Agricultural Irrigation Area Prediction Based on Improved Random Forest Model

Guangda Gao

China University of Geosciences

Maofa Wang (✉ wangmaofa2008@126.com)

Guilin University of Electronic Technology

Hongliang Huang

Zhuhai College of Jilin University

Weiyu Tang

Zhuhai College of Jilin University

Research Article

Keywords: Random Forest, Extra Tree Regression, Irrigation area, Parameter Adjustment, Cross validation

Posted Date: February 1st, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-156767/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Agricultural Irrigation Area Prediction based on Improved Random Forest Model

Guangda Gao^{1, #}, Maofa Wang^{2, *, #}, Hongliang Huang^{3, *, #}, Weiyu Tang³

¹ School of Information Engineering, China University of Geosciences (Beijing), Beijing, China, 100083

(gaogd@cugb.edu.cn)

² Guangxi Key Laboratory of Trusted Software, Guilin University of Electronic Technology, Guilin 541004, China

(wangmaofa2008@126.com)

³ School of Computer, Zhuhai College of Jilin University, Zhuhai 519041, China

(huanghongliang2020@126.com, tangweiyu2000@stu.jluzh.edu.cn)

These authors contributed equally to this work

* Corresponding authors: Maofa Wang (wangmaofa2008@126.com, huanghongliang2020@126.com)

Abstract: The food problem is a major problem of common concern in the world, and the prediction of irrigation area can promote the solution of food and agricultural problems. In this paper, the data of grain production and irrigation area in the world are analyzed. An improved Random Forest Regression model is proposed and applied to the prediction of irrigation area. Based on ordinary Random Forest and Limit Tree Regression algorithm, an improved random forest prediction model for irrigation area in China is proposed. Firstly, the arithmetic mean value (AMM) of mean square error (MSE) and mean absolute error (MAE) was used as the evaluation index of the improved impure function and irrigation area prediction effect. Then, the grid search method is used to determine the optimal number of decision trees (70 trees and 30 trees respectively) in ordinary random forest and limit tree regression, and a new improved random forest model is established. After following, the model is compared with other prediction models, and 10 fold cross validation shows the rationality of the model. Finally, the error analysis of the improved Random Forest model shows that the prediction error is small. It is expected to be applied in the annual analysis of irrigation area in China.

Keywords: Random Forest; Extra Tree Regression; Irrigation area; Parameter Adjustment; Cross validation

0 Introduction

In recent years, the problem of food production and food supply has once again become the focus of attention all over the world. Agriculture is the source of food^[1]. Irrigation area is an important factor that determines the development of agriculture and grain harvest. Irrigation is an important land management measure in agriculture^[2]. The prediction of irrigation area is of great help to the development of agriculture.

Irrigation area is an important basic agricultural data in the database of Food and Agriculture Organization of the United Nations (FAO). This paper establishes the prediction model of irrigation area based on the data provided by FAO.

Random Forest regression is an important prediction algorithm in data analysis and data mining^[3]. Random Forest regression model can solve nonlinear problems. In the field of machine learning, random forest is a classifier which contains multiple decision trees, and its output category is determined by the mode of the output categories of multiple trees. The concept of random forests evolved from Random Decision Forests proposed by Ho of Bell Laboratories in 1995^[4]. Then Leo Breiman and Adele Cutler^[7] developed and deduced random forest algorithm. Combined with Breimans' idea of "Bootstrap aggregation" and Ho's^[4] "Random Subspace Method", the set of decision trees is constructed. According to the set of decision trees, the random forest is obtained. The output of the forest is Bagging or Boosting of the output results of multiple trees. This model can solve the problem of nonlinear prediction and avoid the problems of large error and over fitting.

Based on the random forest and limit tree regression algorithm, this paper improves its impure function, obtains a new evaluation index Average – MSE – MAE, and studies it. At the same time, an improved prediction method for irrigation area is obtained^[8].

The work and innovation of this paper are as follows:

- (1) Based on Random Forest and Limit Tree regression model, the impure function and Random Forest model of CART algorithm are improved.
- (2) Combined with grid search, the number of decision trees is adjusted to get better parameters.
- (3) The cross validation method is used to evaluate and train the model for many times, which reduces the probability of model error.
- (4) Combined with Random Forest and Limit Tree regression prediction model, a new improved random forest model is proposed and compared with other prediction models.
- (5) It is feasible to select China as a significant agricultural country.

1 Materials and methods

1.1 Grain types and production analysis

This study is based on two agricultural data from FAO statistical database (FAOSTAT). They are the detailed data of agricultural irrigation area from 1961 to 2017 and food products of various countries in the world from 1961 to 2007. In order to study and predict the national irrigation area (unit:1000 Ha) and establish the prediction model of irrigation area, this paper is based on China, a major grain producing country. This research is mainly divided into the following steps: raw data preprocessing, screening of important agricultural data in China, establishment of mathematical prediction model, experimental evaluation, improvement of prediction model, experimental evaluation and comparative analysis. The simplified process is shown in Fig.1.

Fig.1 Research simplified flow chart

First of all, through Matplotlib and Seaborn, the statistical analysis of the world's food types is carried out, as shown in Fig.2. It can be seen that the top three of the world's total grain production are: excluding butter, eggs, excluding beer and excluding butter, which are far more than the total number of other grains. This also shows that the demand for dairy products is relatively large for all countries in the world. According to the actual situation, we can know that the production and planting of grain cannot do without the irrigation area of the country. The following is the data exploration and statistical analysis of irrigation area in various regions of the world.

Fig.2 Top 25 agricultural products around globe

According to the original data of irrigation area collected by FAOSTAT, the data are processed by data De duplication, cleaning and transposition. Using Matplotlib and Seaborn for visualization, the change trend image of irrigation area is obtained, as shown in Fig.3. We can see that China ranks first in irrigation area and production quantity. According to the analysis and practice, China is a big grain producing country. Therefore, it is meaningful to study the irrigation area in China. The following is a more detailed data analysis of China's grain and irrigation area.

Fig.3 Top gross agriculture producing nations from 1961 to 2017

Based on the statistical analysis of China's grain change from 1997 to 2007 and data visualization, the growth trend of grain production in China is obtained, as shown in Fig.4. It can be seen that the grain production in China basically presents a linear trend of growth, and roughly presents a fluctuating growth. The demand for grain is increasing year by year.

In order to study and predict China's agricultural production, Random Forest model is improved based on Random Forest and Limit Tree algorithm. We studied, analyzed and predicted the irrigation area^[2].

Fig.4 China's production and India's production

1.2 Decision Tree Learning

The implementation of Random Forest regression needs to integrate Decision Tree algorithm. The concept of Decision Tree is introduced first.

Decision tree is composed of a decision graph and output results, which is used to create a plan to achieve the goal. In machine learning, decision tree is a prediction model. The decision tree is used as the prediction model to predict the class label of samples, which is also called classification tree or regression tree.

The goal of decision tree is to create mathematical model and use samples to predict the target value of samples. The training process of a decision tree^[10] is: according to the selected index, the proposed training set is split into multiple subsets, and then the multiple subsets are split continuously and recursively. This process is called recursive segmentation, until the subset gets the target value, and finally stop recursion.

The year is assumed to be the variable x , and the forecast variable Y is the irrigation area of China (target value). Then the data is presented as follows:

$$(x, Y) = (x_1, x_2, x_3, \dots, x_n, Y) \quad (1)$$

Where n is the n -th year.

Decision tree is mainly divided into regression tree and classification tree^[7]. It is usually called CART. In this study, the regression tree in the decision tree is used for prediction (the output of the regression tree is a real number).

CART algorithm can solve the problem of classification and regression prediction. The tree constructed by CART regression tree is a binary tree, and all the inputs that fall on the same leaf node have the same output. When constructing the decision tree learning model, each leaf node in CART regression tree represents the continuous predictive value of the target.

Selecting and measuring segmentation variables and points is an important step in decision tree learning. Generally, the impure of segmented nodes is used to measure, and there are two important impure functions in regression task. Next, two common impure function models are introduced ^[11].

(1) Mean Square Error

First, given the data set.

$$\mathbf{Data} = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_i, y_i), \dots, (\mathbf{x}_n, y_n)\} \quad (2)$$

Where $\mathbf{x}_i$ in $\mathbf{Data}$ is a b-dimensional vector. In other words, $\mathbf{x}_i$ contains b features.

The goal of regression prediction is to construct function $\hat{F}(x)$ and fit elements of data set. The calculation formula of mean square error^[13] of prediction model is Equation (3).

$$\mathbf{H}(\mathbf{x}) = \frac{1}{n} \sum_{k=1}^L (\hat{F}(\mathbf{x}_i) - y_i)^2 \quad (3)$$

In order to minimize the mean square error^[14], the objective function of minimizing the mean square error is established, namely Equation (4).

$$\min \frac{1}{n} \sum_{i=1}^n (\hat{F}(\mathbf{x}_i) - y_i)^2 \quad (4)$$

Where $\hat{F}(\mathbf{x}_i)$ is the predicted value and y_i is the real value.

Next, assume that a CART regression tree has L leaves, which means that the CART regression tree divides the input x into L units $\omega_1, \dots, \omega_L$ to obtain at most L target prediction values. By improving the objective function with minimum mean square error, the objective function with minimum mean square error of cart regression is obtained. It is shown in Equation (5).

$$\min \frac{1}{n} \sum_{k=1}^L \sum_{\mathbf{x}_i \in \omega_k} (\hat{C}_k - y_i)^2 \quad (5)$$

Where $\hat{C}_k$ is the predicted value of the k-th leaf obtained by CART regression prediction model.

Since the predictive value of multiple leaves can be obtained, the final predictive value is set as the mean value of each leaf containing the training set data, and Equation (6) can be obtained.

$$\hat{C}_k = \text{Average}\{y_i | \mathbf{x}_i \in \text{leaf } F_k\} \quad (6)$$

In order to minimize the mean square error of the tree model in the training set and the sum of the mean square errors of each leaf, it is necessary to select appropriate segmentation variables and points when dividing each node of the regression tree

The heuristic method is used to traverse all the segmentation variables and points, and the

minimum sum of the mean square error is selected for the segmentation. The m -th feature $x^{(m)}$ and the value r_m are selected as the segmentation variables and points. All the segmentation variables and points divide the set into two. Equation (7) and (8) are obtained.

$$\omega_1\{m, r\} = \{x | x^{(m)} \leq r\} \quad (7)$$

$$\omega_2\{m, r\} = \{x | x^{(m)} > r\} \quad (8)$$

According to cart, the optimal segmentation variable m and segmentation point r are obtained. Equation (9) is got as following.

$$\min_{m, r} \left[\min_{c_1} \sum_{x \in \omega_1\{m, r\}} (\hat{c}_1 - y_i)^2 + \min_{c_2} \sum_{x \in \omega_2\{m, r\}} (\hat{c}_2 - y_i)^2 \right] \quad (9)$$

$\omega_1\{m, r\}$ and $\omega_2\{m, r\}$ are obtained by solving.

At the same time, we get the following results that is Equation (10).

$$\hat{c}_j = \frac{1}{N_j} \sum_{x \in \omega_j\{m, r\}} y_i \quad (10)$$

Where $j = 1, 2$.

The segmentation variable m , segmentation point r and prediction value $\hat{c}_j$ are calculated repeatedly until the minimum mean square error is obtained. The input x is divided into L units $\omega_1, \dots, \omega_L$ to obtain the decision tree model. It is shown in Equation (11).

$$F(x) = \sum_{k=1}^L \hat{c}_k I(x \in \omega_k\{m, r\}) \quad (11)$$

The above is the analysis process of mean square error impure function.

(2) Mean Absolute Error

The mean absolute error is similar to the mean square error, which measures the prediction error by the square of the difference between the error value and the true value. Mean absolute error^[15] measures the prediction error by the absolute value of the difference between the error value and the true value. Equation (12) for calculating the mean square error of cart regression tree is as follows.

$$H(x) = \frac{1}{n} \sum_{k=1}^L |\hat{c}_k - y_i| \quad (12)$$

The objective function of cart regression tree is the minimum mean absolute error^[16]. So we get Equation (13).

$$\min \frac{1}{n} \sum_{k=1}^L \sum_{x_i \in \omega_k} |\hat{c}_k - y_i| \quad (13)$$

The procedure of selecting appropriate segmentation variables and points is similar to the mean square error.

In order to improve the prediction model of China's irrigation area, this study is based on China's irrigation area data and years. Based on ordinary Random Forest and Limit Tree regression, an

improved Random Forest model is established.

1.3 Improved Random Forest model based on ordinary Random Forest and Limit Tree algorithm

Bootstrap aggregation, also known as Bagging algorithm^[17], is a group learning algorithm in the field of machine learning^[18]. By integrating with regression algorithm, we can improve the accuracy and stability of prediction, reduce the variance of prediction results and avoid over fitting.

The principle of bagging algorithm is: given a training set of size, select a new training set of size by self-help sampling method (uniform and put back). In the new training set, we use the regression algorithm to get a model, and then get the average value of the model to get bagging results. The process of algorithm steps is shown in Fig.5.

Fig.5 Demonstration of bagging algorithm

Bagging algorithm is applied to tree regression training algorithm. First of all, given the training set $X = \{x_1, \dots, x_t\}$ and the target prediction value $Y = \{y_1, \dots, y_t\}$, bagging algorithm is used to repeat N times. The bootstrap sampling method is used to sample from the training set, and the training is carried out by sampling.

Table 1 Model training steps

Step:

For $n = 1, \dots, N$:

1. Sample, with replacement, n training examples from X, Y ; call there X_n, Y_n .
2. Train a regression tree F_n on X_n, Y_n .

According to the steps in Table 1, until the end of training, the predicted value of the target is the average value after the predicted values of all single regression trees are obtained. The model is shown under Equation (14).

$$\hat{F} = \frac{\sum_{n=1}^N F_n(X')}{N} \quad (14)$$

This is the bagging method of the original tree model.

(1) Random Forest

In order to get a better prediction tree model^[19], we improved it to random forest regression model^[6]. Through the improved learning algorithm, the feature random subset is selected in each candidate splitting and splitting process. This process is called Feature Bagging.

Feature Bagging can greatly reduce the influence of strong correlation between decision trees. In this paper, sklearn is used to call RandomForestRegressor to establish the prediction model of Random Forest.

Random Forest is a model integrated by multiple decision trees, as shown in Fig.6.

Fig.6 Random Forest Simplified

(2) Extreme Random Forest

Then, Limit Tree regression model is established. Similar to ordinary Random Forest, Limit Tree^[20] is also an integrated algorithm of single decision tree. But the difference is that for the split points, it is not by calculating the best point of each feature, that is, it is not based on the ordinary Random Forest method, but by random division. Among all the partition points, the point with the highest score is selected as the partition point of the node. For the regression prediction problem studied in this paper, the number of features selected by each node is the data quantity n (samples). In this paper, we use sklearn to call ExtraTreeRegressor to build the prediction model of extreme Random Forest (hereinafter referred to as Limit Tree).

Next, the improved random forest regression model and limit tree regression model are established for algorithm prediction, and the prediction effect is evaluated.

(3) Improved Random Forest Model

The data of irrigation area in China are standardized^[21] (dimensionless) to make the target value approximately obey $N(0,1)$ normal distribution. The standardized formula is as follows.

$$\tilde{y} = \frac{y - \bar{y}}{\sqrt{\sigma}} \quad (15)$$

Where $\sqrt{\sigma}$ is the standard deviation of China's irrigation area and $\bar{y}$ is the mean value of China's irrigation area.

According to the principle of Random Forest and Limit Tree regression prediction model, RandomForestRegressor and ExtraTreeRegressor were used for experimental evaluation. 75% of China's irrigation area from 1961 to 2017 was used as training set and 25% as test set.

For random forest regression tree, the most important parameter is the number of decision trees. In order to optimize the prediction model, the number of decision trees is adjusted to get better tree model parameters.

In order to improve the evaluation index model by considering both mean square error and mean absolute error, we define a Average – MSE – MAE model (abbreviated as AMM impure function). It is shown in Equation (16).

$$\text{INDEX} = \frac{\text{MSE} + \text{MAE}}{2} \quad (16)$$

The objective function is $\min\{\text{INDEX}\}$. The optimal model of the improved Random Forest is obtained by selecting the appropriate segmentation variables and points.

In order to calculate the optimal parameters, MSE and MAE with the number of decision trees of 1, 10, 20, 30, 40, 50, 60, 70, 80, 90 and 100 are simulated by the grid search method. Then 10 fold cross validation is carried out to get the impure function values of different parameters, as shown in Table 2. It can be seen from the above table that the prediction error of random forest is small. When the number of decision trees is 70, the mean square error and mean absolute error are the smallest, only 0.0083 and 0.0608. Finally, we calculate $\min\{\text{INDEX}_{\text{RF}}\} = 0.0346$.

The prediction effect of limit tree regression is also good. When the number of decision trees is 40, the mean square error is the smallest. When the number of decision trees is 30, the average absolute error is the smallest. When the number of trees is 30 and 40, combined with the mean square error and mean absolute error of the prediction model, it can be concluded that the mean square error and mean absolute error of 30 trees are smaller. We choose 30 trees as the optimal parameters, and the mean square error and mean absolute error are about 0.0036 and 0.0462. Finally, we calculate $\min\{\text{INDEX}_{\text{EF}}\} = 0.0249$.

Table 2 Error analysis based on parameter adjustment

n_estimators	Error	MSE of Random Forest	MAE of Random Forest	MSE of Extreme Forest	MAE of Extreme Forest
	Regression	Regression	Regression	Regression	Regression
1		0.0114	0.0745(Max)	0.0039	0.0493
10		0.0118	0.0712	0.0109	0.0696

20	0.0097	0.0619	0.0036	0.0464
30	0.0111	0.0682	0.0036	0.0462(Min)
40	0.0121(Max)	0.0729	0.0035(Min)	0.0467
50	0.0112	0.0662	0.0039	0.0484
60	0.0096	0.0652	0.0113	0.0734
70	0.0083(Min)	0.0608(Min)	0.0066	0.0572
80	0.0102	0.0656	0.0108	0.0706
90	0.0098	0.0653	0.0081	0.0598
100	0.0087	0.0628	0.0113(Max)	0.0737(Max)

In conclusion, the number of decision trees in random forest is 70, and the number of decision trees in limit tree regression is 30. Taking these two adjusted parameters as the better parameters of the research experiment, a new improved Random Forest model is finally established.

2 Experimental Evaluation

2.1 Comparative Analysis

In order to widely verify the effect of the improved random forest regression model established in this paper, it is compared with other prediction models^[22].

10 fold cross validation based on Sklearn. The model scores of KNeighbors, LinearRegression, DecisionTree, SVR, AdaBoost, GradientBoosting and original Bagging^[9] were obtained and compared^[23]. As shown in Table 3 .

Table 3 Ten fold cross-validation of nine models

Model \ Times											Average
	1	2	3	4	5	6	7	8	9	10	
Bagging	0.9898	0.9612	0.9880	0.9873	0.9392	0.9914	0.9816	0.9595	0.9959	0.9975	0.9791
ExtraTree	0.9942	0.9570	0.9821	0.9826	0.9557	0.9766	0.9655	0.9160	0.9928	0.9875	0.9710
RandomForest	0.9938	0.9690	0.9758	0.9862	0.9379	0.9944	0.9655	0.9841	0.9956	0.9972	0.9800
KNeighbors	0.9762	0.9468	0.9891	0.9765	0.9183	0.9672	0.9686	0.9777	0.9855	0.9673	0.9673
LinearRegression	0.7888	0.7682	0.6613	0.8460	0.8574	0.8598	0.7501	0.7174	0.8384	0.3869	0.7474
DecisionTree	0.9939	0.9386	0.9606	0.9807	0.9559	0.9949	0.9655	0.9433	0.9942	0.9958	0.9723
SVR	0.9557	0.9071	0.9591	0.9593	0.8822	0.9339	0.9383	0.9315	0.9759	0.8562	0.9299

AdaBoost	0.9861	0.9349	0.9569	0.9696	0.9449	0.9915	0.9561	0.9258	0.9905	0.9953	0.9652
GradientBoosting	0.9938	0.9386	0.9606	0.9807	0.9559	0.9949	0.9655	0.9433	0.9942	0.9958	0.9723

We visualized and compared the experimental model fitting results. At the same time, we used Sklearn and Matplotlib to evaluate the model, and get the image, as shown in Fig.7. The model scores of RandomForest Model and ExtraTree Model reached 99%. Compared with the other seven models, the model scored the highest. This shows that the integration of Random Forest and Limit Tree regression algorithm can reduce the error and enhance the prediction effect. There is a theoretical basis for this. It provides mathematical model support for predicting irrigation area.

Fig.7 Comparison and fitting curve of nine prediction models

2.2 Fitting Evaluation

The improved Random Forest regression model and Limit Tree regression model are used to fit the real value with the predicted value, as shown in Fig.8. The predicted value of the improved random forest and limit tree regression prediction model is basically consistent with the real value, which greatly reduces the prediction error and improves the prediction effect. Through the fitting experiment, it can be seen that the predictive value of random forest regression model is relatively high, while the predictive value of limit tree regression model is relatively low. In order to improve the prediction accuracy, a new improved random forest model is established. The model expression is Equation (17).

$$\hat{F}(x)' = \text{Average}\{RF(D = 70), LT(D = 30)\} \quad (17)$$

Where $\hat{F}(x)'$ is the predicted irrigation area, RF(D = 70) is the predicted value of Random Forest model with 70 decision trees, and LT(D = 30) is the predicted value of Limit Tree model with 30 decision trees.

Fig.8 Random Forest Regression with 70 decision trees and ExtraTreeRegressor with 30 decision trees

In order to better evaluate the effect of the improved random forest regression model in predicting irrigation area, k-fold cross validation was carried out^[24].

Let $k = 10$, based on the improved random forest regression model, 10 fold cross validation is carried out. The model determination coefficient^[22], mean square error^[13] and mean absolute error^[15] are obtained, as shown in Fig.9. After ten fold cross validation, the coefficient of determination of the improved random forest model is above 95%, and the mean value is 98%, and the mean square error is 135.52 (unit:1000 Ha). The average absolute error is only 6.08 (unit:1000 Ha). It shows that the effect of the improved random forest regression prediction model is significant.

Fig.9 R-Squared Value, Mean Square Error and Mean Absolute Error of improved Random Forest prediction model based on ten experiments

3 Conclusion

Based on the annual agricultural data provided by FAO, this paper analyzes the influencing factors of irrigation area. In this paper, mean square error MSE and mean absolute error MAE of random forest impure function are improved, and a new impure function (AMM) is defined. The number of optimal decision trees of common random forest and limit tree regression is determined by AMM and grid search method. Finally, a new improved random forest prediction model is established.

It is used to forecast the irrigation area of different countries in different years. Compared with other prediction models, the rationality of the model is verified. Through 10 fold cross validation, the improved prediction algorithm has the advantages of high fitting degree, low mean square error and mean absolute error, and strong adaptability to different training sets. Experiments show that the proposed improved prediction model has smaller error and better robustness. It can effectively predict the annual change of irrigation area in various countries, and has a certain application value for agricultural production in various countries.

Acknowledge

This work was supported by Guangxi Key Laboratory of Trusted Software (KX202049), Guangxi science and technology base and talent project (2020AC19134), National Natural Science Foundation of China (41504037), as well as the National key R & D plan of China (2018YFC1503806). In the process of the project, we got the professional help and guidance from Professor Li Jingchao, Professor Wang Chengxi, Dr. Zheng Xiao and other teachers from the development center of China Geological Survey. We also expressed our thanks for the valuable opinions on the algorithm used in the project.

[References]

- [1] Huang Y. Research status and applications of nature-inspired algorithms for agri-food production[J]. International Journal of Agricultural and Biological Engineering, 2020, 13(4): 1-9.
- [2] Towa J J, Xiangping G. Effects of irrigation and weed-control methods on growth of weed and rice[J]. International Journal of Agricultural and Biological Engineering, 2014, 7(5): 22-33.
- [3] Rencher, Alvin C.; Christensen, William F., Chapter 10, Multivariate regression – Section 10.1, Introduction, Methods of Multivariate Analysis, Wiley Series in Probability and Statistics 709 3rd, John Wiley & Sons: 19, 2012 [2019-05-14], ISBN 9781118391679
- [4] Tin Kam Ho. Random decision forests. Proceedings of 3rd International Conference on Document Analysis and Recognition (Montreal, Que., Canada: IEEE Comput. Soc. Press). 1995, 1: 278–282. ISBN 978-0-8186-7128-9. doi:10.1109/ICDAR.1995.598994
- [5] Tin Kam Ho. The random subspace method for constructing decision forests. IEEE Transactions on Pattern Analysis and Machine Intelligence. Aug./1998, 20 (8): 832–844.
- [6] Ho, Tin Kam. A Data Complexity Analysis of Comparative Advantages of Decision Forest Constructors (PDF). Pattern Analysis and Applications. 2002: 102–112 [2019-02-16].
- [7] Breiman, Leo; Friedman, J. H., Olshen, R. A., & Stone, C. J. Classification and regression trees. Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software. 1984. ISBN 978-0-412-04841-8.
- [8] Ling J, Xu J, Lin H, et al. Prediction and fusion algorithm for meat moisture content measurement based on loss-on-drying method[J]. International Journal of Agricultural and Biological Engineering, 2020, 13(4): 198-204.
- [9] Liu Y, Lyu Q, He S, et al. Prediction of nitrogen and phosphorus contents in citrus leaves based on hyperspectral imaging[J]. International Journal of Agricultural and Biological Engineering, 2015, 8(2): 80-88.

-
- [10] Quinlan, J. R., (1986). Induction of Decision Trees. Machine Learning 1: 81-106, Kluwer Academic Publishers
- [11] Zheng CHU, Jiong YU. Performance Prediction Based on Random Forest for the Stream Processing Checkpoint[J]. Journal of Electronics and Information Technology, 2020, 42(6): 1452-1459.
- [12] Liu Junming He Xiaotong Wang Pengxin, Huang Jianxi. Early prediction of winter wheat yield with long time series meteorological data and random forest method[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2019, 35(6): 158—166. (in Chinese with English abstract)
- [13] Lehmann, E. L.; Casella, George (1998). Theory of Point Estimation (2nd ed.). New York: Springer. ISBN 978-0-387-98502-2. MR 1639875.
- [14] A modern introduction to probability and statistics : understanding why and how. Dekking, Michel, 1946-. London: Springer. 2005. ISBN 978-1-85233-896-1. OCLC 262680588.
- [15] Willmott, Cort J.; Matsuura, Kenji (December 19, 2005). "Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance". Climate Research. 30: 79–82. doi:10.3354/cr030079.
- [16] Hyndman R J, Koehler A B. Another look at measures of forecast accuracy[J]. International journal of forecasting, 2006, 22(4): 679-688.
- [17] Breiman L. Bagging predictors[J]. Machine learning, 1996, 24(2): 123-140.
- [18] Breiman L. Random forests[J]. Machine learning, 2001, 45(1): 5-32.
- [19] H. Lan and Y. Pan, "A Crowdsourcing Quality Prediction Model Based on Random Forests," 2019 IEEE/ACIS 18th International Conference on Computer and Information Science (ICIS), Beijing, China, 2019, pp. 315-319, doi: 10.1109/ICIS46139.2019.8940306.
- [20] Geurts P, Ernst D, Wehenkel L. Extremely randomized trees[J]. Machine learning, 2006, 63(1): 3-42.
- [21] Dodge, Yadolah, and Daniel Commenges, eds. The Oxford dictionary of statistical terms. Oxford University Press on Demand, 2006.
- [22] Fen X, Honghai H, Xiaofeng D, et al. Nutritional compositions of various potato noodles: Comparative analysis[J]. International Journal of Agricultural and Biological Engineering, 2017, 10(1): 218-225.
- [23] K. Nugroho et al., "Improving Random Forest Method to Detect Hatespeech and Offensive Word," 2019 International Conference on Information and Communications Technology (ICOIACT), Yogyakarta, Indonesia, 2019, pp. 514-518, doi: 10.1109/ICOIACT46704.2019.8938451.
- [24] Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection[C]//Ijcai. 1995, 14(2): 1137-1145.
- [25] Chang J S, Luo Y F, Su K Y. GPSM: A generalized probabilistic semantic model for ambiguity resolution[C]//30th Annual Meeting of the Association for Computational Linguistics. 1992: 177-184
- [26] Silvano Di Zenzo. Pattern recognition: a statistical approach: Devijver P A and Kittler J Prentice-Hall, Englewood Cliffs, NJ, USA (1982) pp 448. 24.95[J]. Image and Vision Computing, 1985, 3(2):87-88.
- [27] de Almeida Bressiani D, Srinivasan R, Jones C A, et al. Effects of spatial and temporal weather data resolutions on streamflow modeling of a semi-arid basin, Northeast Brazil[J]. International Journal of

Agricultural and Biological Engineering, 2015, 8(3): 125-139.

- [28] Zia S, Du W, Spreer W, et al. Assessing crop water stress of winter wheat by thermography under different irrigation regimes in North China Plain[J]. International Journal of Agricultural and Biological Engineering, 2012, 5(3): 24-34.

Figures

Figure 1

Research simplified flow chart

Figure 2

Top 25 agricultural products around globe

Figure 3

Top gross agriculture producing nations from 1961 to 2017

Figure 4

China's production and India's production

Figure 5

Demonstration of bagging algorithm

Random Forest Simplified

Figure 6

Random Forest Simplified

Figure 7

Comparison and fitting curve of nine prediction models

a. RandomForestRegression: number of decision trees = 70 b. ExtraTreeRegressor: number of decision trees = 30

Figure 8

Random Forest Regression with 70 decision trees and ExtraTree Regressor with 30 decision trees

Figure 9

R-Squared Value, Mean Square Error and Mean Absolute Error of improved Random Forest prediction model based on ten experiments