Journal on Big Data
DOI:10.32604/jbd.2021.016317
Article

Tech Science Press

Grain Yield Predict Based on GRA-AdaBoost-SVR Model

Diantao Hu, Cong Zhang*, Wenqi Cao, Xintao Lv and Songwu Xie

Wuhan Polytechnic University, Wuhan, 430023, China
*Corresponding Author: Cong Zhang. Email: hb_wh_zc@163.com

Received: 30 December 2020; Accepted: 07 April 2021

Abstract: Grain yield security is a basic national policy of China, and changes in grain yield are influenced by a variety of factors, which often have a complex, non-linear relationship with each other. Therefore, this paper proposes a Grey Relational Analysis–Adaptive Boosting–Support Vector Regression (GRA-AdaBoost-SVR) model, which can ensure the prediction accuracy of the model under small sample, improve the generalization ability, and enhance the prediction accuracy. SVR allows mapping to high-dimensional spaces using kernel functions, good for solving nonlinear problems. Grain yield datasets generally have small sample sizes and many features, making SVR a promising application for grain yield datasets. However, the SVR algorithm's own problems with the selection of parameters and kernel functions make the model less generalizable. Therefore, the Adaptive Boosting (AdaBoost) algorithm can be used. Using the SVR algorithm as a training method for base learners in the AdaBoost algorithm. Effectively address the generalization capability problem in SVR algorithms. In addition, to address the problem of sensitivity to anomalous samples in the AdaBoost algorithm, the GRA method is used to extract influence factors with higher correlation and reduce the number of anomalous samples. Finally, applying the GRA-AdaBoost-SVR model to grain yield forecasting in China. Experiments were conducted to verify the correctness of the model and to compare the effectiveness of several traditional models applied to the grain yield data. The results show that the GRA-AdaBoost-SVR algorithm improves the prediction accuracy, the model is smoother, and confirms that the model possesses better prediction performance and better generalization ability.

Keywords: Grey Relational Analysis (GRA); Support Vector Regression (SVR); Adaptive Boosting algorithm (AdaBoost); grain yield prediction

1 Introduction

Grain security is an important strategic guarantee for the formation of China's "new development pattern" [1]. The sudden worldwide spread of the COVID-19 and the floods, locusts and man-made disasters in 2020 have not only caused a serious crisis for the world economy, but have also had a major impact on the world's grain security. In today's interconnected and mutually influencing "globalized" world, no country is immune, and the world economic recession and food security problems will inevitably have a certain negative impact on China's economic development and grain yield. In the study of grain security, it is often necessary to have information on the expected grain yield over a period of time to help the government and the grain industry to formulate measures in advance to deal with various situations that may arise, thus forecasting grain yield is important to ensure national grain security [2]. At present, domestic and foreign scholars have proposed various models for grain yield prediction currently [3–6]. Such as exponential smoothing prediction, linear regression model, BP neural network, time series prediction method, support vector machine (SVM) and weather prediction method, remote sensing prediction method, and so on. Among them, time series method, exponential smoothing prediction and linear regression model are the data processing methods based on linear model, while grain yield data tend to show non-linear characteristics. Based on traditional agriculture, there are weather prediction methods, remote sensing prediction methods. However, the longitudinal and latitudinal span of China is large and vast, and the prediction results using traditional agricultural methods are not generalizable. BP neural network algorithms suffer from slow learning and the tendency to fall into local optimal solutions, which affects the accuracy of prediction results. The support vector regression algorithm is a support vector machine for regression prediction, the penalty factor C and the selected kernel function directly affect the prediction results.

In order to accurately identify the interrelationships and statistical patterns of grain yield and its influencing factors. Grey Relational Analysis (GRA) has been shown by related scholars to be effectively applied in grain yield prediction [7–8]. GRA can solve the problem of correlation between non-linear data. Using data on projected food production for a given year as an example. A number of factors such as the area of cultivated land, the amount of fertilizer used in agriculture, the total power of machinery, the number of agricultural reservoirs and the area affected by the disaster are generally taken as input variables, and food production is taken as output. However, because the relationship between grain yield and its influencing factors is complex and some of the influencing factors have little impact on the final outcome, it is not reasonable to use all the influencing factors directly. First, the GRA method is used to normalize each feature to derive the correlation coefficients, and only the more highly correlated impact features are retained to improve the prediction sample relevance. Secondly, the kernel function in Support Vector Regression (SVR) can be used to map to higher dimensional spaces [9], which is a good solution for nonlinear problems. Better fit to small samples, more adaptable to the inadequate sample size of grain yield data. Finally, the SVR is simultaneously combined with the Adaptive Boosting (AdaBoost) algorithm to train multiple base learners as a basic training method [10]. Ultimately, the strong learners are obtained by some combination strategy. The generalization capacity is further tuned to ensure the accuracy of the grain yield prediction model. This will provide a reliable basis for China's grain production policy.

The remainder of this article consists of the following sections: Section 2 details the grey correlation analysis algorithm, the support vector regression algorithm, and the adaptive enhancement algorithm, also combined algorithmic model. Section 3 describes the application of the GRA-AdaBoost-SVR model for grain yield forecasting in China and compares it with traditional algorithms. Section 4 is a summary of the work and the outlook for the future.

2 GRA-AdaBoost-SVR

2.1 Grey Relational Analysis (GRA)

The gray system is relative to the white system and the black system. It was originally proposed by Deng Ju long, a professor of control science and engineering. According to cybernetic conventions, color generally represents the amount of information we know about a system. White represents sufficient information. A mechanical system, for example, where the relationships between the elements can be determined, is a white system. A black system, on the other hand, represents a system whose structure is unknown to us. The gray color is somewhere in between, indicating that we only have a partial understanding of the system [11].

Grey Relation Analysis (GRA) is a method of statistical analysis of multiple factors. In a grey system, if you want to know how strongly or weakly an item is influenced by other factors [12], you can use some mathematical methods to obtain an analysis result that shows which of the factors is more relevant to a particular indicator.

The calculation process and processing steps for grey correlation analysis of the grain yield dataset can be decomposed as follows:

Grain yield is assumed to be the reference series.

$$X_0' = (X_0'(1), X_0'(2), \cdots, X_0'(m))^T \tag{1}$$

The influencing factors are the comparison series.

$$X_i' = (X_i'(1), X_i'(2), \cdots, X_i'(m))^T, i = 1,2,\cdots,n \tag{2}$$

where n is the number of characteristic data series and m is the number of indicators.

Pre-processing of variables, standardization of data sets, narrowing of measures to simplify calculations and elimination of adverse effects, here a normalized approach is used.

$$x_i(k) = \frac{x_i'(k) - x_{i\,mean}'}{x_{i\,max}' - x_{i\,min}'} \tag{3}$$

Calculate the absolute difference $\Delta_{0i}(k)$, the maximum value of the absolute difference Δ_{max} and the minimum value of the absolute difference Δ_{min}, and then calculate the correlation coefficient $\epsilon_i(k)$ between total food production and the influence factor.

$$\Delta_{0i}(k) = |x_0(k) - x_i(k)| (k = 1,\cdots,m\ i = 1,\cdots,n) \tag{4}$$

$$\Delta_{max} = \max_i \max_k \Delta_i(k) \tag{5}$$

$$\Delta_{min} = \min_i \min_k \Delta_i(k) \tag{6}$$

$$\epsilon_i(k) = \frac{\Delta_{min} + \rho\Delta_{max}}{\Delta_{0i}(k) + \rho\Delta_{max}} \tag{7}$$

where ρ is the discrimination coefficient, which takes the value in (0, 1), and if ρ is smaller, the greater the difference between the correlation coefficients and the stronger the discrimination ability. The introduction of this coefficient can greatly reduce the distortion caused by large values of the maximum absolute difference and increase the significance of the difference between correlation coefficients. Usually, ρ is 0.5.

The mean of the correlation coefficients of the reference series and the comparison series were calculated separately to reflect the correlation between grain yield and the influencing factors. The highest ranking is the main influencing factor, ranked in descending order.

$$r_{0i} = \frac{1}{m}\sum_{k=1}^{m} \epsilon_i(k)\ (i = 1,\cdots,n) \tag{8}$$

2.2 Support Vector Regression (SVR)

The first theoretical approach to support vector machines (SVM) was developed in 1963 by the ATE-T Bell Laboratory research group, led by Vanpik. Support Vector Regression (SVR) is the application of support vectors in the field of functional regression [13–14]. Unlike SVM, SVR seeks the optimal hyperplane not so that the two sample points are most apart, but so that the total deviation of all sample points from the hyperplane is minimized, and finding the optimal hyperplane is equivalent to finding the maximum interval. SVR creates a "spacing band" on both sides of the linear function with a spacing of ϵ \epsilonϵ (tolerance bias, an empirical value set by hand). Losses are not computed for all samples that fall into the interval band, i.e., only the support vector has an effect on its function model. Finally, the optimized model is derived by minimizing the total loss and maximizing the interval.

The application of support vector regression in the prediction of grain yield data is shown below:

Suppose the characteristic vector of influencing factors for the grain in year K is $X_m = (x_{m1}, \cdots, x_{mn})$ and the grain yield is Y_m. Given a training sample: $D = \{(X_1, Y_1), \cdots, (X_m, Y_m)\}, Y_i \in R$, so the support vector machine function is a nonlinear function.

$$f(x) = w\varphi(x) + b \tag{9}$$

In the above equation, $\varphi(x)$ is a non-linear mapping of the input space X to a higher dimensional space, the variables reflect the complexity of the function, and w and b are the parameters to be determined. Tolerant deviations are assumed to be ϵ, Thus, the SVR problem can be reduced to the following format:

$$\min_{w,b} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m} l_\epsilon(f(x_i), y_i) \tag{10}$$

where l_ϵ is the loss function and C is the penalty factor, introducing the relaxation variable ξ_i , $\hat{\xi}_i$, which optimizes to the equation.

$$\begin{cases} min\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}(\xi_i + \hat{\xi}_i) \\ s.t. f(x_i) - y_i \le \epsilon + \xi_i \\ y_i - f(x_i) \le \epsilon + \hat{\xi}_i \\ \xi_i \ge 0, \hat{\xi}_i \ge 0 \\ i = 1,2,\cdots,m \end{cases} \tag{11}$$

Introduce the Lagrange equation and then derive the partial derivatives for $w,\ b,\ \xi_i,\ \widehat{\xi}_i$, so that the partial derivative results in zero, which gives the above dual optimization form.

$$\begin{cases} max -\frac{1}{2}\sum_{i,j=1}^{n}(\alpha_i + \hat{\alpha}_i)(\alpha_j + \hat{\alpha}_j)K(x_i, x_j) \\ -\sum_{i=1}^{n}\alpha_i(\epsilon - y_i) - \sum_{i=1}^{n}\hat{\alpha}_i(\epsilon + y_i) \\ s.t. \sum_{i,j=1}^{n}(\alpha_i - \hat{\alpha}_i) = 0 \\ 0 \le \alpha_i, \hat{\alpha}_i \le \mathrm{C}(i = 1,2,\cdots,n) \end{cases} \tag{12}$$

Solve the above equation to obtain the final support vector machine regression function.

$$f(x) = \widehat{\omega}\varphi(x) + \hat{b} = \sum_{i=1}^{m}(\hat{\alpha}_i - \alpha_i)K(x_i, x_j) + b \tag{13}$$

where $K(x_i, x_j) = \phi(x_i) \cdot \phi(x_j)$ is the nuclear function, and the role of the nuclear function is to project the original input space into the high-dimensional feature space to solve the problem of linear indistinguishability of the original space. The common nuclear functions are linear, radial base, and Sigmoid kernels.

2.3 Adaptive Boosting Algorithm (AdaBoost)

Adaptive Boosting (AdaBoost) first gives a weight value to the initial sample, inputs the sample and the corresponding weights into the base learning method, and trains a base learner. The distribution of training sample weights is adjusted according to the base learner's performance, and the next base learner is trained based on the adjusted weights. The process is repeated until the number of base learners reaches a pre-specified value. All base learners are finally combined according to a binding strategy to obtain the final strong learner.

The process of applying the AdaBoost regression algorithm to the grain yield dataset is as follows:

The input grain yield data training set: $T = \{\{x_i, y_i\}\}_{i=1}^{m}$, the base learning algorithm is τ; the number of base learners is K, and the output final strong learner is $f(x)$.

Distribution of weights for initialized training sample samples.

$$\hat{T}(1) = (w_{11}, \cdots, w_{1m}); w_{1i} = \frac{1}{m}; i = 1,2,\cdots, \tag{14}$$

For iterative rounds there is $k = 1,2,\cdots,K$

Train the base learner $G_k = \tau(T, \hat{T})$ using a T_k with weights;

Calculate the maximum error of the sample on the training set.

$$E_k = \max|y_i - G_k(x_i)| \ \ i = 1,2,\cdots,m \tag{15}$$

To calculate the relative error for each sample, here the linear error was chosen.

$$e_{ki} = \frac{|y_i - G_k(x_i)|}{E_k} \tag{16}$$

Calculate the regression error rate of the base learner G_k on the training set, weighting coefficients, update the distribution of weights for the training set sample T_{k+1}.

$$\begin{cases} \varepsilon_t = \Sigma_{i=1}^{m} w_{k_i} e_{k_i} \\ \alpha_t = \frac{\varepsilon_t}{1-\varepsilon_t} \\ Z_k = \Sigma_{i=1}^{m} w_{k_i} \alpha_k^{1-e_{k_i}} \\ w_{k+1,i} = \frac{w_{k_i}}{Z_k} \alpha_k^{1-e_{k_i}} \end{cases} \tag{17}$$

A linear combination of base learners is constructed, and the binding strategy uses the median of the weights of the base learners is taken. The base learner is used as the basic method for the strong learner, and the final strong learner is as follows:

$$f(x) = \Sigma_{k=1}^{K} \left(\ln \frac{1}{\alpha_k}\right) g(x) \tag{18}$$

where $g(x)$ is the median of all $\alpha_k G_k(x), k = 1,2,\cdots,K$.

2.4 The Establishment of GRA-AdaBoost-SVR Model

Support vector regression (SVR) linear indistinguishability is made possible by transforming samples with linearly indistinguishable low-dimensional input space into a high-dimensional feature space using a nonlinear mapping algorithm [15], which makes it possible to analyze the nonlinear features of the samples linearly using a linear algorithm in a high-dimensional feature space. However, its performance depends heavily on the choice of kernel functions and parameters. The AdaBoost algorithm trains multiple base learners, and by combining strategies to get the final strong learner, the sample generalization is better. The AdaBoost algorithm is sensitive to anomalous samples, and anomalous samples may receive higher weights in iterations, which affects the prediction accuracy of strong learners. Moreover, the commonly used base learning algorithm is decision tree, but decision tree is very ineffective in dealing with nonlinear problems, and the prediction accuracy varies greatly. However, the AdaBoost algorithm is sensitive to anomalous samples, and anomalous samples may receive higher weights in iterations, which affects the prediction accuracy of strong learners [16]. Moreover, the commonly used base learning algorithm is decision tree, but decision tree is very ineffective in dealing with nonlinear problems, and the prediction accuracy varies greatly.

This paper therefore combines the three GRA-AdaBoost-SVR approaches to address the following issues:

Using SVR alone for sample learning, the model performance depends on the selected kernel function and kernel parameters. The use of SVR as the base learner of AdaBoost, however, reduces the impact of the choice of kernel functions and parameters in the SVR algorithm [17]. It also solves the poor solving ability of AdaBoost's traditional algorithm for nonlinear problems. Make the AdaBoost-SVR algorithm suitable for dealing with nonlinear feature data prediction and ensure the generalization ability of the model. Finally, the GRA method is introduced to solve the outlier problem of AdaBoost-SVR algorithm and ensure that the input features are the main relevant features, which improves the prediction accuracy and also ensures the smoothness and generalization ability of the model.

The GRA-AdaBoost-SVR model for grain yield prediction can be developed by following these steps:

Step 1: The original dataset was processed using GRA, using the grain production data as a reference series and the remaining influences as a comparison series to generate correlation coefficients according to Eqs. (1)–(8).

Step 2: Sort the correlation coefficients from largest to smallest, select the subsequences with correlation coefficients greater than n as input variables, and update the training prediction sample.

Step 3: In the updated data set selected m years as training samples, according to Eq. (14) to assign initial weights to each set of samples.

Step 4: Initialize the SVR algorithm, select the appropriate penalty factor C and kernel function, and use the SVR as a base learner for training according to Eqs. (9)–(13).

Step 5: Iteratively train K rounds according to Eqs. (14)–(17), and update the weights by calculating the corresponding parameters to get K base learners.

Step 6: After training K-rounds, build a strong learner according to the binding strategy, and obtain Eq. (18) as the final output function.

The flow chart of the algorithm is shown in Fig. 1.

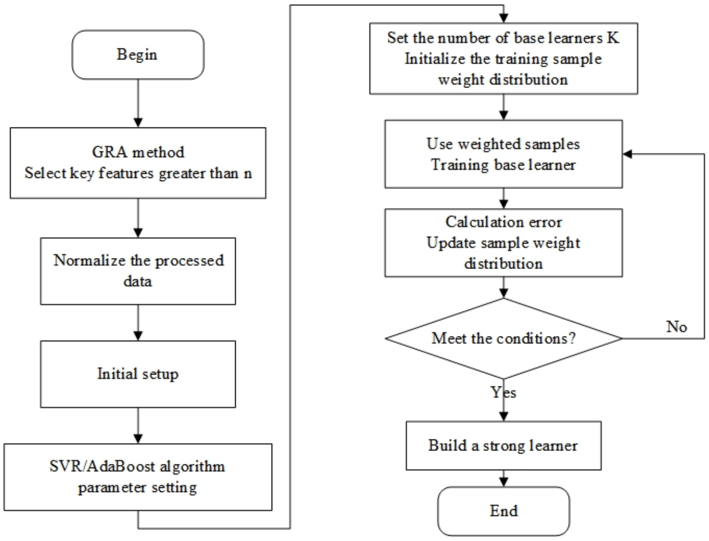
Begin
GRA method
Select key features greater than n
Normalize the processed data
Initial setup
SVR/AdaBoost algorithm parameter setting
Set the number of base learners K
Initialize the training sample weight distribution
Use weighted samples
Training base learner
Calculation error
Update sample weight distribution
Meet the conditions?
No
Yes
Build a strong learner
End

Figure 1: Flowchart of the GRA-AdaBoost-SVR algorithm model

3 Experiments and Analysis

3.1 Data Collection

The relevant data sources for this paper are the website [18] of the National Bureau of Statistics of China and the China Grain Development Report. The national grain dataset for 2000–2018 was selected by reviewing relevant references and related materials. The grain production data are shown in Fig. 2. From the dataset, we can see that the grain production data showed a steady trend in 2000–2003, the data in 2003–2013 showed an overall linear upward trend, and after 2014 China's grain production tends to be stable. Therefore, a simple linear or time series model cannot fit the grain production data well.

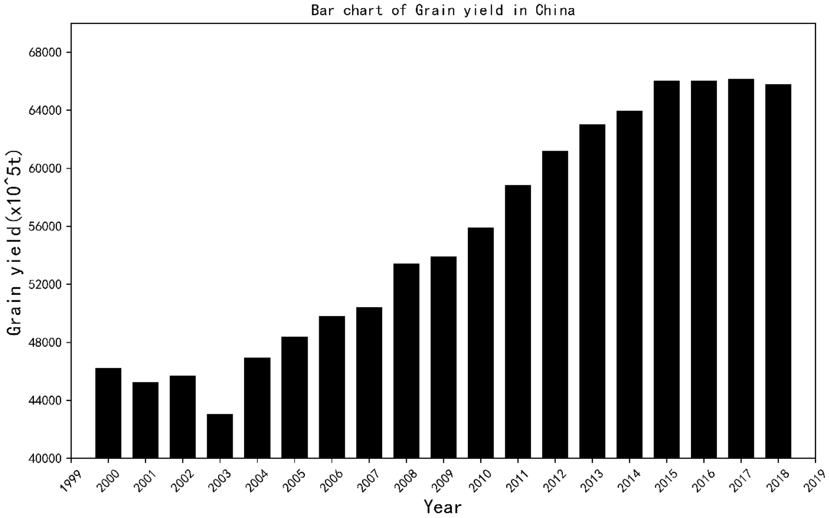

Figure 2: Grain yield in China, 2000–2018

By reading the relevant references and considering them all, the function inputs selected to influence the grain yield prediction model were 11 influencing factors: grain crop sown area x_1, effective irrigation area x_4, number of agricultural reservoirs x_7, and rural electricity consumption x_{11}. Output as a function of grain yield. See Tab. 1 for details.

Table 1: Factors influencing grain yields

Feature number	Feature name	Unit of measurement
x_1	Area planted with grain crops	$10^4\ m^2$
x_2	Amount of agricultural fertilizer applied	$10^8\ kg$
x_3	Total power of agricultural machinery	$10^8\ w$
x_4	Effective irrigated area	$10^4\ m^2$
x_5	Agricultural diesel use	$10^8\ kg$
x_6	Amount of pesticides used	$10^3\ kg$
x_7	Number of agricultural reservoirs	1
x_8	Agricultural plastic film use	$10^3\ kg$
x_9	Area of crops affected	$10^4\ m^2$
x_{10}	Price indices of agricultural inputs	Prior year = 100
x_{11}	Rural electricity consumption	$10^{11}\ w/h$
y	Grain yield	$10^8\ kg$

3.2 Application of the GRA-AdaBoost-SVR Model

The dataset of this study is based on the grain production data of China from 2000–2018. The 2000–2016 data were used as a training set and the 2017–2018 data year as a test set. In this paper, 11 characteristics from Tab. 1 were selected as reference series with grain yield as the reference series and the remaining 11 influences as comparison series. The grain dataset is processed according to step 1 of the GRA-AdaBoost-SVR model to generate the gray correlation coefficient plots, as shown in Fig. 3.

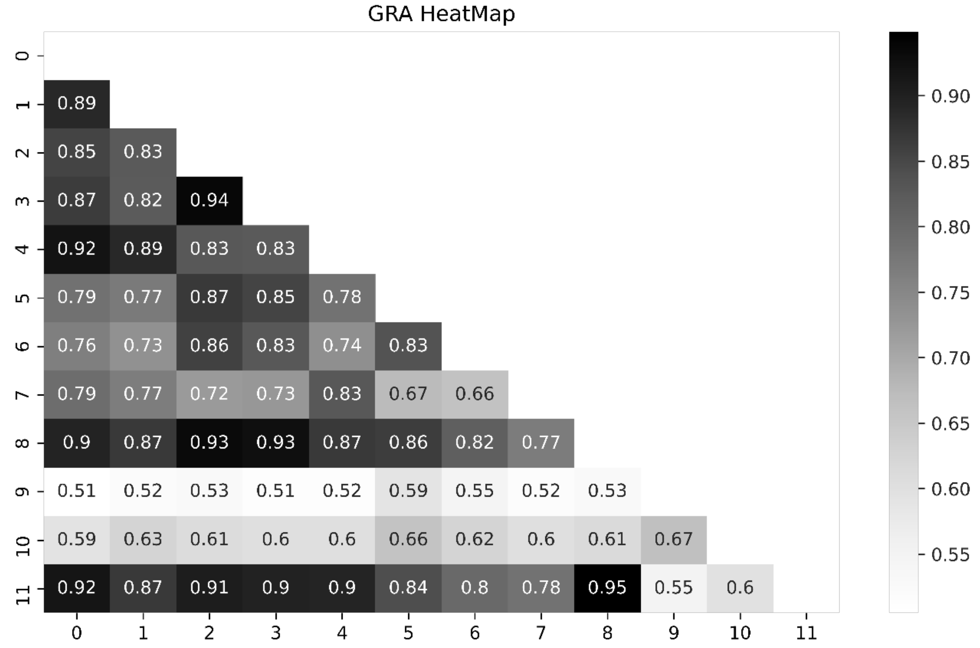

Figure 3: Grey correlation coefficient plots of grain yield characteristics

The 0th column is the grain production data, and the 1st through 11th columns are the characteristics that affect grain generation, corresponding to a total of 11 influences from $x_1 - x_{11}$ in Tab. 1. The corresponding correlations are then ranked according to Step 2, and the ranking results are shown in the following table.

Table 2: Ranking of associations between grain yield and influencing factors

Feature number	Feature name	Correlation coefficients	Sort
x_4	Effective irrigated area	0.920751	1
x_{11}	Rural electricity consumption	0.917346	2
x_8	Agricultural plastic film use	0.895166	3
x_1	Area planted with grain crops	0.889663	4
x_3	Total power of agricultural machinery	0.865302	5
x_2	Amount of agricultural fertilizer applied	0.854724	6
x_7	Number of agricultural reservoirs	0.792434	7
x_5	Agricultural diesel use	0.785625	8
x_6	Number of agricultural reservoirs	0.76215	9
x_{10}	Price indices of agricultural inputs	0.593	10
x_9	Area of crops affected	0.505111	11

From Tab. 2, according to Step 2, set n to 0.85 and select the influencing factors greater than 0.85 as input variables, there are six factors. They are x_1, x_2, x_3, x_4, x_8, x_{11}, and the corresponding independent variables are Area planted with grain crops, t Amount of agricultural fertilizer applied, Total power of agricultural machinery, Effective irrigated area, Agricultural plastic film use, Rural electricity consumption. These six items are used as inputs to the GRA-AdaBoost-SVR model, with grain yield as a function output.

Using the grey relational analysis (GRA) in the GRA-AdaBoost-SVR model, six highly correlated characteristic variables were screened and used as input variables for the model. The comparison algorithms were conventional: BPNN, SVR, AdaBoost-SVR. All conventional algorithms have 11 input feature variables. The input layer, intermediate layer, and output layer nodes of the BPNN algorithm are set to 11, 7, and 1, respectively [19]. The SVR algorithm kernel functions used in the experiments all used radial basis functions (RBF) [20]. There are two model parameters that need to be determined in the SVR algorithm: the penalty factor C and the kernel parameter σ^2. It has been shown that the selection of the model parameters has a great influence on the fit of the model [21]. In this paper, to avoid large differences in the function fitting effect. Set the parameter range C $\in$ [1,100] with a step size of 10; set C $\in$ [1,100] with a step size of a multiple of 10. The parameters were adjusted using the grid search CV method. The optimal search results for the parameters C=20 and σ=0.1 were obtained. The base learner training method in the AdaBoost algorithm uses the support vector regression algorithm, and the number of base learners all set to 50 [22–23].

The above parameter settings are used in the model used in this paper. The BPNN algorithm uses 11 feature variables as input, uses input training set year data as the result of algorithm training, uses gradient descent to update the learning rate, and least squares as the loss function to obtain the prediction function after 500 rounds of training. The SVR and AdaBoost-SVR algorithms also both use 11 feature variable inputs and use least squares as the loss function to obtain the prediction function. The GRA-AdaBoost-SVR model proposed in this paper, after building the model following the steps in Section 2.4, uses grey correlation analysis, which employs six feature variables with higher correlation, and ditto for inputting the training set year data to obtain the prediction function. After the prediction functions were obtained for all four methods, the prediction results were obtained by inputting the characteristic variables of the test set years. The results of comparing the four models trained and tested are shown in Fig. 4.

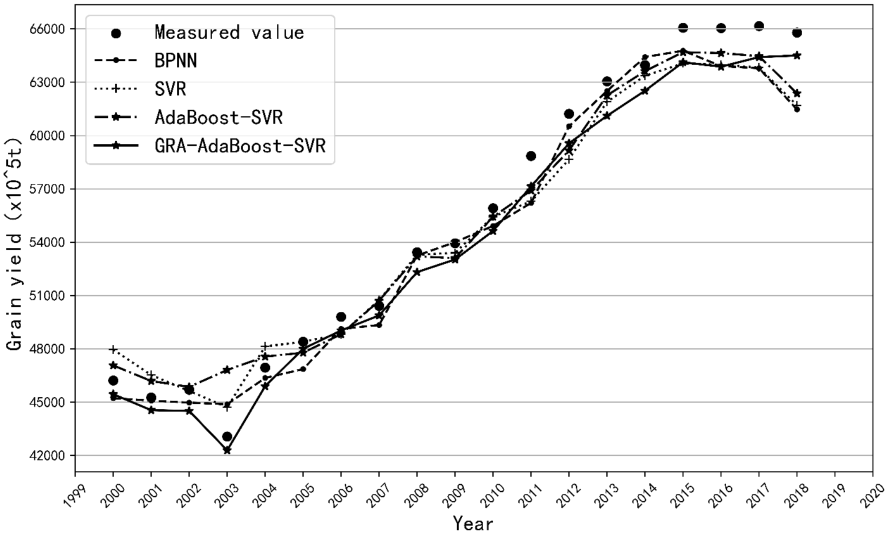

Figure 4: Comparison of the results of the 4 model training predictions

3.3 Results and Analysis

Fig. 4 presents the results of the different methods for predicting grain production. The GRA-AdaBoost-SVR algorithm in the figure is smoother and closer to the true data for each year. The overall fluctuations of the three model algorithms, BPNN, SVR, and AdaBoost, are greater. For example, in the 2003 projections, the projections made are more closely aligned with the real values because GRA-AdaBoost-SVR uses feature extraction, but the other algorithms all have large gaps from the real year. Between 2003 and 2013, there is an overall linear upward trend in food production, with little difference between the four models. However, simple models, such as BPNN and SVR algorithms can better fit the true value. During 2014–2018, the four models have a large difference in algorithms. The GRA-AdaBoost-SVR algorithm basically matches the grain production trend, but several other algorithms produce predictions that differ from the true value due to irrelevant factors.

In this paper, two indicators of Root Mean Square Error (RMSE) mean Relative Error were selected to assess the quality of the prediction results of the grain production dataset [24]. They were calculated as follows:

$$RMSE = \sqrt{\frac{1}{N}\Sigma_{i=1}^{N}(y^{(i)} - \hat{y}^{(i)})^2} \tag{19}$$

$$Relative\ Error = \frac{y^{(i)} - \hat{y}^{(i)}}{y^{(i)}} \tag{20}$$

where $y^{(i)}$, $\hat{y}^{(i)}$ are actual and predicted grain production, respectively, and N denotes the size of the data set. The smaller the RMSE, the smaller the relative error value, the better the prediction [25]. The results of the trained model on the test set are shown in Tab. 3. Tab. 4 gives the results of the model's performance on the entire dataset.

Tab. 3 and Tab. 4 show the three evaluation indicators of mean relative error, maximum relative error, and RMSE. It can be concluded that the GRA-AdaBoost-SVR algorithm significantly outperforms the other three algorithms. Although in some years the BP algorithm is more closely aligned with actual values. But in general, the GRA-AdaBoost-SVR algorithm predicts smoother and closer to the true value overall. The maximum relative error of the SVR, AdaBoost-SVR algorithm is more than twice as large as that of the GRA-AdaBoost-SVR algorithm. It shows that the algorithmic model does improve the prediction accuracy of the traditional algorithm.

Table 3: Comparison of the results of the four model test sets

Year	Measured grain yield (10^5 kg)	BPNN		SVR		AdaBoost-SVR		GRA-AdaBoost-SVR	
		Prediction	Relative error（%）	Prediction	Relative error (%)	Prediction	Relative error (%)	Prediction	Relative error (%)
2017	66160	63023	4.74	63824	3.53	64467	2.56	64410	2.65
2018	65789	60982	7.31	61703	6.21	62370	5.2	64500	1.96
Test set mean Relative Error		6.03		4.87		3.88		2.31	

Table 4: Effectiveness of the four models in predicting food production

Predictive model	Mean relative error	Max relative error	RMSE
BPNN	2.22	7.31	1677.07
SVR	2.37	6.21	1709.13
AdaBoost-SVR	2.18	8.69	1549.02
GRA-AdaBoost-SVR	2.17	3.29	1332.37

In summary, compared with traditional BPNN and SVR algorithms, AdaBoost-SVR has higher prediction accuracy and better generalization ability. Addressing the small sample size and multiple characteristics of grain yield datasets. The GRA method is used for correlation analysis of comparative sequences on the dataset to reduce input features and increase the weight of relevant features in prediction. The GRA-AdaBoost-SVR algorithm is proposed to further improve the stability of the algorithm and enhance the prediction accuracy with relatively small errors and better results.

4 Conclusion

In this paper, a GRA-AdaBoost-SVR model is proposed which can firstly extract the input features in grain yield prediction accurately using the GRA method, secondly map the input features to a high dimensional space by an SVR algorithm to solve the nonlinear features of the grain yield dataset, and finally combine the AdaBoost algorithm to improve the generalization capability of the SVR algorithm to solve the SVR algorithm The problem of over-reliance on the selection of parameters in the model. It makes the model improve the accuracy of prediction while safeguarding the generalization ability of the model. The GRA-AdaBoost-SVR method was finally utilized to develop the grain yield prediction model. The GRA method is used to extract key features affecting grain yield, select more highly correlated feature factors as inputs, and accurately analyze the main factors while reducing the cost required for grain yield prediction. The base learning algorithm used in the model is SVR, and the final strong learner is obtained after iterative training to obtain the grain yield prediction results.

The results show that the GRA-AdaBoost-SVR combination method has higher accuracy and reliability in predicting grain yield in China, and the maximum relative error is 50% of that of the traditional model, which makes the model more stable and can be effectively applied to grain production prediction research. It provides a new way for grain production prediction at home and abroad. At present, due to the small data volume samples used in the algorithm, it has less impact on the training speed, but the training time will gradually increase when the data volume samples increase, so in future work, we can consider enhancing the learning time of the GRA-AdaBoost-SVR model algorithm, by streamlining the algorithm model steps and optimizing the parameters of the algorithm to achieve the optimization of the model, while improving the training time of the algorithm, so that the model predicts with greater accuracy and generalization capability.

Funding Statement: This work was support in part by Research on Key Technologies of Intelligent Decision-Making for Food Big Data under Grant 2018A01038, in part by the National Science Fund for

Youth of Hubei Province of China under Grant 2018CFB408, in part by the Natural Science Foundation of Hubei Province of China under Grant 2015CFA061, in part by the National Nature Science Foundation of China under Grant 61272278, and in part by the Major Technical Innovation Projects of Hubei Province under Grant 2018ABA099.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] G. Q. Cheng and M. D. Zhu, "The impact of the new coronary pneumonia epidemic on food security: trends, impacts and responses", *China Rural Economy*, vol. 5, pp. 13–20, 2020.

[2] X. Y. Gao and F. Han, "Grain yield prediction by support vector machine based on hybrid intelligent algorithm", *Journal of Jiangsu University (Nature Science)*, vol. 41, no. 3, pp. 301–306, 2020.

[3] G. Velu, J. Crossa, R. P, Singh, Y. F. Hao, S. Dreisigacker *et al.,* "Genomic prediction for grain zinc and iron concentrations in spring wheat," *Theoretical and Applied Genetics,* vol. 129, no. 8, pp. 1595–1605, 2016.

[4] A. Haghighattalab, J. Crain, S. Mondal, J. Rutkoski J, R. P. Singh *et al.,* "Application of geographically weighted regression to improve grain yield prediction from unmanned aerial system imagery," *Crop Science*, vol. 57, no. 5, pp. 2478, 2017.

[5] C. S. Zong, H. W. Zheng and L. S. Wang, "Improved particle swarm optimized BP neural network grain yield prediction model," *Computer Systems Applications,* vol. 27, no. 12, pp. 204–209, 2018.

[6] X. Q. Tian, "Grain yield prediction based on multiple linear regression," *Science and Technology Innovation and Application*, vol. 16, pp. 3–4, 2017.

[7] L. Q. Rong, F. Chen and H. Ouyang, "Research on grain yield prediction in Guangxi based on GRA & BPNN," *China Agricultural Resources and Zoning,* vol. 38, no. 2, pp. 105–111, 2017.

[8] C. S. Xiang and L. F. Zhang, "Grey theory and Markov fusion model for grain yield prediction," *Computer Science,* vol. 40, no. 2, pp. 245–248, 2013.

[9] S. Q. Zhao and L. S. Shao, "A coal demand prediction model based on MVO-SVR-AdaBoost for China," *Journal of Liaoning University of Engineering and Technology (Natural Science),* vol. 39, no. 4, pp. 366–374, 2020.

[10] R. Z. Tian, "Housing price prediction based on multiple machine learning algorithms," *China New Communications,* vol. 21, no. 11, pp. 228–230, 2019.

[11] S. M. Zai, J. Wen and W. F. Wu, "Grain yield prediction model based on grey correlation analysis in Liaoning province," *Water Conservation and Irrigation,* vol. 5, pp. 64–66, 2011.

[12] W. Y. Zhao, R. L. Fu, J. Q. He and R. M. Li, "Comprehensive evaluation of the development level of agricultural modernization in various provinces in my country," *Chinese Journal of Agricultural Machinery Chemistry,* vol. 39, no.12, pp. 94–100, 2018.

[13] S. J. Yang and Y. B. Li, "Grey correlation analysis of agricultural mechanization and grain yield in Jilin province," *Chinese Journal of Agricultural Machinery Chemistry,* vol. 39, no. 8, pp. 101–107, 2018.

[14] D. Zhao, "Research and application of machine learning method based on group intelligence optimization," Ph.D. dissertation, Jilin University, Jilin, 2017.

[15] Y. Li, C. T. Hou, J. G. Wu, B. R. Liu and Z. Q. Li, "An overview of the machine learning potential," Beijing Mechanics Society, 2020.

[16] Z. D. Zhou, D. W. Li and G. S. Li, "Application of AdaBoost-SVR model based on stepwise regression in cost prediction of offshore wind power projects", *Solar Energy Journal,* vol. 41, no. 7, pp. 259–264, 2020.

[17] W. Fang, L. Pang and W. N. Yi, "Survey on the application of deep reinforcement learning in image processing," *Journal on Artificial Intelligence*, vol. 2, no. 1, pp. 39–58, 2020.

[18] National Bureau of Statistics of the People's Republic of China. [Online]. Available: http://www.stats.gov.cn/.

[19] Y. Y. He, N. Gao, F. H. Wang, S. F. Ru and J. B. Han, "Stock price integration prediction based on SVR under EMD decomposition," *Journal of Northwestern University (Natural Science Edition)*, vol. 49, no. 3, pp. 329–336, 2019.

[20] X. Zhuang and F. Han, "Optimization of BP neural network for grain yield prediction based on mixed group intelligence algorithm," *Journal of Jiangsu University (Natural Science Edition)*, vol. 40, no. 2, pp. 209–215, 2019.

[21] Y. P. Yuan, "Research on the prediction of the development level of agricultural machinery equipment in Heilongjiang reclamation area and its contribution to grain production," Ph.D. thesis, Heilongjiang Bayi Agricultural Reclamation University, Heilongjiang, 2014.

[22] C. Oswald, "Artificial intelligence, machine learning, and deep learning," *Mercury Learning & Information*, pp. 2–13, 2020.

[23] V. Pagani, T. Guarneri, D. Fumagalli, E. Movedi, L. Testi *et al.,* "Improving cereal yield forecasts in Europe–the impact of weather extremes," *European Journal of Agronomy*, vol. 89, pp. 2–13, 2017.

[24] A. Ceglar, A. Toreti, R. Lecerf, M. V. D. Velde and F. Dentener, "Impact of meteorological drivers on regional inter-annual crop yield variability in France," *Agricultural and Forest Meteorology*, vol. 216, pp. 58–67, 2016.

[25] W. Fang, F. Zhang, Y. Ding and J. Sheng, "A new sequential image prediction method based on LSTM and DCGAN," *Computers, Materials & Continua*, vol. 64, no. 1, pp. 217–231, 2020.