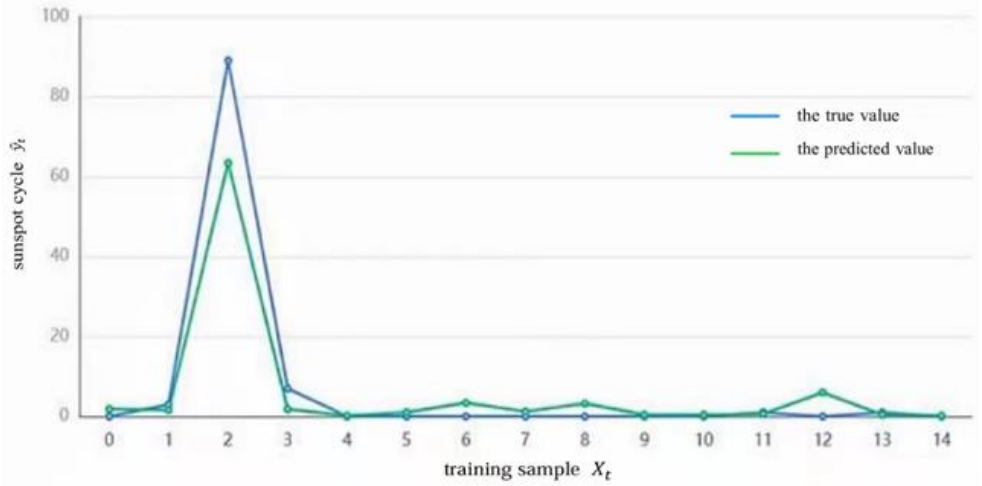

Figure 2. K-nearest neighbors (KNN) predictions of the test data.

From Table 2, it can be concluded that the end time of the current cycle is around June 2030, the start of the next solar cycle is around June 2030, and its end time is around July 2040.

Table 2. Result table of forecast end time.

	Solar Cycle Start Time Prediction	End Time Forecast
The Current Period	December 2019	June 2030
Next Cycle	June 2030	July 2024

3.2. Result of XGBoost

In the process of establishing the XGBoost regression model, firstly, after the data shuffling, 70% of the data is extracted as the training set, and the remaining part is used as the test set, and the XGBoost regression model is established by using the training set data retrieved from the overall data. After the XGBoost model is established, the feature importance of the model is calculated, and the established XGBoost regression model is applied to the training and test data, and then the evaluation results of the XGBoost model are obtained, and the evaluation results are analyzed. The L1 regular term is 0 and the L2 regular term is 1. Since the XGBoost model is random, the effect of each operation is different, and if the model has been determined, it can be used multiple times in the subsequent training model.

After the XGBoost model is established, the relevant values of the corresponding model symbolic fitting effect are calculated and obtained, such as MSE and RMSE these five parameter values, and quantitative indicators are used to measure MAE the MAPE prediction effect of the R^2 XGBoost model [11]. Among them MSE RMSE and MAE, is 0.001 0.035 and 0.012. The smaller the values are, the higher the accuracy of the model is. R^2 is 1, the accuracy of the model is high, indicating that the model is more reasonable and the fitting effect is better. Therefore, the XGBoost regression model is used to predict the start time of the solar maximum for the next solar cycle. The results are counted in Table 3.

Table 3. Model evaluation results of XGBoost.

	MSE	RMSE	MAE	MAPE	R^2
Training set	0.001	0.035	0.012	0.056	1
Test set	1.28	1.131	1.092	13.067	0.986

Figure 3 illustrates the start time of the solar maximum for the next solar cycle predicted by the XGBoost model.

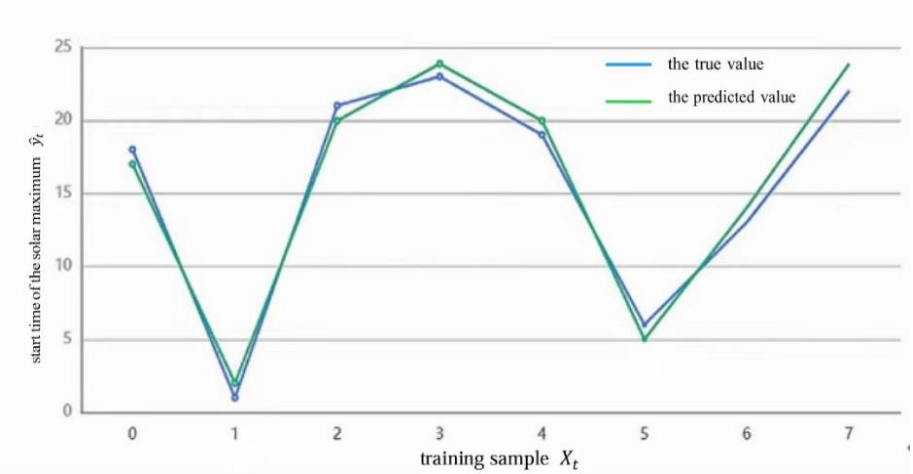

Figure 3. Model prediction effect.

Through the observation of the images, it is found that the line chart of the true value and the predicted value is equal or approximately equal in terms of the changing trend and the size of the specific value, and within the allowable error range, it can be approximately regarded as the true value and the predicted value are equal everywhere, which shows that the fitting effect of the XGBoost model is excellent and the prediction result is reasonable.

Solar activity statistics began in 1700 and have gone through 24 cycles since then. The maximum duration of solar motion for 24 periods is collected, and a line chart, Figure 4, is drawn to observe its trend and predict whether there is a periodic change pattern.

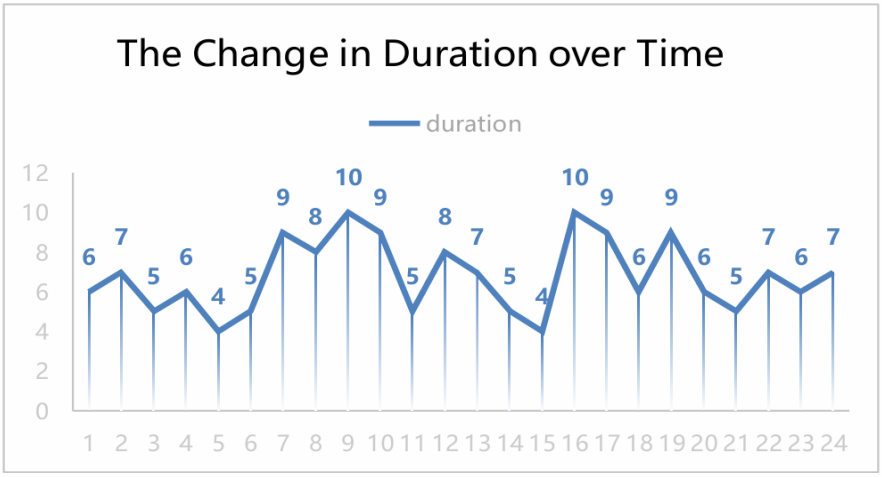

Figure 4. The Duration of the Solar Maximum of the Solar Cycle.

From Figure 4, it can be seen that the duration of the solar maximum of the solar cycle fluctuates greatly, and the time of the first few cycles is relatively short. From the fifth cycle, the duration of the solar maximum begins to rise sharply, reaches the peak in the ninth cycle, and then begins to fluctuate up and down in the interval of 4-10. The general trend is similar to a trigonometric function, with a certain periodic change law. In terms of local observations, after the 11th period, every four groups are taken as a group, and it is found that there are peaks and small values in each group. It can be inferred that the duration of the solar maximum of the solar cycle is regular, and the duration of the solar maximum of the subsequent cycle can be predicted by model prediction.

The XGBoost model achieves high level through the model correlation rationality test. Therefore, the XGBoost regression model is used to predict the duration of the solar maximum for the next solar cycle. The results of modeling test are shown in Table 4.

Table 4. Model evaluation results of XGBoost regression.

	MSE	RMSE	MAE	MAPE	R^2
Training set	0.001	0.023	0.016	0.221	1
Test set	2.502	1.582	1.352	23.467	-0.741

Figure 5 illustrates the duration of the solar maximum for the next solar cycle predicted by the XGBoost model.

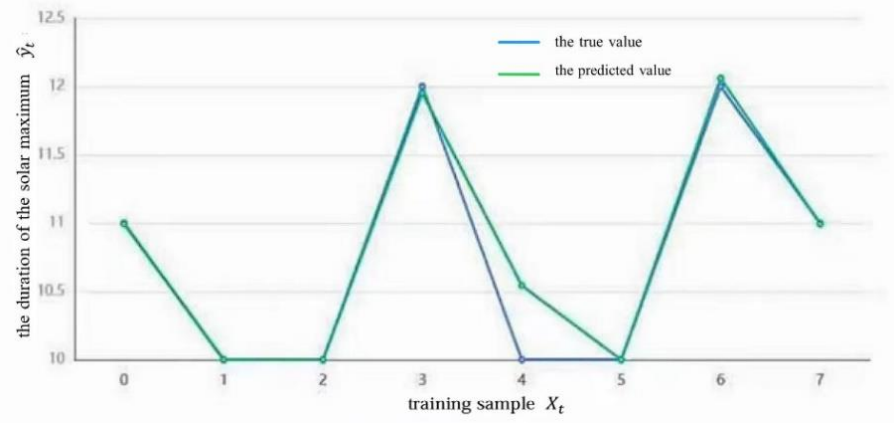

Figure 5. The model predicts the situation.

In Figure 5, the blue line represents the true value and the green line represents the forecast value. Through the observation of the images, it is found that the line chart of the true value and the predicted value has the same change trend in a certain interval of the ordinate, the degree of fitting is high, and the degree of numerical deviation is relatively small, which is in line with the analysis of the above line chart, and there is a certain regularity in the duration of the solar maximum of the solar cycle in a certain period in 24 cycles. It can be seen that the fitting effect of the XGBoost model is in a good state and the prediction results are reasonable. The forecast and prediction of the solar maximum start time are shown in the Table 5.

Table 5. Result Table of XGBoost model fitting.

	Solar Maximum Start Time _ Forecast	Duration _ Prediction
Next Cycle	August 2034	8

Table 5 displays the XGBoost model fitting result, the predicted start time of the solar maximum in the next cycle is August 2034, and the duration of it is 8 months.

It should also be noted that although the model solution and test effect of the XGBoost model are highly satisfactory, there will inevitably be some errors in the model establishment process. XGBoost cannot obtain definite equations like traditional models, and the model is usually evaluated by testing the prediction accuracy of the data, so the results of the model establishment and solution may deviate from the prediction results with reference to all the actual data.

4. Conclusion

In conclusion, sunspot prediction is of great significance for space weather, ionosphere state, short-wave radio communication, and satellite communication, so it has a wide range of promotion value in scientific research and technical applications. To predict the end of the current cycle date, this paper uses the KNN regression model. When exposed to new data, the KNN regression model changes to accommodate the new data points, so it is suitable to apply in multiple types of problems to better predict the given data. By collecting the relevant data of the past 24 cycles, the KNN regression model is established, predicting that the current cycle will end around June 2030, and the

next solar cycle will start around June 2030 and end around July 2040. The XGBOOST regression and time series models are established for prediction, and it is found that XGBOOST regression model has better fitting effect. The next solar cycle maximum is predicted to begin in August 2034 and last for eight months.

However, there are still some problems. The disadvantage of the KNN algorithm is that it has high memory requirements, because the algorithm stores all the training data, and the prediction stage may be slow, so it is sensitive to irrelevant functions and data size. Nevertheless, the XGBOOST regression model is used in this paper to predict the multi-sample data and can be extended to data analysis and prediction in the fields of economics, finance, meteorology, etc. The parameters can be optimized through appropriate algorithms and techniques, such as gradient descent method, grid search, and other methods, to improve the accuracy and stability of the model. Capture more data features and complex relationships to make more accurate forecasts of demand for different sample types.

References

[1] SUN Wei, YU XiaoXiao, LIU FuYao, ET al.2024. Numerical simulation and prediction of the sunspot magnetic field and polarity index [J]. Progress in Geophysics, 2024, 39(3): 885-895.

[2] ZHOU Meilin, ZHONG Libo. Sunspot Data Collection and Experimental Validation for McIntosh Classification [J]. Astronomical Research and Technology, 2023, 20 (04): 353-368.

[3] CHENG Shu, SHI Yaolin, ZHANG Huai. Prediction of sunspot changes based on neural network [J]. Journal of University of Chinese Academy of Sciences, 2022, 39 (05): 615-626.

[4] Lockwood M, Owens M J, Barnard L A, Application of historic datasets to understanding open solar flux and the 20th-century grand solar maximum. 1. Geomagnetic, ionospheric, and sunspot observations [J]. Frontiers in Astronomy and Space Sciences, 2022, 9:960775.

[5] Amrita P, Soumya R, Arindam S, et al. Prediction of solar cycle 25 using deep learning based long short-term memory forecasting technique [J]. Advances in Space Research, 2022, 69(01):798-813.

[6] Uddin, S, Haque, I, Lu, H. et al. Comparative performance analysis of K-nearest neighbour (KNN) algorithm and its different variants for disease prediction[J]. Scientific Reports, 2022, 12:6256.

[7] Zhang Shichao, Challenges in KNN Classification [J]. IEEE Transactions on Knowledge and Data Engineering, 2022, 34(10): 4663-4675.

[8] MU Meihong, ZHAO Gang, HE Bingshun, Qing Li, et al. XGBoost-based method for flash flood risk assessment [J]. Journal of Hydrology, 2021, 598: 126382.

[9] Mu Chengyu. Research on credit risk assessment model of small, medium and micro enterprises based on Xgboost regression [J]. Business News,2022(15):119-122

[10] Petrovay, K. Solar cycle prediction [J]. Living Reviews in Solar Physics, 2020, 17(2):1-93.

[11] Davide C, J M W, Giuseppe J. The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. [J]. Peer J. Computer science, 2021, 7 e623-e623.