Prediction of learning behavior based on improved random forest algorithm

Xiaona Xia\textsuperscript{1,2,3*}

\textsuperscript{1}Chinese Academy of Education Big Data, Qufu Normal University, Qufu, Shandong, 273165, China
\textsuperscript{2}School of Information Science and Engineering, Qufu Normal University, Rizhao, Shandong, 276826, China
\textsuperscript{3}Faculty of Education, Qufu Normal University, Qufu, Shandong, 273165, China
*Corresponding author’s e-mail: xiaxn@qfnu.edu.cn

Abstract. The research of learning behaviors is the premise of improving the learning process and results. The scale, composition and quality of the corresponding dataset will have the key impacts on research methods. Based on big dataset of learning behaviors, this study demonstrates the rules of learning behaviors, uses intelligent learning algorithms to analyze data and predict trends. Based on data characteristics, this paper evolves random forest algorithm and constructs the data structure. Taking the assessment results of learners as the test targets, the data samples are divided into two parts: train set and test set. Through the train set, we design the data model. With the help of the test set, the optimal model is finally selected, that is, the learning interaction activities are predicted and estimated. The research shows that learning behaviors have predictability and guidance, transferability and adjustability, groupness and personality. Based on this conclusion, the data analysis scheme of learning platforms is designed, which will help to improve the platform construction and learning behaviors.

1. Introduction

Learning behaviors are the basis of accurate teaching and personalized learning guidance, and is the key data to realize the prediction and decision feedback\cite{1}. The tracking and analysis of learning behavior is of great significance to the research of education big data issues\cite{2}. Generally, there are two types of learning behaviors: online and offline\cite{3}. The data generated by learners with the help of online learning platform is online data, which is generally tracked, collected, screened and stored by computer programs and database technology\cite{4}, with the integrity and long-term of data management, the online data will not be worn out; offline data is formed by the traditional face-to-face classroom teaching and experimental practice process, its storage and management are not stable, easy to wear or missing. Therefore, about data-driven learning behaviors \cite{5}, complete online data has more advantages.

Applying online data analysis technology, statistical methods and intelligent decision algorithms to the research of online education big data are the main ideas of studying learning behaviors. We extract the learning rules, behavior characteristics and the relationship between learning interaction activities and learning results, so as to mine the problems existing in the teaching and learning process and the decision-making measures\cite{7}, which puts forward higher requirements for the periodicity and effectiveness of the learning behavior dataset to the data processing technology\cite{8}.

The research of online learning behavior is inseparable from the support of large-scale learning behavior dataset[9]. If we want to optimize the research process of data serialization, we need effective data structure, logical rule and relationship. On the one hand, the dataset should be relatively complete and large-scale, involving multiple learning period for data analysis and comparison; on the other hand, the dataset should be open and shared[10], so as to realize the multi-dimensional comparative experiment and demonstration. In this respect, it is subject to the basic business logic of online learning platform system[11]. It is of great significance to select the appropriate dataset for the research method design and conclusion verification.

Based on the open dataset OULAD, this study analyzes and demonstrates online learning behaviors and designs an appropriate data classification and prediction algorithm. Therefore, based on the attributes of data and the characteristics of learning behaviors, the random forest algorithm is optimized and applied to the model design and data prediction of samples. The results of the study can provide feasible technical guidance of learners' potential learning behaviors and the mining of personalized preferences, and also provide method reference for the improvement of online teaching mode and the implementation of decision feedback.

2. Related work

At present, the application of data mining technology, intelligent computing model and machine learning algorithm to the analysis of learning behaviors can realize decision-making prediction based on education big data[12], there have been some achievements that can be used for references and provide multi-dimensional space for research methods and technologies of learning behaviors[13].

In recent years, the National Institutes of health of the United States has launched the "big data to knowledge" program and established the corresponding training coordination center, striving to develop personalized data science training resources for biomedical researchers. The platform is supported by erudite, the discovery engine of educational resources, which collects and classifies data science training resources, including tutorials and research lecture videos, textbooks and other online resources. Despite the formation of so many potential learning resources, they are highly heterogeneous in quality, difficulty, format and theme, and the constraints of data processing are harsh, so it is difficult to develop effective data mapping and association. The development of data science is inseparable from data extraction, data integration, machine learning, information retrieval, natural language processing and other technologies, so as to realize the automatic collection, integration, description and organization of existing online data resources[14].

Yuan Liping has proposed to introduce width learning method into education data mining technology[15], and strive to obtain meaningful rules and patterns, effective knowledge and information in data mining. In line with the people-oriented concept to realize the ontology value, with the help of education big data mining to realize the tool value, through privacy protection to realize the normative value, and to maintain the human-computer cooperation mode, the intelligent value of width learning is realized.

Shi Yafei et al. Design a recommendation model based on learning profile for accurate and personalized learning path generation[16]. According to the learning state of learners, the model can adaptively adopt personalized tracking mechanism and flexible strategy to recommend meta learning element list to learners. Learners can actively select the most suitable meta learning element to promote the learning process.

How to use data and models to predict learning results and find out the existing rules is a hot issue in the field of learning analysis. Sun Fanqin sums up the main factors affecting the academic level of learners by combing relevant literature[17]. Combined with the in-depth processing of the original data, advanced behavior indicators are obtained. Neural network, decision tree and linear regression algorithm are used to model and analyze the data, the authors come to the conclusion that learning attitude, learning timeliness and engagement are the main factors affecting online academic achievement, frustration tolerance is the secondary factor, and interaction level, positive level and stage effect do not affect the final academic performance.
Kyle M. L. Jones uses intelligent learning algorithm to track, aggregate and analyze learner files, as well as the digital and analog behaviors of students in the information system. It is suggested that learning analysis technology should be used to study learning behaviors in higher education. For online learning process, machine learning methods should be taken to promote and adapt to optimization, so that learners can master and monitor personal privacy in time\cite{18}.

Liao, Soohyun Nam describes a modeling method of binary classification of support vector machine, which uses clicker data to predict the final exam results in the third week of the semester. Teachers use peer teaching method to automatically collect clicker data\cite{19}. By training the course data of one semester, they can predict the possible results of the next semester. This modeling method has obvious advantages, for example, it can deal with light-weight learning data sources, achieve the early detection of difficult students, and realize multiple predictions for multiple courses with different curriculum arrangements (different final exams, minor changes in course content, etc.).

The development of artificial intelligence technology not only provides new technical means for the study of learning behavior, but also promotes the development of teaching process to the personalized, precise and intelligent direction of "learner centered". It also plays a positive role in hybrid teaching, which is mainly reflected in promoting the construction of personalized teaching resources, promoting interactive teaching and infiltrative emotional teaching, and promoting classroom teaching Integration of teaching and online learning\cite{20}.

Jiang Bo et al. propose an automatic recognition algorithm of learning puzzles based on facial expression based on the puzzlement inducing experiment of online evaluation. By setting different difficulty test questions to induce the learners to produce confused emotions, and using video equipment to track the learners' facial expressions, extract key features of the facial region are extracted, machine learning algorithm is used to achieve confusion recognition\cite{21}.

Ming Liu et al. Develop a machine learning method to make a reflective statement on the work arrangement of pharmaceutics learners by binary classification. Using the emotional, cognitive and linguistic features of language query and word number analysis, combining with the emotional and rhetorical features of the academic writing analysis platform, four common statistical classifiers are trained for the corpus with 301 sentences. Experiments show that the performance of the algorithm can be improved to some extent by emotional analysis of academic writing and rhetoric of reflective actions\cite{22}.

Qazdar A designs a performance prediction framework based on machine learning algorithm for learning behavior in 2016-2018 of a high school in Morocco\cite{23}. The model is trained and tested by using the learner data collected by the school's management system “Massar”. The prediction results show that the model can accurately predict the learning achievement of learners.

Beaulac et al. Analyze the big dataset of each undergraduate course in a canadian university in the past 10 years \cite{24}. From the perspective of machine learning classification algorithm, two classifiers are constructed. First, using the first two semesters of courses completed by learners predicts whether they will get an undergraduate degree. Second, for those who have completed the course, their major will use the first few courses they have registered again. The results show that the two accurate classifiers and the importance analysis of one variable can provide useful information for university management. Fisnik dalipi et al. use machine learning technology to predict the dropout of MOOC learners, analyze the challenges of prediction model in learning behavior prediction, and based on the results of data analysis, put forward valuable opinions and suggestions, which can help to develop useful and effective solutions to MOOC dropout problems\cite{25}.

These research results focus on the learning process, and introduce machine learning and intelligent algorithm strategies to different degrees in data analysis, which can promote the study of the rules, relationships and rules contained in the learning process. Some of these data are collected based on certain contract conditions in a certain period, and others are limited data of online platform obtained by registered account within the scope of authority. In both cases, the data are partial online or offline data in a controllable environment, and the integrity and continuity cannot be guaranteed. Learning analysis and problem research based on this kind of data are mainly reflected in the scheme planning
of top-level decision-making method, the application and practice of appropriate algorithm for local
data. It is difficult to completely desensitize the dataset, which is not conducive to the comparative
verification of serialization based on the same dataset.

In view of the complex relationship and special structure of learning behaviors, in addition to ensuring
the large scale of dataset, it is also necessary to establish standardized data processing and
standardized transposition, and the premise is effective data desensitization and privacy protection.
Therefore, the open and shared education big data is more conducive to the research and
demonstration of intelligent algorithm and learning strategy. It can also put forward different problem
hypotheses for the same dataset, and construct the mathematical model and implementation algorithm
from different perspectives. The data is trained and tested. If it is a demonstration of the same problem
in the same dataset, we can compare the analysis and prediction quality of the models and algorithms
by the differences of the results. However, few of the existing achievements are based on the open and
shared education big data, and data selection has greater autonomy and limitations, and there are not
unified education big data for the design of appropriate algorithms suitable for data characteristics and
problem researches.

3. Data cleaning and standardization of learning behavior of OULAD

OULAD contains all data related to static description and dynamic interaction of courses, learners and
learning tasks. The whole dataset covers two years, four periods and seven courses. The tracking of
learning behaviors and the collection of data are independently completed by the online learning
platform. All learners complete the remote and open learning process and learning assessment with the
help of online technology. The data can reflect strong autonomy and initiative.

In this study, the learning behavior data of CourseB is selected as the sample, and the final assessment
results of learners are the observation objects of data decision analysis. According to statistics, the size
of CourseB data subset is shown in Figure 1, and the graph is divided into two statistics. One is the
scale of dataset. It can be seen that the data in the third period is less, the data in the second and fourth
periods are larger, and learning interaction records are frequent; the other is the number of
participating sites corresponding to learning activities in each learning period, and the distribution of
the first three semesters is similar, However, in the fourth period, there is no positive correlation
between the interaction of online learning behavior and the online resources involved.

![Figure 1](image1.png)

**Figure 1** CourseB Subset Distribution

![Figure 2](image2.png)

**Figure 2** CourseB Activity Distribution

In OULAD, the distribution of learning interaction activities of participants in CourseB in
each period is as shown in Figure 2. In 2013 and 2014, CourseB involved 12 kinds of learning
interaction activities. However, there are differences in actual participation. The group intensive areas
are relatively concentrated, and several interaction activities such as homepage, quiz and subpage have
obvious approximate distribution in the four periods. Compared with the first three periods, the
amount of online content in D is significantly different. The individual participation of learning
interaction activities such as glossary, oucollaborate, ouelliminate, questionnaire and sharedsubpage is
obvious, and oucontent appears groupness.

In order to test the impact of learning interaction activities on learning results, it is necessary to realize
the data structure transposition with learning interaction activities as the basic data attribute, and form
the dataset to be tested through association and mapping with the final assessment results. Among them, the final assessment conclusions are divided into four categories: distinction, pass, fail and withdrawn. The participation of each learning interaction activity is classified and counted by using “sum_click”. In each period, the learning interaction activities are taken as independent variables, the final assessment result of the course is taken as the test dependent variable, the course code, learning interaction activities and final assessment results of the course learners are taken as one record, and the four learning periods corresponding to four datasets are tested. On this basis, the values of each independent variable and dependent variable are filtered, NA and NULL are normalized, and all data with zero participation in learning interaction activities or blank final assessment results are deleted. Table 1 and Table 2 are respectively statistical description indicators of four datasets. There are ten learning interaction activities in four periods. The relatively concentrated activities with high and low click rate are different in trend distribution. The blank cells in the table indicate that there is no participation record. The data measures of A, B, C and D are 1447, 1731, 1261 and 1921 respectively. The number of learners in learning interaction activities is relatively stable, the growth of learners in D is obvious, and the change of learning interaction activities is also large. According to the data distribution in Figure 1 and Figure 2, the learning behaviors are not directly related to the choice of learning interaction activities and the actual participation of learners, and online learning behaviors have strong individual autonomy and individuality. From the perspective of learning periods, D has the largest difference in participation in learning interaction activities, and the learning behaviors present obvious personalized characteristics. From the analysis of learning interaction activities, forumng and homepage are significantly different. In other cases, learning behaviors basically present the same trend.

### Table 1. Statistical Indicators of Learning Interaction Activities in 2013

|           | S.E | Variance | Kurtosis | Skewness | Confidence (95%) |
|-----------|-----|----------|----------|----------|-----------------|
|           | A   | B        | A        | B        | A   | B      | A    | B | A | B |
| forumng   | 4.53| 3.49     | 29671.91 | 21143.29 | 51.67 | 78.08 | 6.05 | 7.50 | 8.88 | 6.85 |
| glossary  | 0.04| 0.03     | 2.70     | 1.94     | 41.13 | 41.25 | 5.43 | 5.34 | 0.08 | 0.07 |
| homepage  | 0.94| 1.01     | 1279.42  | 1766.36  | 22.47 | 90.55 | 3.46 | 7.03 | 1.84 | 1.98 |
| oucollaborate | 0.02 | 0.65 | 164.27 | 10.80 | 0.04 | 0.65 | 164.27 | 10.80 | 0.04 |
| oucontent  | 0.03| 0.12     | 1.41     | 25.66    | 24.72 | 47.45 | 3.25 | 5.77 | 0.06 | 0.24 |
| ouelluminate | 0.03 | 1.21 | 71.71 | 6.93 | 0.06 | 1.21 | 71.71 | 6.93 | 0.06 |
| questionnaire | quiz | 0.02 | 0.02     | 0.57     | 0.50 | 16.26 | 15.85 | 3.40 | 3.47 | 0.04 | 0.03 |
|           | 0.29| 0.21     | 120.60   | 75.19    | 24.83 | 14.60 | 3.77 | 2.95 | 0.57 | 0.41 |
| resource  | 0.00| 0.00     | 0.02     | 0.02     | 319.69 | 223.75 | 15.44 | 14.02 | 0.01 | 0.01 |
| sharedsubpage | subpage | 0.40 | 0.34     | 231.45   | 197.03  | 13.19 | 10.76 | 2.69 | 2.62 | 0.78 | 0.66 |
| url       | 0.16| 0.16     | 36.54    | 42.21    | 10.51 | 10.43 | 2.37 | 2.50 | 0.31 | 0.31 |

### Table 2. Statistical Indicators of Learning Interaction Activities in 2014

|           | S.E | Variance | Kurtosis | Skewness | Confidence (95%) |
|-----------|-----|----------|----------|----------|-----------------|
|           | C   | D        | C        | D        | C   | D      | C    | D | C | D |
| forumng   | 5.82| 4.91     | 42758.91 | 46401.74 | 55.96 | 42.11 | 6.31 | 5.34 | 11.42 | 9.64 |
| glossary  | 0.03| 0.17     | 1.45     | 53.90    | 33.95 | 114.42 | 5.22 | 8.56 | 0.07 | 0.33 |
4. Principle and implementation rules of improved random forest algorithm

Whether learning interaction activities will have impacts on the final assessment results, and how much influence different learning interaction activities bring, it is necessary to perform iterative learning analysis and demonstration for the test dataset, retrieve the optimal data sampling method, complete multiple data classification and fitting through different levels, according to the analysis results, we can make a decision-making. This process is the integrated learning mechanism of intelligent decision-making, needs strong data generalization performances, and ensures that the individual learners of the integrated processes are relatively independent as much as possible.

About one dataset, after the sampling is completed, several different subsets can be generated, and then a learner can be trained from each subset. The different training data determines that there may be differences. However, if the samples are completely different, the learner can only be built based on a small part of the training data. If data learning is not comprehensive enough, the quality cannot be guaranteed. Therefore, the decision-making analysis of learning interaction activities can consider the implementation of overlapping samples.

Because the final learning results are divided into four categories, but the existing decision-making algorithms are mainly based on two categories of observation objects. On the one hand, it is necessary to properly merge the final learning results of the dataset; on the other hand, the data analysis process should not be the direct application of the existing algorithms, and it needs to improve algorithms. In the decision-making algorithm, the accuracy and robustness of the random forest classifier have relative advantages. Based on the basic mechanism of the random forest classifier, the data structure and attribute characteristics are optimized.

4.1. Bagging

Bagging is a typical application of parallel integrated learning method[26], based on bootstrap sampling [27]. The sampling process is described as follows: given a dataset containing $M$ samples, first randomly take a sample and put it into the sampling set, then put the sample back into the initial dataset, and the sample may be selected in the next sampling. After $m$ random sampling, a set containing $M$ samples is obtained. Some samples in the initial training set appear many times in the sampling set, some never appear. According to the demonstration results in [27], about 63.2% of the samples in the initial training set can be included in the sampling set. For learning interaction dataset, based on bagging, we can get $T$ sub sets with $m$ training samples. Each sample set can train a basic learner, which can form $T$ basic learners. Taking the final assessment as the observation variable with $T$ basic learners, the prediction output of learning process is selected by random voting method. The learning analysis process of bagging is described as algorithm 1. $\Pi(\cdot)$ is an indicator function. If the
result of ‘‘=’’ is true, the return value is 1, the result of ‘‘=’’ is false, and the return value is 1. The flow of base learning algorithm can be found in [28].

If the time complexity of learning process is $O(m)$, and the time complexity of the sampling process is $O(s)$, the time complexity of bagging is approximately $T(O(m)+O(s))$, and $T$ is a small constant. The time complexity of training a bagging integration is the same as that of directly using the base learning algorithm, which shows that the implementation of the bagging algorithm is efficient. In the self-sampling process of bagging, 63.2% of the samples are used as the training set, and the remaining 36.8% of the samples can be used as the out of bag estimate. Therefore, it is necessary to record the training samples used. Supposing $D_t$ is the actual training set of $h_t$, $H_{oob}^t(x)$ is the out of bag prediction, then:

$$H_{oob}^t(x) = \arg\max_{y \in Y} \sum_{i=1}^{T} \left(h_i(x) = y\right) \cdot \left(x \not\in D_t \right)$$

The out of bag estimate of bagging generalization error is as follows:

$$\varepsilon_{oob} = \frac{1}{|D|} \sum_{(x,y) \in D} \left(H_{oob}^t(x) \neq y\right)$$

Algorithm 1

| Input: Taining set $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\}$ |
| base learning algorithm $\mathcal{E}$ |
| Training times of data set $T$ |
| Process: |
| 1. For $t = 1$ to $T$ do |
| 2. $h_t = \mathcal{E}(D_t, D_{\neg t})$ |
| 3. End for |
| Output: $H(x) = \arg\max_{y \in Y} \sum_{i=1}^{T} \left(h_i(x) = y\right)$ |

4.2. Improved random forest

Random Forest (RF) is one direct application of Bagging. RF is decision tree, its supporting is based on the integration of Bagging, random attribute selection mechanism is introduced. Assuming that each node of the base decision tree has $d$ attributes, a subset containing $k$ attributes is randomly selected from the attribute set of the node, and then an optimal attribute is selected from the subset for partition, $k$ controls the random introduction degree. If $k = d$, the construction of the base decision tree is the same as that of the traditional decision tree; if $k = 1$, one attribute is randomly selected for division, generally, we select the recommended value, that is $k = \log_2 d$.

Random forest is an improvement of bagging. The diversity shows not only the sample disturbance, but also the attribute disturbance. As a result, the generalized performance of the final integration is further improved with the increase of the differences of learners. In addition, the convergence of random forest is similar to bagging, the number of individual learners increases, and the random forest usually converges to a lower generalized error.

The improved random forest algorithm aims at multi-objective classification, through training multiple decision trees, we generate models and iteratively train data, synthesizing multiple decision trees, that can determine the best analysis results and ensure the analysis efficiency and accuracy. The improved random forest classifier is described as algorithm 2.

According to Algorithm 2, $t$ decision trees are generated. For each test sample, the optimal one is selected as the final classification result of the random forest by synthesizing the classification results.

There are two possible types of target attributes:

1. The target attribute is numerical type: we need to take the average value of a decision tree as the classification result.
(2) The target attribute is the category type: the minority is subject to the majority, and the category with the most classification results of a single tree is taken as the classification result of the whole random forest. The goal attribute of the learning interaction dataset is that the final assessment of the learning process belongs to the category, and the second processing scheme is implemented during data training process.

Algorithm 2

| Input | the number of decision tree \( r \), the number of input attributes \( m \) when each node of the decision tree is split. |
|-------|---------------------------------------------------------------|
| Output | Output \( t \) decision trees. |
| Process | (1) \( i = 1 \) /* \( i \) is the control variable used to accumulate the number of decision trees generated */ |
|         | (2) If \( N \) is the number of training samples, then a single decision tree needs randomly to select \( m \) training samples from \( N \) training sets. |
|         | (3) Set the number of input attributes of training samples is \( M \), \( m \) is far less than \( M \). When each node of one decision tree is split, randomly select \( m \) attributes from \( M \). From \( m \) input attributes, then select an optimal attribute to split, \( m \) will not change during the construction of the decision tree. |
|         | (4) Each decision tree divides iteratively according to the process of (3), until all training samples of the node belong to the same class. No more pruning. |
|         | (5) \( i = i + 1 \) |
|         | (6) if \(( i \leq t )\) goto (2) |
|         | else break |
|         | End Process |

5. Data test and empirical analysis

Based on the improved random forest classification algorithm, four datasets are classified and predicted. The measurement of data processing results needs two indicators: frequency and dimension. Frequency is the learning activity in a certain period, which is reflected in the interaction processes between learners and online learning platform, expressed by “click rate” in OULAD. Dimension refers to the participation of each learner in all learning interaction activities of one course in a certain period. If there is participation data for an activity learner, the dimension component is recorded as 1, otherwise it is recorded as 0.

Frequency is the activity of learners about the online learning process. Combined with the characteristics of OULAD, before data analysis, the calculation model of learners’ participation frequency is defined. The specific characteristics are as follows:

If a course involves \( m \) learning interaction activities in a learning period, which is expressed as \( A_1, A_2, \cdots, A_m \); \( n \) learners have completed the learning processes and produced the final assessments, which are expressed as \( S_1, S_2, \cdots, S_n \). The interaction value (click rate) between each learner \( S_i (1 \leq i \leq n) \) and each learning interaction activity \( A_j (1 \leq j \leq m) \) is expressed as \( (S_i, A_j) \), \( \sum_{i=1}^{n} (S_i, A_j) \), that is the total participation frequency of each activity, the total participation frequency of the learning interaction activities of the course in the period is expressed as \( \sum_{j=1}^{m} \sum_{i=1}^{n} (S_i, A_j) \). The
participation weight of each learning interaction activity is \( \lambda_j = \frac{\sum_{i=1}^{n}(S_j, A_i)}{\sum_{j=1}^{m} \sum_{i=1}^{n}(S_j, A_i)} \), \( \sum_{j=1}^{m} \lambda_j = 1 \). Then the weighted participation frequency calculation model of \( S_j \) is \( P_j = \frac{1}{m} \left( \sum_{j=1}^{m} \lambda_j (S_j, A_j) \right) \).

The calculation model of learners’ participation frequency also accounts for the overall distribution of learning interaction activities in a complete learning period. Mapping the fitting weights of local data from a global perspective is beneficial to data analysis and prediction that can meet the relatively optimal data estimation.

For the measurement of the participation dimension of learning interaction activities, the calculation model is expressed as \( \sum_{j=1}^{m} \Pi \left( (S_j, A_j) \right) \), if the value of \( (S_j, A_j) \) is not equal to 0, \( \Pi \left( (S_j, A_j) \right) = 1 \), otherwise \( \Pi \left( (S_j, A_j) \right) = 0 \).

Frequency and dimension are used to generate the observation variables of four datasets to be tested. The data analysis of Algorithm 1 and Algorithm 2 is completed from two aspects of "learning interaction activity-frequency-final assessment result" and "learning interaction activity -dimension-final assessment result". Each dataset is divided into training set and test set according to the proportion of 7:3. “Distinction” in the final_result value is combined into "pass" in the learner course assessment for calculation.

The outliers are tested and screened by \( Z \) test. If we want to test whether the difference between the average number \( \bar{X} \) of a sample and the average number \( u_0 \) of a known population is significant, \( Z \) calculation formula is \( Z = \frac{\bar{X} - u_0}{s/\sqrt{n}} \), \( \bar{X} \) is the average number of the test sample, \( u_0 \) is the average number of the known population, \( s \) is the standard deviation of the sample, and \( n \) is the sample capacity.

According to the statistical indicators, the learning interaction activities with \( Z \in (-\infty, 3] \) or \( Z \in [3, +\infty) \) are classified as outliers and deleted, the relevant parameters of the algorithms are optimized for data imbalance. According to the implementation of the algorithms, the sampling scheme is adjusted, the data is trained repeatedly, and the best learning result of the decision tree is selected. The relation curve between the algorithm execution error rate of four datasets and the number of decision trees are shown in Figure 3 (1) - (4). From the four subgraphs, it can be seen that when the number of decision trees of four datasets reaches about 200, the number of effective decision trees converges, and the prediction results tend to be stable with high accuracy.
The influence of learning interaction activities on the observation value of each node in the decision tree is measured by MeanDecreaseGini, which calculates the influence of each learning interaction activity through Gini index. The greater the value, the greater the importance of the activity. Among them, Gini is the index of the optimal feature of the classical decision tree CART. Suppose there is $K$ classes, the probability of the sample belonging to $k$ class is $p_k$, then the Gini index of the probability distribution is defined as $G(p) = \sum_{k=1}^{K} p_k (1 - p_k) = 1 - \sum_{k=1}^{K} p_k^2$, $\sum_{k=1}^{K} p_k = 1$.

The classification prediction analysis between learning interaction activities and final assessment results show that the overall distribution of group learning interaction activities is relatively consistent, there are differences in individual activities, individual autonomous activities are obvious, but the participation data is very small. Learning interaction activities will affect learning results, but the key problem is that learners should select effective learning interaction activities to improve learning results. There is no positive correlation between online participation and learning results. This conclusion has not been proved by complete and real dataset in previous research results.
Because there are serious data imbalances in the four datasets, the data distribution proportion of pass, fail and withdrawn as the three classification prediction targets is maintained at 3:1:1, and the participation degree distribution of learning interaction activities is greatly different. There are abnormal values and noises in datasets, which affect the overall distribution, and the relevant statistical indicators are not ideal. The decision classification and learning prediction of data do not really reflect the essential relationship between online learning behaviors and learning results, which causes obstacles to prediction. The traditional classification learning algorithm applied to the analysis of dataset is generally not suitable. For this reason, the data has been fully standardized. The empirical results show that the improved random forest prediction accuracy rate is the highest. Table 3 shows the algorithm prediction accuracy rate of four periods, which is more than 90%. The prediction results can be used for references.

Table 3. Prediction Accuracy

| Predict Accuracy | A      | B      | C      | D      |
|------------------|--------|--------|--------|--------|
|                  | 0.9187605 | 0.9050523 | 0.9160494 | 0.9482484 |

6. Conclusion and reflection

Because learning behaviors are related to learners' autonomous learning process and guiding learning objectives[29], it makes it difficult for data tracking and management, and there are many nonstandard and unstructured data in the quantitative description of learning behaviors[30]. Learning as a correlation process between learners and knowledge, whether or not realizes data-driven behavior prediction through decision-making learning analysis, that is the significance of education big data research. In this study, aiming at the learning behaviors of multiple observation variables, we improve the random forest decision algorithm, select or design effective data cleaning, standardization, equilibrium and outlier location[31], and realize the intelligent learning analysis of complex learning behaviors. Through data training and testing, the following conclusions are drawn:

6.1. Predictability and guidance of learning interaction activities

Based on the characteristics of learning interaction activities, an effective learning analysis algorithm is designed and implemented, which can predict the learning interaction activities with high accuracy. In this study, the dataset is divided into two parts according to the proportion of 7:3: training set and test set. Based on the training set, the estimation model of learning interaction activities and final assessment is established. Through the test set, the performance of the model, the stability and predictability of the model are evaluated.

The empirical study of complex learning interaction activities can predict the relationship between learning behaviors and learning results, and judge the potential learning interaction activities of
learners, so as to drive sustainable behaviors through the tracking and analysis of data rules and the adjustment and optimization of data structure, meanwhile, we can improve or design adaptive intelligent learning analysis based on data characteristics[32], which can promote the formation and deepening of learners' good learning behaviors in online learning, open learning and distance learning. The predictability of learning interaction activities can provide strong feasibility support for business demand analysis of online platform construction. There are many existing online learning platforms, but platform architecture with adaptive data analysis and prediction mechanism business is still in the demonstration stage, the main reason is that it is difficult to obtain the learning process data [33]. This study uses OULAD to carry out data analysis and decision-making prediction. The research results have reference value for method design and practical scheme.

6.2. The potential mobility and adjustability of learning interaction activities
From the statistical indicators and analysis results of the third and fifth parts of this study, we can see that there is a certain correlation between learning interaction activities. "Homepage -subpage-URL" is an interactive activity realized with the help of web pages, "oucontent-resource-glossary" provides online learning resources, and quiz is a flexible test method of learning process, which corresponds to the final assessment. “forumg” is an effective means of online interaction processes between learners and lecturers. These learning interaction activities themselves have goal coupling in the preferences and participation modes of learning behaviors. The participation of one learning interaction activity will jointly improve the participation of other related activities, which is similar to the other learning interaction activities and has a similar synchronous trend.

Therefore, with the help of intelligent classification and clustering algorithm, we can realize the class division and formation of online learning interaction activities. For the activities of the same class, about the settings of learning interaction activities of online learning platform, we should set the planning of similar classes. When learners choose learning interaction activities, they belong to the same class of activities, can carry out predictive rank of association, and recommend to learners[34]. In this way, on the one hand, it can improve the participation of online learning platform related learning interaction activities, the generation process of data can partially ensure the relative integrity and balance of data distribution, and give full play to the role of learning interaction activities without excessive abnormal values and noises; On the other hand, we expand the effective learning style of learners. According to the nature and difficulty of learning contents, try to choose the approximate learning interaction activities, which will not cause the "low efficiency" of learning results because of the "invalid" of a learning interaction activity. This study has demonstrated that the frequency and dimension of participation in learning interaction activities do not necessarily have a significant impact on learning outcomes. Effective prediction and timely recommendation of key learning interaction activities are conducive to improving learning outcomes.

6.3. Groupness and personality of learning interaction activities
Learners participate in online course learning, the same course has group tendency and individual tendency in different periods. When using algorithms to train and test the learning behavior datasets of CourseB in different periods, MeanDecreaseGini shows this conclusion. Combined with the data statistics of the third part, the participation degree variance of learning interaction activities with larger influence degree is large, and the participation of learning interaction activities is seriously polarized. Pass, Fail and Withdrawn are the three assessment results, among the three classification objectives of learning interaction activities, the data distribution ratio is basically maintained at 3:1:1. Online learning based on self-study makes the learning process scattered in different places inseparable from individual autonomy. About one fifth of learners leave a large number of learning interaction activities, but give up the final assessment of the course.
Figure 5. Learning Behavior Tracking and Prediction

Based on the datasets of learning platform, the online learning process and related attributes are analyzed. According to the research objectives, we can put forward the problem hypothesis, select the data needed for the problem demonstration, establish the association requirements of the learning behavior characteristics according to the internal data relationship, and on this basis, carry out data merging, fitting and data cleaning. We can also standardize the dataset to be tested, select or design appropriate algorithms to train data, use the training results to predict the learning behavior and make decisions, with the help of recommendation strategy, the feedback and guidance of potential learning interaction activities are realized, so as to achieve the improvement of learning characteristics. Taking the data analysis mechanism of online learning behaviors as the decision-making business, it is an important link to realize personalized learning behaviors, probability prediction and decision feedback.

7. Summary

Research on learning behaviors are the premise of improving learning process and optimizing learning results. The scale, composition and quality of datasets will have a key impact on the choice of research methods and the decision-making of research results, and also increase the difficulty of empirical research on learning behavior. This study takes the big dataset of learning behavior of OULAD as the research object, selects the learning behavior data of CourseB as the sample, and takes the final assessment results of learners as the test objective to verify the impact of learning interaction activities. The sample is divided into two parts: training set and test set, and the data model is constructed and validated respectively. According to the multi-objective classification and prediction requirements, the random forest algorithm is improved. For the model prediction accuracy of complex behavior data, the algorithm is more than 90%. The data analysis results are effective and the prediction results are meaningful. The research shows that learning interaction has predictability and guidance, potential mobility and adjustability, groupness and personality. Based on this conclusion, the data analysis and prediction process of online learning platform is designed, which has reference value in both platform construction and learning behavior improvement.
Acknowledgment
This study is supported by National Natural Science Foundation of China (NSFC) "Optimal Design and Data Analysis of Several Kinds of Factor Tests" (Grant NO. 11771250); Shandong Provincial Social Science Planning and Management Office "Educational Decision Platform Architecture Supported by Big Data" (Grant NO. 19CZKJ08); Shandong Province Quality Core Curriculum of Postgraduate Education "Advanced Software Engineering" (Grant NO. SDYKC18079); Natural Science Foundation of Shandong Province "Social Clothing Research and application of decision-making mechanism of service recommendation" (Grant NO. ZR2016FM45).

References
[1] Ambite, J. L., Fierro, L., Gordon, J., Burns, G., Geigl, F., Lerman, K., et al. (2019). Bd2k training coordinating center's erudite: the educational resource discovery index for data science. IEEE Transactions on Emerging Topics in Computing, pp.1-1.
[2] Baojian Zhang, Xiaokong Zhang, Bo Pei, Guoqiang Sun (2019). Perceived value, cognitive process and behavioral intention: a configuration analysis of MOOC Learning. Distance Education in China, (9): 72-82, 93.
[3] Beaulac, Cédric, Rosenthal, J. S.. (2018). Predicting university students' academic success and major using random forests, 1-25.
[4] Breiman, Leo (1996). Bagging predictors. Machine Learning, 24(2):123-140.
[5] Breiman Leo (2001). Random Forests. Machine Learning, 45(1):5-32.
[6] CHEN Jin-yin, WANG Zhen, CHEN Jin-yu, CHEN Zhi-qing, ZHENG Hai-bin(2019). Design and Research on Intelligent Teaching System Based on Deep Learning. Computer Science, (B06):550-554.
[7] DAI Yonghui, XU Bo, CHEN Haijian (2018). Artificial Intelligence on the Promotion of Blending Learning and the Construction of Its Ecological Chain. Modern Distance Education Research, (2): 24-31.
[8] Dalipi, F., Imran, A. S., Kastrati, Z. (2018). MOOC Dropout Prediction Using Machine Learning Techniques: Review and Research Challenges. EDUCON’2018.
[9] DENG Li-juan, ZHANG Bei (2019). Students' Behavior Data Analysis Based on the Cloud Classroom Administration Platform. Journal of Jimei University, 20(03): 63-69.
[10] Efron, B., Rogosa, D., Tibshirani, R.. (2015). Resampling Methods of Estimation. International Encyclopedia of the Social Behavioral Sciences, pp.492-495.
[11] Ferguson R(2012). Learning analytics: Drivers, developments and challenges. International Journal of Technology Enhanced Learning, 4(5/6): 304-317.
[12] Hai-Ling, L., Jun-Huai, L., Jun, P.. (2018). Big Data Technology Applied to Learning Behavior Evaluation System. 2018 International Conference on Big Data and Artificial Intelligence (BDAI). IEEE, pp. 10-17.
[13] Li, H. , Zhang, J., Zhao, X. , Si, J., Huang, B. . (2019). Relationships among epistemological beliefs, self-regulated learning and academic procrastination in college students: a moderated mediation model. Psychological Development and Education, pp. 557-565.
[14] Hubalovsky, S., Hubalovska, M., Musilek, M. . (2018). Assessment of the influence of adaptive e-learning on learning effectiveness of primary school pupils. Computers in Human Behavior, 92: 691-705.
[15] JIANG Bo, LI Wanjian, LI Zhixuan, YE Yun(2018). Automatic Learning Confusion Recognition Using Facial Expressions. Open Education Research, (4):101-108.
[16] Jones, K. M. L.. (2019). Learning analytics and higher education: a proposed model for establishing informed consent mechanisms to promote student privacy and autonomy. International Journal of Educational Technology in Higher Education.
[17] Li, Y. , Xing, S., Bao, H., Su, Y., Yuan, Z.. (2019). The effects of learning analytics tools on teacher interventions in online collaborative learning. China Educational Technology, (02): 80-86.
[18] Liao, S. N., Zingaro, D., Thai, K., Alvarado, C., Griswold, W. G., Porter, L.. (2019). A robust machine learning technique to predict low-performing students. ACM Transactions on Computing Education, 19(3): 1-19.

[19] Liu, M., Shum, S. B., Mantzourani, E., Lucas, C.. (2019). Evaluating Machine Learning Approaches to Classify Pharmacy Students' Reflective Statements. Artificial Intelligence in Education. Springer, Cham. pp. 220-230.

[20] Mirzaei, M., Sahebi, S.. (2019). Modeling Students’ Behavior Using Sequential Patterns to Predict Their Performance. Artificial Intelligence in Education. pp. 350-353.

[21] PENG Zi-han, WANG Yun-wu(2019). On Influencing Factors of Group Cooperative Learning Behavior—Based on Blended Learning Mode. Heilongjiang Researches on Higher Education, (6): 141-147.

[22] Qazdar, A., Er-Raha, B., Cherkaoui, C., Mammass, D.. (2019). A machine learning algorithm framework for predicting students performance: a case study of baccalaureate students in morocco. Education and Information Technologies, 24(1): 3577–3589.

[23] Tengzhong, R., Xianghui, M.. (2019). Analysis on higher-order discrete MPT model of MOOC learning behavior data. Statistics Decision, (10): 14-17.

[24] Saqr, M.. (2018). A literature review of empirical research on learning analytics in medical education. International Journal of Health Sciences, 12(2), 80-85.

[25] Schweder, S.. (2019). Mastery goals, positive emotions and learning behavior in self-directed vs. teacher-directed learning. European Journal of Psychology of Education, (3).

[26] Sedrakyan, G., Malmberg, J., Verbert, K., Sanna Järvelä, Kirschner, P. A.. (2018). Linking learning behavior analytics and learning science concepts: designing a learning analytics dashboard for feedback to support learning regulation. Computers in Human Behavior, 107: 1-15.

[27] Shi Yafei, Peng Hongchao, Tong Mingwen(2019). Research on Generative Paths Recommendation Strategies of Precise Personalized Learning Based on Learning Profile. China Educational Technology, (05): 84-91.

[28] Starcic, A. I.. (2019). Human learning and learning analytics in the age of artificial intelligence. British Journal of Educational Technology, 50(6): 2974 – 2976.

[29] Sun Faqin, Feng Rui(2019). A Research on Online Learning Achievement Factors Based on Learning Analysis. China Educational Technology, (03):48-54.

[30] Wang Yang, Liu Qingtang, Zhang Wenchao, David Stein(2019). The Research on Online Learning State Analysis Model and Its Application Driven by Data. Journal of Distance Education, (2):74-80.

[31] Xinyi Shen, Jianwei Wu, Yanxia Zhang, Ying Li, Yuchun Mu(2019). Towards an evaluation model of online learning behavior and learning effectiveness for MOOCAP learners. DISTANCE EDUCATION IN CHINA, (7): 38-46 and 93.

[32] Yang Jiumin, Zhang Yi, Li Li, Pi Zhongling (2019). The Interaction of an Instructor's Guided Behaviors and Learners' Prior Knowledge in Predicting Learning from Video. China Educational Technology, (7):74-81.

[33] Yuan Liping, Chen Chuannan(2019).The Value and Significance of Broad Learning. China Educational Technology, (05):34-40.

[34] ZHANG Muhua, WEI Ning, LI Yan(2019). An Empirical Study on Analyzing Learners’Learning Behaviors Supported by BYOD. Modern Distance Education, (3): 3-12.

[35] Zong Cheng(2019). An Empirical Analysis of the Influencing Factors of Students’ Learning Behavior on the Teaching Resource Database of Vocational Education. JOURNAL OF NATIONAL ACADEMY OF EDUCATION ADMINISTRATION, (8): 73-80.