Study on Feature Engineering and Ensemble Learning for Student Academic Performance Prediction

Du Xiaoming[1]

School of Economics and Management
Jiangsu University of Science and Technology
Zhenjiang, China

Chen Ying[2], Zhang Xiaofang[3], Guo Yu[4]

Zhangjiagang Campus
Jiangsu University of Science and Technology
Suzhou, China

Abstract—**Student academic performance prediction is one of the important works in the teaching management, which can realize accurate management, scientific teaching and personalized learning by mining important features affecting the academic performance and accurately predicting academic. Due to the subjectivity of feature extraction and the randomness of hyperparameters, the accuracy of academic performance prediction needs to be improved. Therefore, in order to improve the accuracy of prediction, an academic prediction method based on Feature Engineering and ensemble learning is proposed, which makes full use of the advantages of random forest in feature extraction and the ability of XGBoost in prediction. Firstly, the feature importance is calculated and ranked by using the random forest method, and the optimal feature subset combined with the forward search strategy. Secondly, the optimal feature subset is input into the XGBoost model for prediction. The sparrow search algorithm is used to optimize the XGBoost hyperparameters to further improve the accuracy of academic prediction. Finally, the performance of the proposed method is verified through the experiments of the public data set. The experimental results show that the academic prediction method designed is better than the single learner prediction method and other integrated learning prediction methods. The accuracy result jumps to 82.4%. It has good prediction performance and can provide support for teachers to teach according to students' aptitude.**

Keywords—Academic performance prediction; feature engineering; ensemble learning; random forest; XGBoost

I. INTRODUCTION

With the application and popularization of emerging technologies such as internet of things, big data and artificial intelligence, the intellectualization of education has been developing rapidly, which promoting the reform of education and triggering the change of educational paradigm[1]. At present, many colleges have carried out intelligent campus construction and collected a large number of data generated by educational activities [2]. By machine learning technology, how to mine the mode and value contained in the data has become one of the urgent problems to be solved. Academic performance prediction refers to the use of relevant theories such as pedagogy, computer science and statistics, which analyzing the data generated in the process of students' learning and predicting their future academic performance [3]. Through academic prediction, managers can effectively manage the school, carry out academic early warning for students with academic risk in advance, establish a dynamic early warning

mechanism, and accurately guide students out of difficulties. Teachers can realize scientific teaching; predict students' academic status and teaching effect in advance, so as to purposefully optimize teaching activities, formulating differentiated teaching plans, meeting the needs of students at different levels, and truly teaching students according to their aptitude. Students can realize learning personalization; find out in advance which behaviors are beneficial to learning and which behaviors affect learning effect in advance. In this way, it could consolidate beneficial behaviors and break bad habits. Therefore, whether from the perspective of efficient school management, scientific teaching of teachers or personalized learning of students, to investigate the characteristics of affecting students' performance and create a high-accuracy prediction model has become an important research direction under the background of educational intelligence.

Currently, the research on the prediction of students' academic performance has some problems such as: subjectivity of feature extraction and poor prediction accuracy. This paper proposes an academic prediction model based on feature engineering and ensemble learning. Firstly, the out of bag estimation method of random forest is used to calculate and rank the feature importance, and the forward sequence method is used to search the optimal feature subset, which could deal with the randomness of the selection of random forest feature subset. Then the sparrow search optimization algorithm is used to adjust the hyperparameters of XGBoost to obtain the optimal combination of hyperparameters. Finally, the optimal feature subset is input into the optimized XGBoost model for academic prediction. The two kinds of ensemble learning methods are combined orderly, making full use of the advantages of random forest in feature extraction and the ability of XGBoost in prediction, which could enhance the generalization and effectiveness of academic prediction methods.

II. RELATED WORK

With the extensive use of data acquisition equipment, the collection of student data has expanded from single-mode learning data to multi-mode data，which including learning behavior, life behavior and psychological behavior [4]. The data show exponential development in volume and high dimensionality in characteristics. The essence of prediction is to find the mapping relationship between features and targets. Because the original feature set contains associated features and redundant features, it is often necessary to extract features in order to achieve better prediction effect. Too little feature

extraction will lead to "under fitting", which will affect the prediction accuracy. Too much feature extraction will lead to "over fitting", which will not only increase the calculation difficulty, but also reduce the prediction accuracy. In order to extract the optimal feature subset, some researchers manually select the feature subset through domain knowledge or expert experience. For example, Hu and others divide the features into static features and dynamic features, which takes students' basic information features as static features and students' behavior features (early rising behavior, borrowing behavior, etc.) as dynamic features, and predict learning performance according to the selected features [5]. Fan divides the characteristics into three types: tendency characteristics, human-computer interaction characteristics and interpersonal interaction characteristics. It is pointed out that the prediction ability of propensity characteristics is strong in the early stage of learning. With the progress of learning, the prediction ability of human-computer interaction characteristics and interpersonal interaction characteristics have gradually enhanced [6]. Li constructed a student behavior analysis model including five dimensions: Students' basic information, classroom learning, extracurricular learning, campus life and entertainment [7]. Some researchers use filtering or packaging based feature engineering methods to automatically select feature subsets. For example, based on filtered correlation analysis and information gain method, Chen and others calculate the Pearson correlation coefficient between features and scores, sorting them in descending order according to the results, and select the first 9 features from the 16 features as the main influence features [8]. However, it does not make the optimal selection of features. Through correlation analysis, information gain ratio and chi square analysis, Febro has selected 14 features from 29 original features to form the optimal subset, and verified that the prediction result of the feature subset is better than that of the whole feature [9]. Cao extracted the characteristics of students' regularity and preciseness from the campus life data. It is found that students' regularity is positively correlated with their grades, and preciseness is significantly correlated with their grades [10].Wen proposed a hybrid feature selection method based on packaging. This method first generates candidate feature sets through scoring and sorting, and then uses heuristic methods to generate the final results [11]. However, the computational complexity of this method is exponential and needs longer running time. The above research screened the features, but did not consider the redundancy between features, and did not check the dimension of the optimal features.

In terms of prediction methods, machine learning methods have gradually replaced probability and statistics methods and gradually become the main research methods, including linear regression, logical regression, decision tree, support vector machine, neural network, integration method and deep learning [12-14]. Wu compared four different performance prediction methods: decision tree, Bayesian network, neural network and support vector machine, and found that the performance prediction model of Bayesian network has high accuracy and recall [15]. Liu used support vector machine to predict students' grades [16]. Wang used correlation analysis and regression analysis to study the predictive effect of big five personality traits and individual intelligence on academic achievement [17]. Considering the influence of the spatial and temporal characteristics of students' behavior data, Du proposed a serial hybrid deep learning algorithm of CNN and LSTM to predict learners' performance [18]. Cao proposed LSTM depth neural network method to predict learning achievement [19].Ding uses the methods of random forest, SVM, KNN, decision tree and naive Bayes to predict students' academic performance. The results show that the prediction performance of random forest algorithm is the best [20]. Yao proposed a multi task learning achievement prediction framework based on learning ranking algorithm. [21]. The above studies mostly use the single classifier method for prediction. It is found that the integrated learner has better performance and higher accuracy than the single learner. Ensemble learning methods include two types: boosting and bagging. Boosting methods include AdaBoost, GDBT and XGBoost etc. Some researchers have applied ensemble learning to many fields and achieved good results. Hao used XGBoost model to predict whether learners can complete the course and obtain certificates [22]. Xu used XGBoost model to automatically identify students' classroom behavior [23]. Cao uses XGBoost to predict the online short rent market price. The experimental results show that XGBoost is better than the integrated learning method of LightGBM and AdaBoost [24]. When using ensemble learning methods, the above researchers often use default hyper parameters or set hyper parameters based on experience. Because there are many kinds of hyperparameters, these methods often cannot obtain the optimal hyperparameters combination, which will affect the prediction accuracy.

III. METHODOLOGY

A. The Framework of Academic Prediction

This paper designs the academic prediction framework, as shown in Fig. 1. Academic prediction mainly includes four processes. First, academic data preprocessing. It consists of three parts: clean up, convert and normalize the data. Second, feature extraction. The random forest model is used to rank the importance of data features, and the optimal feature subset is extracted according to the forward search strategy. Third, model training. Train XGBoost model based on sparrow search optimization algorithm. Fourth, performance evaluation. The performance of the model is evaluated according to the evaluation metrics.

B. Feature Engineering

Feature engineering is an important link in the process of machine learning prediction. It can effectively remove the associated features and redundant features, and use the appropriate search strategy to extract the optimal feature combination, which is helpful to reduce the complexity and improve the accuracy of the prediction method. Random forest (RF) is an integrated learning method based on decision tree. Its embedded feature importance evaluation mechanism has the function of analyzing the correlation between features. RF has the advantages of simplicity and good robustness in feature extraction. RF belongs to bagging method. Samples are randomly selected from the original data for basic learner training. The unselected data is called out of bag (OOB), which can be used as a test set. The error predicted according to OOB

is called generalization error. For a feature, the generalization error is calculated after its eigenvalues are randomly disrupted. If the difference between the two generalization errors is small, it means that the feature is not important, otherwise it means that the feature is important. When calculating the feature importance of random forest, the out of bag data is used to calculate the generalization error before and after the disturbance of feature data, and the difference between the two generalization errors is calculated. The calculation steps of RF characteristic importance are as follows:

1) $D = \{(X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n)\}$ is the original data set $X_i \in R^T, Y_i \in R$, T represents the number of features and n represents the number of samples. m (m < n) samples are randomly selected from D in k times to generate k training sets, k OOB sets and k decision trees.

2) Calculate the generalization error e_t of OOB dataset corresponding to the t-th decision tree.

3) Keep other eigenvalues of OOB unchanged, randomly disrupt the order of eigenvalues of the i-th feature, and recalculate the generalization error e_t^i.

4) Repeat steps 2) - 3) to traverse the whole forest and calculate the importance of the i-th feature. As shown in formula (1):

$$\varepsilon^i = \frac{1}{k} \sum_{t=1}^{k} e_t - e_t^i \tag{1}$$

Through the above steps, the feature importance is calculated and sorted in descending order according to the importance. In order to extract the optimal feature subset, a forward search strategy is adopted. In the first round, the feature subset containing one feature is selected from the ordered feature set $F = \{f_1, f_2, ..., f_t\}$. Obviously, the first feature f_1 is the selected feature subset of a single feature. In the i round, the first i features are formed into feature subsets, and their operation effects are compared with those of the first i-1 feature subsets. If the prediction accuracy is not as good as that of the first i-1 feature subset, the operation will be stopped, and the first i-1 feature subset is the best feature subset.

C. XGBoost Algorithm

XGBoost is an improved gradient boosting decision tree algorithm, which takes CART as the base learner and combines many base learners into high-performance integrated learners. For dataset $D = \{(\mathbf{x}_i, y_i)\}(i = 1, 2, ...n)$, where $\mathbf{x}_i$ represents the eigenvalue vector of sample i and y_i represents the label of sample i. Assuming that the XGBoost model contains T base learners, the predicted value of sample i is:

$$y_i = \sum_{t=1}^{T} f_t(\mathbf{x}_i), f_t \in F \tag{2}$$

F is all base learner spaces and $f(\mathbf{x})$ is the base learner function.

GBDT approximates the objective function by first-order Taylor expansion, while XGBoost approximates the objective function by second-order Taylor expansion to accelerate the convergence and improve the accuracy of the algorithm. In addition, in order to control the structural complexity of the model, XGBoost adds a regular term to the objective function to prevent the algorithm from over fitting and improve the generalization performance. The XGBoost objective function is:

$$obj(\theta) = L(\theta) + \Omega(\theta) \tag{3}$$

$L(\theta) = \sum_{i=1}^{n}(y_i - y_i)$ is the error function.

$\Omega(\theta) = \sum_{t=1}^{T} \Omega(f_t)$ is a regular item.

In the process of XGBoost training, the next base learner is trained according to the residuals of the previous trained model to minimize the objective function. After much iteration, an integrated model with high accuracy is generated. The objective function at iteration t is:

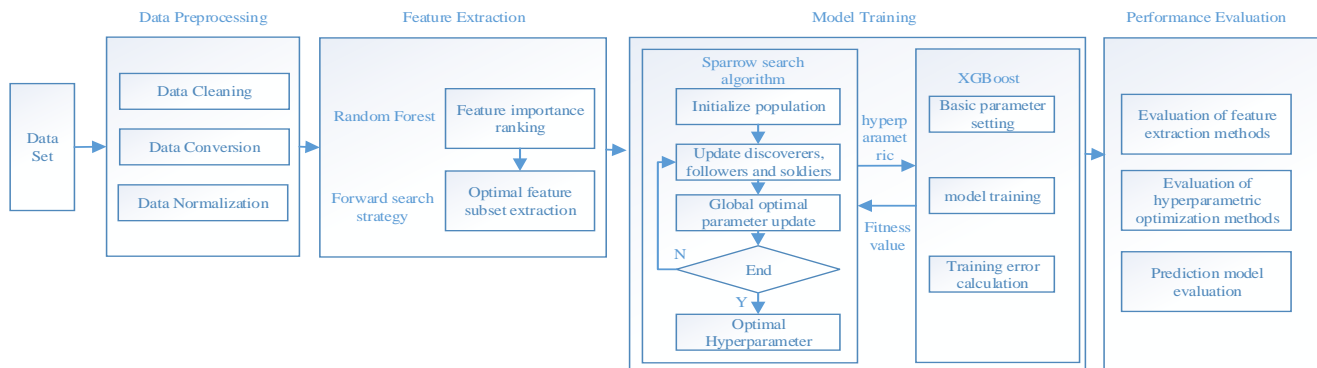$$obj^t = \sum_{i=1}^{n}((y_i - y_i^{t-1}) - f_t(\mathbf{x}_i)) + \Omega(f_t) + C \tag{4}$$

Fig. 1. Academic Prediction Framework.

f_t is a new basic learner function. C is constant. Carry out the second-order Taylor expansion of formula (4):

$$obj^t \approx \sum_{i=1}^{n} [L(y_i - y_i^{t-1}) + g_i f_t(\mathbf{x}_i) + \frac{1}{2} h_i f_t^2(x_i)] + \Omega(f_t) + C \tag{5}$$

$g_i = \partial_{y^{t-1}} L(y_i, y^{t-1})$ is the first derivative.

$h_i = \partial_{y^{t-1}}^2 L(y_i, y^{t-1})$ is the second derivative. The complexity of XGBoost consists of the number of leaves and the structure of the tree, so the regularization term can be defined as:

$$\Omega(f_t) = \gamma T + \frac{1}{2} \lambda \sum_{j=1}^{T} w_{j(x)}^2 \tag{6}$$

γ and λ are constant. T is the number of leaf nodes. $w_{j(x)}$ is the real value of the leaf node to the sample. Therefore, formula (5) can be changed to:

$$obj^t \approx \sum_{i=1}^{n} [L(y_i - y_i^{t-1}) + g_i f_t(\mathbf{x}_i) + \frac{1}{2} h_i f_t^2(x_i)] + \gamma T + \frac{1}{2} \lambda \sum_{j=1}^{T} w_{j(x)}^2 \tag{7}$$

The constant C does not affect the maximization of the objective function, so it can be omitted.

Define the sample set contained in the leaf node j:

$$I_j = \{i | q(x_i) = j\} \tag{8}$$

According to formula (8), formula (7) is expressed as:

$$obj^t = \sum_{j=1}^{T} [(\sum_{i \in I_j} g_i) w_j + \frac{1}{2}(\sum_{i \in I_j} h_i + \lambda) w_j^2] + \gamma T \tag{9}$$

$$G_j = (\sum_{i \in I_j} g_i), H_j = (\sum_{i \in I_j} h_i)$$

$$obj^t = \sum_{j=1}^{T} [G_j w_j + \frac{1}{2}(H_j + \lambda) w_j^2] + \gamma T \tag{10}$$

The minimization of the objective function is transformed into the minimization of the quadratic function of w_j, and the optimal w_j^* is solved:

$$w_j^* = -\frac{G_j}{H_j + \lambda} \tag{11}$$

$$obj^t = \frac{1}{2} \sum_{j=1}^{T} \frac{G_j}{H_j + \lambda} + \gamma T \tag{12}$$

D. Sparrow Search Algorithm for Optimizing Hyperparameters

Although XGBoost improves GBDT, optimizes the convergence speed and improves the accuracy, the determination of XGBoost hyperparameters is still the key problem to improve its performance. The prediction effect is often not high when the hyperparameters are set according to experience. Therefore, it is necessary to set the hyperparameters through the optimization algorithm. Sparrow search algorithm (SSA) is a bionic intelligent optimization algorithm that simulates the foraging behavior and anti-predation behavior of sparrows [25]. Compared with other intelligent optimization algorithms, it has better global search ability, less iterations and high prediction accuracy. In the process of sparrow foraging, it is divided into discoverer, follower and police soldier. The discoverer updates his position according to the foraging rules and guides the population to forage. The follower obtains the food around the discoverer or competes for the food of other individuals and updates the position. When the sparrow group realizes the danger, it will carry out anti predation behavior and update the corresponding position. The discoverer update location rule is:

$$X_{i,j}^{t+1} = \begin{cases} X_{i,j}^t \exp(-\frac{i}{\alpha \, iter_{max}}), & R_2 < ST \\ X_{i,j}^t + QL, & R_2 \geq ST \end{cases} \tag{13}$$

$X_{i,j}^t$ represents the position of the i-th sparrow in the j-th dimension at the t-th iteration. $\alpha \in (0, 1]$ is a random number. $iter_{max}$ is the total number of iterations. $R_2 \in [0, 1]$ is the warning value. $ST \in [0.5, 1]$ is the safety value. Q is a random number with normal distribution. L is the matrix of $1 \times d$, whose values are all 1. $R_2 < ST$ indicates that the sparrow is in a safe state and can expand the foraging range. $R_2 \geq ST$ indicates that when a predator is found, all sparrows should fly away quickly.

The rule for followers to update the location is:

$$X_{i,j}^{t+1} = \begin{cases} Q \exp(-\frac{X_{worst}^t - X_{i,j}^t}{i^2}), & i > n/2 \\ X_p^{t+1} + |X_{i,j}^t - X_p^{t+1}| A^+ \cdot L, & i \leq n/2 \end{cases} \tag{14}$$

n represents the number of all sparrows. X_p and X_{worst} represent the position of the current discoverer and the worst position of all sparrows, respectively, L is the matrix of $1 \times d$, whose values are all 1. A^+ represents the pseudo inverse matrix. When $i > n/2$, it means that followers with low fitness value are difficult to capture food and need to fly to other places for feeding.

When aware of the danger, the sparrow updates the location rule as follows:

$$
X_{i,j}^{t+1} = \begin{cases} X_{best}^t + \beta \left| X_{i,j}^t - X_{best}^t \right|, & f_i > f_g \\ X_{i,j}^t + k \left[\dfrac{X_{i,j}^t - X_{worst}^t}{(f_i - f_w) + \varepsilon} \right], & f_i = f_g \end{cases} \tag{15}
$$

X_{best}^t is the best position among all sparrows at present. f_i f_g and f_w represent the current sparrow fitness value, the worst fitness value among all sparrows and the best fitness value among all sparrows. β is the random value obeying (0,1) normal distribution. $k \in [-1, 1]$ is a random number. When $f_i = f_g$, it indicates that the sparrow is on the edge of danger and needs to be close to other sparrows.

IV. EXPERIMENTAL OF PREDICTION MODELS

The experimental environment of this paper is 64 bit Windows 7 operating system, the CPU is i5-3317u, the RAM is 4GB, the programming language is python, and the compilation environment is PyCharm.

A. Dominant Set and Data Preprocessing

This paper uses the score data set collected by the learning management system (LMS) of the University of Jordan. The data set contains 480 student records of 12 courses in 2 semesters, and each record includes a total of 16 features. These 16 characteristics are divided into four categories: demographic characteristics, knowledge background characteristics, parental behavior characteristics and learning behavior characteristics, as shown in Table I. It is divided according to 16 eigenvalue categories, including 5 binary features, 7 nominal features and 4 numerical features. The grades of each student's course are divided into three grades: L (0 to 69), M (70 to 89)) and H (90 to 100). There are 127 students with L, 211 students with M and 142 students with H.

Data preprocessing is a very important step in machine learning. Standardized data processing can eliminate the impact of different data dimensions on prediction accuracy, and help to improve prediction performance while maintaining data distribution. It usually includes data cleaning (missing value processing), data conversion and data normalization. There is no missing value in the data set. Five binary feature data are transformed into {0,1}, and seven nominal features are mapped into quantitative feature values. Finally, all data are normalized by min-max method.

TABLE I.   CLASSIFICATION OF DATA CHARACTERISTICS

Attribute type	Feature
Demographic Characteristics	Gender,Nationality, Place of birth,Parent responsible for student
Knowledge Background Characteristics	Educational Stages, Grade Levels, Section ID,Topic,Semester
Parental Behavior Characteristics	Parent Answering Survey,Parent School Satisfaction
Learning Behavior Characteristics	Raised hand, Visited resources, Viewing announcements,Discussion groups,Student Absence Days

B. Model Evaluation Metrics

The prediction results of this experiment are divided into three levels, which belong to multi classification problem. Accuracy and kappa coefficient are used as evaluation metrics. The value range of the two metrics is [0,1]. The larger the metric value, the better the prediction performance. The calculation formulas are:

$$
Accuracy = \frac{d'}{z} \tag{16}
$$

$$
Kappa = \frac{Accruacy - P_e}{1 - p_e} \tag{17}
$$

$$
Pe = \frac{d_1 \times d_1' + d_2 \times d_2' + d_3 \times d_3'}{z \times z} \tag{18}
$$

z is the total number of samples. d' is the number of samples with correct predictions. d_i' is the number of samples with correct prediction of type i. d_i is the total number of class i samples.

C. Feature Extraction

Based on the preprocessed data, the relative importance of features is calculated from equation (1) by using the random forest method. The results are shown in Fig. 2. Among the 16 features, the importance gap of each feature is obvious, and the five most important features belong to learning behavior features. It shows that students' learning behavior in class has a great impact on course performance. Too many or too few features will affect the prediction accuracy. In order to obtain the optimal feature subset, remove the unimportant features in turn according to the feature importance in Fig. 2, and calculate the Kappa index values of different feature subsets, as shown in Fig. 3. When the number of feature sets increases from 1 to 12, the Kappa index shows an increasing trend as a whole, reaches the maximum when the number of feature sets is 12, and shows a downward trend when the number of feature sets increases from 12 to 16. The main reason is that when the number of features is relatively small, the model training is insufficient, which affects the prediction accuracy. When the number of features is too large, the complexity of the model increases, resulting in over fitting training, which will also

reduce the prediction accuracy. Therefore, this paper selects the top 12 features of feature importance ranking for subsequent model prediction.

D. Comparison of Hyperparameters Optimization Methods

XGBoost parameters are divided into general parameters, lifter parameters and task parameters. General parameter setting is the overall function of the model, lifter parameter setting is the basic learner function, and task parameter setting is the optimization step. The parameters to be adjusted in this experiment and their adjustment range are shown in Table II.

In order to verify the efficiency of sparrow search algorithm in XGBoost hyperparametric optimization, comparative experiments are carried out by using manual experience method, grid search algorithm, random search algorithm and hyperparametric optimization method based on sparrow search algorithm. The experimental results are shown in Table III.
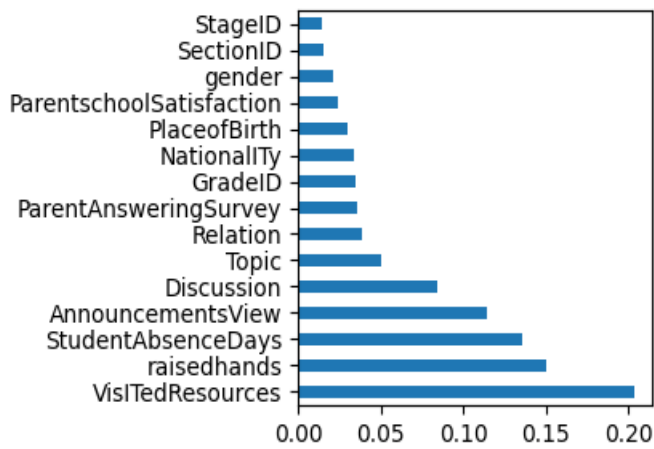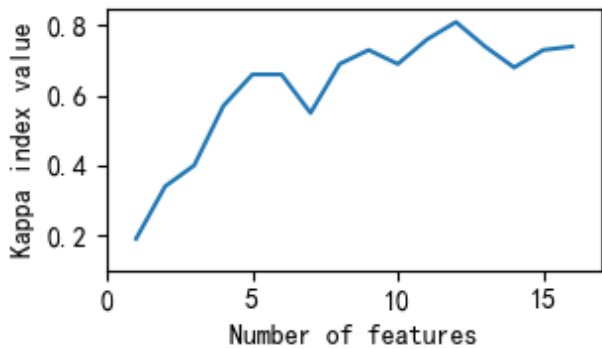

Fig. 2. Feature Importance.

Fig. 3. Kappa Value.

TABLE II. HYPERPARAMETERS AND RANGES

hyperparameters	ranges	content
n_estimator	（100,1500）	Number of decision trees
learning_rate	（0.01,0.15）	Learning rate
max_depth	（1,10）	Maximum depth of decision tree
min_child_weight	（1,10）	Minimum weight of leaf node

TABLE III. KAPPA VALUES OF DIFFERENT METHODS

	manual experience method	grid search algorithm	random search algorithm	sparrow search algorithm
Kappa values	0.7755	0.8083	0.7924	0.8242

It can be found from Table III that the Kappa index value of the hyperparametric optimization method based on sparrow search algorithm is the highest, followed by the grid search algorithm, and the worst is the manual experience method. The results of sparrow search algorithm are 4.87%, 1.59% and 3.18% higher than manual empirical method, grid search algorithm and random search algorithm respectively in Kappa index. It is verified that different combinations of super parameters have a great impact on prediction performance. Theoretically, when the number and value range of super parameters to be optimized are large, there will be many combinations of super parameters. The grid search algorithm needs to exhaust the whole parameter combination space, and the time complexity is very high; Random search algorithm makes random sampling in a given space, which has fast search speed, but it is easy to miss some better parameter combinations; The sparrow search algorithm gradually obtains the optimal solution after iteration and updating the position, which has the characteristics of fast convergence and high accuracy.

E. Comparison of Different Machine Learning Methods

In order to verify the prediction performance of the method designed in this paper, it is compared with the mainstream single machine learning method and integrated learning method. Single machine learning methods include support vector machine (SVM), decision tree (DT) and logistic regression (LR). Integrated learning methods include gradient boosting decision tree (GBDT), XGBoost (XGB) with default value and the algorithm in this paper (SSA - XGB).

In order to avoid the contingency caused by random data division, five experiments are used to calculate the accuracy respectively, and the data set is divided into training set and test set according to 8:2. The experimental results are shown in Fig. 4, and the average accuracy is shown in Table IV.

TABLE IV. AVERAGE ACCURACY OF DIFFERENT ALGORITHMS

	SVM	DT	LR	GBDT	XGB	SSA- XGB
Accuracy	0.776	0.694	0.775	0.793	0.795	0.824

The experimental results show that the SSA- XGB achieves the best effect in performance prediction, followed by XGBoost with default value, and the worst is decision tree method.

Compared with XGB and GDBT, the SSA- XGB is 2.9% and 3.1% higher. The three integrated learning methods are better than the three single learner methods. Theoretically, the ensemble classifier is composed of multiple base learners. The prediction error of one base learner can be corrected by other base learners. The prediction error of a single classifier cannot be corrected. XGBoost algorithm improves the shortcomings of GBDT, such as the second-order Taylor expansion of the

objective function and the addition of regularization term. Compared with the default XGBoost algorithm, SSA-XGB shows that the optimization of hyperparameters is helpful to improve the prediction performance.
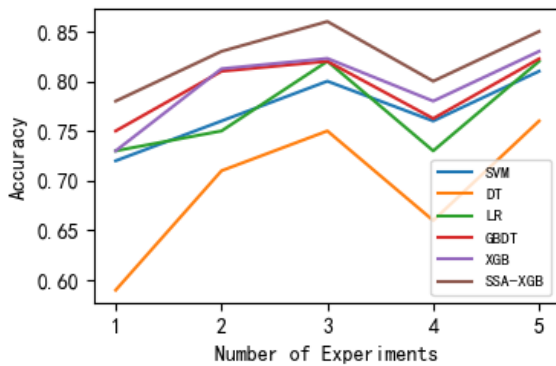

Fig. 4.   Comparison of Accuracy of Different Algorithms.

V.   CONCLUSION AND FUTURE WORK

The high dimension of academic data and the complex optimization of XGBoost hyperparameters are devoted to the paper research respectively. The important characteristics of academic data are extracted by random forest. The random forest algorithm using forward search strategy can effectively extract the feature subset and help to improve the prediction accuracy. The features which have a great impact on students' performance are mainly students' learning behavior characteristics, whose significance to 67 percent of all characteristics. Therefore, teachers should pay more attention to students' learning behavior status and help to improve students' academic performance. The adjustment of hyperparameters can improve the prediction performance of ensemble learning. Sparrow search optimization algorithm is more efficient than other methods of adjusting hyperparameters. Compared with other prediction methods, XGBoost prediction method has the perfect performance. The method designed in this paper enriches the methods of academic prediction in the field of educational data mining, and has a certain reference value for teachers to teach students according to their aptitude and students' personalized learning.

In future, the effects of different categories of guidance on students' academic performance can be studied respectively according to the characteristics of different categories. In addition, whether the combination of different features or the extraction of higher-order features is more conducive to academic prediction is also worth investing.

ACKNOWLEDGMENT

REFERENCES

[1]   Hang Hu,Yang Yang, "Pathways and strategies for deeper learning evaluation: a multi-modal data analysis," Distance Education in China, Vol. 9, PP.13-19, 2022.

[2]   Song Dan,Liu DongBo, and Feng Xia, "Course Performance Prediction and Course Early Warning Research Based on Multi-source Data Analysis," Research in Higher Education of Engineering, Vol. 1, PP. 189-194, 2020.

[3]   Zhou Q, Mou C, and Yang D, "Research progress on educational data mining: A survey," Journal of Software, Vol. 26,No.11, PP. 3026-3042,2015.

[4]   Chen Kaiquan, Zhang Chunxue, et al., "Multi-modal Learning Analysis , Adaptive Feedback and Human-computer Coordination of Artificial Intelligence in Education,"Journal of distance education, Vol. 37,No.5, PP. 24-34,2019.

[5]   Hang Hu, Shuang Du, and Jiarou Liang, "Towards a prediction model of learning performance:informed by learning behavior big data analytics," Distance Education in China, Vol.4, PP. 8-20, 2021.

[6]   Yizhou Fan, and Qiong Wang, "Prediction of academic performance and risk: a review of literature on predicative indicators in learning analytics,"Distance Education in China, Vol. 1, PP. 5-15,2018.

[7]   Li Y Z ,and Hao Z, "Smart Campus Big Data Education Application Based on Students' Behavior Analysis Model," China Educational Technology, Vol. 7, PP. 33-38,2018.

[8]   Zijiang Chen, and Xiaoliang Zhu, "Research on Prediction Model of Online Learners' Academic Achievement Based on Educational Data Mining," China Educational Technology,Vol 12, PP. 75-81,2017.

[9]   Febro J D, "Utilizing Feature Selection in Identifying Predicting Factors of Student Retention," International Journal of Advanced Computer Science and Applications, Vol. 10, No. 9 , PP.269-274,2019.

[10] Cao Y, Gao J, Lian D, et al., "Orderness Predicts Academic Performance: Behavioral Analysis on Campus Lifestyle," Journal of The Royal Society Interface, 2017.

[11] Wen Xiao, Ping Ji, and Juan Hu, "RnkHEU: A Hybrid Feature Selection Method for Predicting Students' Performance.Scientific Programming,". Scientific Programming, PP.1-16,2021.

[12] Rizvi S, Rienties B, and Khoja SA, "The role of demographics in online learning; a decision tree based approach," Comput Educ,Vol 137, PP. 32-47,2019.

[13] Ashraf M, Zaman M, and Ahmed M, "An Intelligent prediction system for educational data mining based on ensemble and filtering approaches," Procedia Comput Sci,Vol 167,pp. 1471-1483,2020.

[14] Huang AY, Lu OH, and Huang JC, "Predicting students' academic performance by using educational big data and learning analytics: evaluation of classification methods and learning logs," Interactive Learn Environ,Vol 28,No 2, PP.206-230,2020.

[15] Qing Wu, and Ruguo Luo, "Predicting the Students' Performances and Reflecting the Teaching Strategies Based on the E-Learning Behaviors,"Modern Educational Technology,Vol. 27,No. 6, PP 18-24,2017.

[16] Liu B P,Fan T C, and Yang H, "Research on application of early warning of students' achievement based on data mining,"ournal of Sichuan University(Natural Science Edition), Vol. 56,No.2, PP. 267-272,2019.

[17] Wang Jinyang,Wang Mengcheng, and Dai Xiaoyang, "Intelligence and personality predict high school students' academic performance: a big five Perspective,"Chinese Journal of Clinical Psychology, Vol. 19,No. 6, PP. 824-826,2011.

[18] Du Xiaoming,Ge Shilun, and Wang Nianxin, "Academic Prediction Based on CNN_LSTM Hybrid Neural Network,"Modern Educational Technology, Vol.31,No.12, PP. 69-76,2021.

[19] Cao Hongjiang， and XIE Jin, "LSTM-based Learning Achievement Prediction and Its Influencing Factors,"Journal of Beijing University of Posts and Telecommunications(Social Sciences Edition). Vol. 22,No. 6, PP. 90-100,2020.

[20] Ding Xinfang,Nie Jing, and Zhang Bin, "Psychological factors and academic performance: a classified prediction model of machine learning ,"Journal of Psychological Science, Vol. 44,No. 2, PP. 330-339,2021.

[21] Yao H , Lian D , Cao Y , et al. "Predicting Academic Performance for College Students: A Campus Behavior Perspective," ACM transactions on intelligent systems, Vol. 10,No.3, PP. 1-21,2019.

[22] Hao Siyuan, and Xie Taifeng, "Study on the Application of Machine Learning in Predicting the Academic Completion Rate of MOOC

Learners," Mathematics in Practice and Theory,Vol.49,No.21, PP. 84-94,2019.

[23] Xu Jiazhen, Deng Wei, and Wei Yantao, "Automatic Recognition of Student's Classroom Behaviors based on Human Skeleton Information Extraction,"Modern Educational Technology, Vol.30,No.5, PP. 108-113,2020.

[24] Cao Rui,Liao Bin,Li Min, "Predicting Prices and Analyzing Features of Online Short-Term Rentals Based on XGBoost," Data Analysis and Knowledge Discovery Vol.5,No.6, PP. 51-65,2021.

[25] Xue J, and Shen B, "A novel swarm intelligence optimization approach : Sparrow search algorithm,"Systems Science &Control Engineering, Vol.8,No.1, PP .22-34, 2020.