Research on correlation analysis and prediction model of agricultural climate factors based on machine learning

Yun Deng*, *Cunliang* Cao, and *Shouxue* Chen

College of Information Science and Engineering Guilin University of Technology, Jiangan Road 12, Guilin, China

Abstract. This article uses machine learning technology to analyze the correlation of climate factors that affect crop yields, and conduct prediction and comprehensive evaluation to guide agricultural production. This paper selects early rice crops in Guangxi as the research object. Based on the climatic data of early rice planting areas in Guangxi from 1990 to 2017, a cart decision tree is constructed to generate a random forest model to analyze the correlation between early rice yield and climatic factors in each growth period, and obtain the various growth periods The ranking of the importance of climatic factors on the yield, thus forming the basis for calculating the weights of the climatic factors in each growth period of early rice; based on the climatic data in Guilin, Guangxi from 2008 to April to July 2017, predicted by the long and short-term memory network Guilin's various climate data from April to July 2018.

1 Introduction

Machine learning algorithms have demonstrated their capabilities in many fields. Applying machine learning algorithms to agriculture as well,reading large amounts of data and discovering internal laws and connections will also be helpful to agricultural production.Random forest model[1] through continuous training on the data set, analysis can draw the importance of variables to the dependent variable.Long short-term memory network algorithm (LSTM) is an improvement of recurrent neural network [2]. Its algorithm improvement effectively solves the former problem of gradient explosion. It is used in processing time series data and predicting unknown time series data.Has a very good effect [3].The purpose of training the data through the LSTM model is to discover the inherent laws in the time series data,so as to accurately achieve the prediction results,and make the weather forecasting of agricultural production more accurate.

2 System model design

2.1 Establish cart decision tree model

* Corresponding author : 540035535@qq.com

This article establishes the CART classification algorithm model [4,5,6], the basic idea is:

(1)for each early rice meteorological yield fluctuation data A, for all its possible climatic factor data a in each growth period, divide the data set into two subsets, A=a and A!=a, according to the early rice meteorological yield fluctuation data A Is it possible to take the climate factor data a during a certain growth period and divide it into two parts D_1 and D_2.

$$D_1 = \{(x, y) \in A(x) = a\}, D2 = D - D_1 \quad (1)$$

Under the conditions of early rice meteorological yield fluctuation data A, the Gini index of set D is defined as:

$$Gini(D, A) = \frac{|D_1|}{D} Gini(D_1) + \frac{D_2}{D} Gini(D_2) \quad (2)$$

(2)Traverse all the early rice meteorological yield fluctuation data A, calculate the Gini index of the value a of the climate factor data of all possible growth periods, and select the minimum Gini index e of D to make the data divided into two subsets to have a better effect. The corresponding features and segmentation points are regarded as the optimal division.

(3)To get a new node, call the above steps (1)(2) recursively, until the set conditions are met.

(4) Generate CART decision tree.

2.2 Establish random forest model

In order to obtain the correlation between the meteorological factors in Guangxi and the influence of early rice yield in Guangxi, and to obtain the importance ranking of the characteristic factors, this paper needs to establish a random forest model [7,8,9]. The main idea of the random forest algorithm is shown in **Algorithm 1**: Random forest algorithm.

Input: training set $D = \{(x_1, y_1), (x_2, y_2), (x_3, y_3) \cdots (x_n, y_n)\}$,a total of 140 sample sets.

Output: obb_Score, importance ranking.

(1)Selection of sample set. The 140 sets of acquired data are formed into a sample set, training samples and test samples are divided, and the sample set is extracted with replacement in sequence until a training set of size 140 is obtained. A total of 1000 rounds of extraction are performed,Then the training sets drawn in each round are respectively $T_1, T_2, \cdots, T_{1000}$.(2) The generation of decision tree. There are a total of 24 feature factors in each growth period of early rice. In each round of decision tree generation, d of the 24 features is randomly selected to form a new feature set, and the new feature set is used to generate Decision tree.(3) The combination of models. Since the characteristics of decision trees are independent of each other, their respective importance is also equal, and finally the classification situation is determined by voting on all decision trees.(4) Train the random forest model and save the training model.(5) Verification of the model. Using a cross-validation method, with simple majority voting as the classification result of the sample, the oob_Score score reflects the generalization ability of the random forest model, and outputs the ranking of the importance of meteorological factors.

2.3 Establish long and short-term memory network model

Taking the historical climate data of Guilin in Guangxi as an example, a long and short-term memory network model was established, and the daily average temperature changes during the growth and seedling period of early rice were selected for further analysis and demonstrated in the article to establish a long- and short-term memory

network model [10]. The steps to achieve prediction data through the LSTM model are shown in **Algorithm 2**:LSTM algorithm.

Input: daily average temperature change data from 2017-06-01 to 2017-08-31.

Output: Epoch, loss_, train_score, test_score, predict.

(1)read the csv file, standardize the daily average temperature change data from 2017-06-01 to 2017-08-31, and divide these data into training data and test data.(2)use the standardized data as the input data of the LSTM algorithm.(3)define the LSTM network, define the training LSTM function, and set the values of the parameters in the LSTM function.(4) Using grid structure and optimization algorithm, after LSTM model training, the loss value is obtained. By continuously reducing the loss value, the model is made more accurate. After 100 model iteration training, the model training score and test score are evaluated respectively.(5)save the LSTM model.(6)define the LSTM prediction model function, use the saved LSTM training model, set the step size of the predicted time series, restore the predicted results, and output the predicted data and visualize the data.

3 Experimental results

3.1 Data processing

Select the early rice yield statistics data per hectare per hectare in these five regions in Guangxi from 1990 to 2017. First, use the 5a moving average method to decompose the early rice yield data of Guangxi cities into two parts,namely meteorological yield and trend yield. Meteorological output is calculated by formula to represent the fluctuation of climate output in different years. Obtain the daily temperature, precipitation, sunshine hours, wind speed climate factor data of these 5 regions from 1990 to April to July 2017, and then calculate the climate factors according to the four growth periods of early rice, and then calculate all the climate factors The data is processed according to different growth periods, and each meteorological factor adopts the average value of each growth period, and the five influencing factors are processed with a 5a moving average anomaly percentage, and the high temperature is expressed by the cold accumulated temperature of each growth period. The unit is degrees. Correspond the processed weather data with the fluctuating output data one by one.

3.2 Random forest experiment result

After sorting out the original data, this paper selects the relative meteorological yield of early rice in Guangxi as the dependent variable, taking seedling stage (Seed), turning green tilling stage (Tiller), jointing booting stage (Head), and heading and setting stage (Mature). Average temperature anomaly (TJ), sunshine hour anomaly (XJ), cold accumulated temperature of freezing injury (ACT), high temperature accumulated temperature (AHT), average wind speed anomaly (SJ), precipitation anomaly (WJ) during the growth period The 24 factors of are independent variables. All the processed data sets of 140 groups were randomly divided into experimental data and test data at a ratio of 3:1, the correlation ratio between climate factors and early rice yield obtained by the random forest model was shown in the picture as shown in Figure 1.

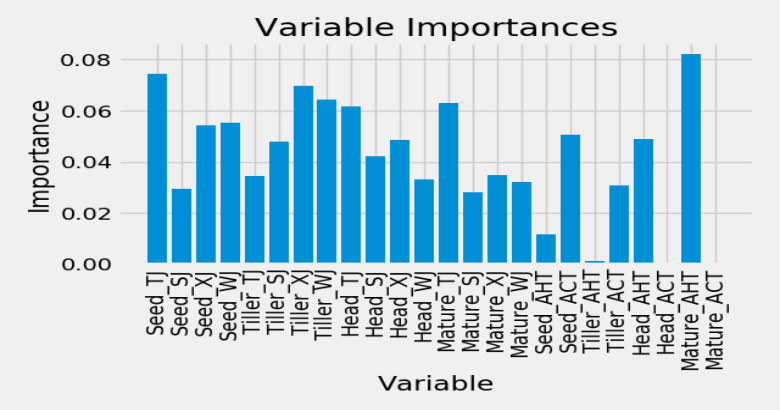

Fig. 1. Correlation diagram of climatic factors affecting early rice yield.

In the figure, the x-axis represents variables, which are different climate factors in each growth period, and the y-axis represents the importance of the variable's influence on the early rice yield trend. It can be seen from the figure that the high temperature at the heading and fruiting stage of early rice has the greatest impact on the output of early rice. From this analysis, it is concluded that the high temperature at the maturity stage of early rice has the greatest impact on the output of early rice due to the failure to harvest in time. From the overall analysis, the environmental factors that have the greatest impact on yield are precipitation and high temperature. Early rice seedling stage temperature and precipitation account for a larger proportion, and the number of sunshine hours and precipitation during the returning green tillering period account for a larger proportion. The temperature and precipitation at the jointing and booting stage have a greater influence; the two meteorological factors, the temperature at the heading stage and the number of sunshine hours, have a greater weight; the growth and development of the seedling stage is greatly affected in terms of the accumulation of low temperature. On the other hand, the jointing and booting stage and the heading stage are longer. In the whole period of early rice growth and development, the high temperature and accumulated temperature factors in the mature period have the greatest influence on the yield trend, and the wind speed factor has little effect on the growth and development of early rice.After sorting out the above analysis, the results are shown in Table 1. It can be intuitively seen that the types of main climatic factors affecting the output of early rice in Guangxi during each growth period.

Table 1. The main climatic factors affecting the yield of early rice.

Various growth periods of early rice	Main climatic factors affecting yield
Seedling stage	T Temperature and rainfall field
Sillering stage	Sunshine hour and rainfall field
Jointing and booting stage	Temperature and sunshine hour
Heading-grain filling	Temperature and sunshine hour

3.3 LSTM experimental result

In the LSTM experiment, read the daily average temperature data of Guilin, Guangxi from 2008 to April 1 to April 30, 2017, and normalize the data to divide the data set into 4:1 Training data and test data, through the previously established LSTM model, setting the

parameters of the LSTM model, training the data set, continuously correcting the model by reducing the loss value and saving the trained LSTM model, through 100 iterations of the model, the saved LSTM training model as shown in figure 2.

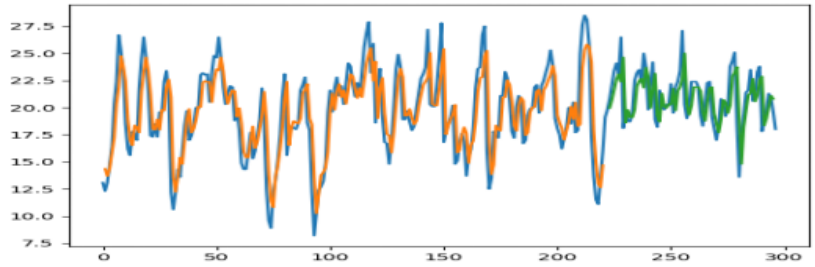

Fig. 2. LSTM training model diagram.

In the figure, the x-axis represents the time series data, the unit is day, and the y-axis represents the temperature value of the day in °C. The blue broken line represents the actual change of the historical data temperature, and the orange broken line represents the LSTM model. The trend graph of the training data, the green broken line represents the trend graph of the test data obtained by predicting the training data. The evaluation data of the LSTM model is shown in Table 2.

Table 2. LSTM model evaluation form.

Train_Score	Test_Score	Loss
2.53RMSE	2.25RMSE	0.0158

By calling the LSTM training function and outputting the results, it can be seen that the loss value continues to tend to a value. The loss is ultimately 0.0158. The smaller the loss value, the better the model training effect and the higher the accuracy. Through the predicted value and the actual value, The calculated error score of the training set is 2.53 RMSE, and the error score of the test set is 2.25 RMSE. It is proved that the trained LSTM model has achieved the expected effect. In the actual prediction, call the saved LSTM model, restore the model parameters, select the prediction time step as 30, and then predict the daily average temperature between April 01 and April 30, 2018 Changes. Display the predicted temperature change graph through the matplotlib.pyplot module. The result is shown in Figure 3.

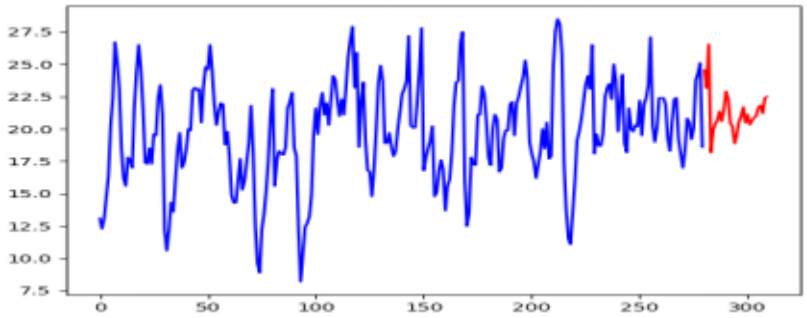

Fig. 3. The temperature change map in the next 30 days predicted based on the LSTM algorithm.

Refer to the above method to predict the wind speed, sunshine hours and precipitation of Guangxi early rice in each growth period. The detailed process will not be listed one by one. The predicted weather data in Guilin, Guangxi from April to July 2018 will be saved , Part of the data is shown in Figure 4.

time(yyyy-mm-dd)	Avg_Tem(°C)	Avg_Wind(m/s)	Avg_Sun(h)	Avg_Prec(mm)
2018-04-01	23.8	1.5	0	3270
2018-04-02	24.5	4.3	0	0
2018-04-03	23.2	3.2	0	3270
2018-04-04	26.5	3.1	0	0
2018-04-05	18.2	2.5	10.5	0
2018-04-06	20	1.8	9.3	0
2018-04-07	20.4	1.3	1.6	0
2018-04-08	20.7	2.1	0	3270
2018-04-09	21.4	1.8	0	0
2018-04-10	20.6	1.3	0	3270
2018-04-11	23.5	0.8	0.5	232
2018-04-12	20.3	0.7	11	14
2018-04-13	18.9	1.5	5.7	0
2018-04-14	20.3	1.9	4.8	0
2018-04-15	17.6	2.5	9.2	0
2018-04-16	19.6	2.6	0	3270

Fig. 4. Predicted weather data of Guilin from April to July 2018.

The LSTM predicts the average data of various climate factors in Guilin from April to July 2018, and compares them with the actual average data of each climate from April to July 2018, reflecting the accuracy of the data predicted by the LSTM model, as shown in Table 3. Data comparison table.

Table 3. Forecast and actual data comparison table.

Climatic factor	Prediction value	Actual values	Accuracy rate
average temperature(° C)	24.7	23.9	96.7%
average wind speed (m/s)	1.8	1.7	94.4%
average sunshine time(h)	4.1	4.3	95.3%
average precipitation(mm)	1188.2	1267.7	93.8%

Comparing the predicted average climate data with the actual average climate data, the accuracy rate is higher than 93%. It can be seen that the LSTM model predicts that the climate data of Guilin from April to July 2018 is accurate.

4 Summary

This study first introduced the random forest experiment in detail, visualized the importance of climatic factors on the yield of early rice in Guangxi, summarized the main climatic factors affecting early rice yield during each growth period of early rice, and analyzed the reasons affecting the yield of early rice in combination with actual climatic conditions. The results of the experiment have practical significance. Based on the climatic data of Guangxi Guilin area from 2008 to 2017 from April to July, the data changes of various climatic factors from April to July of 2018 are predicted, and the data is compared with the actual climatic data of 2018 to get the predicted data The accuracy rate of this research has important theoretical significance and research value for guiding agricultural production.

Thanks to the research financially supported by Guangxi Science and Technology Plan Project.

References

1. Peter Harrington. Machine learning in practice [M]. 2013.
2. Li Jie, Lin Yongfeng. Time series data prediction based on multi-time scale RNN[J]. Computer Applications and Software, 2018, 35(07): 33-37+62.
3. Zhao Deyu. Summary of Deep Learning and Deep Reinforcement Learning [J]. China New Communications, 2019, 21(15): 174-175.
4. Wang Guozhong, Li Zhongyuan, Zhang Jiyu, Zuo Qiting, Cheng Huanling. Analysis of water quality influencing factors of main reservoirs in Henan Province based on decision tree[J]. Journal of Wuhan University (Engineering Edition), 2019, 52(09): 774-781.
5. Xu Xuran, Tu Juanjuan. Air quality prediction system based on decision tree algorithm [J]. Electronic Design Engineering, 2019, 27(09): 39-42.
6. Zhang Bowen, Cui Linli, Shi Jun, Wei Peipei. Research on decision tree classification based on characteristic bands of rice[J]. Anhui Agricultural Sciences, 2017, 45(28): 207-210.
7. Lu Weixue, Wu Hecheng, Wan Liyang. Prediction of precipitation based on PLS fusion random forest algorithm[J]. Statistics and Decision, 2020, 36(18): 27-31.
8. Yang Beiping, Chen Shengbo, Yu Haiyang, An Qin. Remote sensing estimation of rice yield based on random forest regression method[J]. Journal of China Agricultural University, 2020, 25(06): 26-34.
9. Xia Xiaosheng, Chen Jingjing, Wang Jiajia, Cheng Xianfu. Analysis of factors affecting PM_(2.5) concentration in China based on random forest model[J]. Environmental Science, 2020, 41(05): 2057-2065.
10. Zhao Mingzhu, Wang Dan, Fang Jie, Li Yan, Mao Jun. Temperature prediction of subway station based on LSTM neural network [J]. Journal of Beijing Jiaotong University, 2020, 44(04): 94-101.