Early Detecting the At-risk Students in Online Courses Based on Their Behavior Sequences

Shuai Yuan[1], Huan Huang[2(✉)], Tingting He[3], and Rui Hou[4]

[1] National Engineering Research Center for E-Learning, Central China Normal University, Wuhan, China

[2] School of Education, South-Central University for Nationalities, Wuhan, China
huanghuan@mail.scuec.edu.cn

[3] Hubei Provincial Key Laboratory of Artificial Intelligence and Smart Learning, School of Computer, National Language Resources Monitoring & Research Center for Network Media, Central China Normal University, Wuhan, China
tthe@mail.ccnu.edu.cn

[4] College of Computer Science, South-Central University for Nationalities, Wuhan, China

Abstract. Online learning has developed rapidly, but the participation of learners is very low. So it is of great significance to construct a prediction model of learning results, to identify students at risk in time and accurately. We select nine online learning behaviors from one course in Moodle, take one week as the basic unit and 5 weeks as the time node of learning behavior, and the aggregate data and sequence data of the first 5 weeks, the first 10 weeks, the first 15 weeks, the first 20 weeks, the first 25 weeks, the first 30 weeks, the first 35 weeks and the first 39 weeks are formed. Eight classic machine learning methods, i.e. Logistic Regression (LR), Naive Bayes (NB), Radom Forest (RF), K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Iterative Dichotomiser3 (ID3), Classification and Regression Trees (CART), and Neural Network (NN), are used to predict the learning results in different time nodes based on aggregate data and sequence data. The experimental results show that sequence data is more effective than aggregate data to predict learning results. The prediction AUC of RF model on sequence data is 0.77 at the lowest and 0.83 at the highest, the prediction AUC of CART model on sequence data is 0.70 at the lowest and 0.83 at the highest, which are the best models of the eight classic prediction models. Then Radom Forest (RF) model, Classification and Regression Trees (CART) model, recurrent neural network (RNN) model and long short term memory (LSTM) model are used to predict learning results on sequence data; the experimental results show that long short term memory (LSTM) is a model with the highest value of AUC and stable growth based on sequence data, and it is the best model of all models for predicting learning results.

Keywords: Early detecting · The prediction of learning result · Long short term memory

Z. Deze et al. (Eds.): BDTA 2020/WiCON 2020, LNICST 371, pp. 17–33, 2021.
https://doi.org/10.1007/978-3-030-72802-1_2

1 Introduction

In the past decade, online learning has developed rapidly. Thousands of online learning systems have emerged, providing different online learning services for different kinds of learners. Compared with the traditional face-to-face teaching, online learning has many advantages undoubtedly. It breaks the limitation of learning time and space, expands the scale of learners, and effectively improves the autonomy of students. However, there are also some problems in online learning, one of which is that the participation of learners is low [1]. It leads many learners to faile in online courses. To solve this problem, many researchers recently suggest to using big data technology to identify at-risk learners timely and accurately, to provide adaptive learning intervention or support for them [2–4]. According to this, it is of great significance to find an effective learning result predicting method.

Based on the general process of data mining, the basic process of online learning result predicting is as follows: 1) collect learning process data in an all-round way to form a big dataset; 2) select or design some important predicting indicators of learning result based on the learning process data; 3) use a machine learning algorithm to build predicting model of learning result based on the indicators; 4) predict new ones' learning results based on their learning process data. It can be seen that predicting indicator and predicting algorithm are two key components of learning result predicting. For these two components, many scholars have carried out a lot of in-depth researches. In the aspect of predicting indicators, researchers have explored many behavior indicators, such as the total time of online learning, amount of resource views, test scores, and amount of forum posts [5–7]. In the aspect of predicting algorithm, researchers have explored many classic machine learning algorithms, such as Logical Regression (LR), Decision Tree (DT), K-Nearest Neighbor (KNN), Naïve Bayes (NB), Support Vector Machine (SVM), Random Forest (RF), and so on [8, 9]. However, nearly all the existing researches used aggregated data when extracting the predicting indicators, without considering the dynamic pattern of the predicting indicators. Some recent works have shown that some dynamic patterns of learning behavior may reflect the advanced cognitive characteristics of learners, which play an important role in online learning results. Accordingly, if these dynamic patterns are integrated into the predicting model of online learning results, the prediction effect should be improved to a certain extent.

To integrate the dynamic pattern of learning behavior into predicting model and improve predicting accuracy, this paper proposes an online learning result predicting method based on long short term memory (LSTM) neural network. LSTM is an outstanding representative of the recurrent neural network (RNN). RNN is a kind of neural network used to process and predict sequence data. It can mine the hidden dependent relationship or sequential pattern from a large number of sequence data, to achieve accurate predicting of sequential data. LSTM further solves the problem of long-term dependence in sequential data. At present, LSTM has achieved good results in speech recognition, machine translation, and sequential analysis and other applications. Given the remarkable performance of LSTM in processing sequence data, this study tries to apply LSTM in online learning result predicting. Different from the existing predicting methods, this method extracts value sets of predicting indicators based on the online learning behavior data in different time periods and form a sequence data. Based on

the sequence data, it further mines the sequential pattern and its relationship with the learning result by using LSTM.

2 Related Work

Although early warning system for online learning was emerged until recent years, it has been concerned by many researchers since it was put forward. In the past decade, a large number of researches have been carried out on the key issue – learning result predicting. According to the basic process of learning result predicting (collect data, design predicting indicators, develop predicting model, and predict learning result), the existing researches will be examined. In the aspect of data collection, the existing researches mainly used the learning behaviors and test scores recorded in the learning management system [5–7]. However, with the deepening of research in recent years, some researches also used some background information and psychological characteristics of learners through survey, which is also an important basis for learning result predicting. For example, based on the theory of self-regulated learning, Pardo et al. combined the self-regulated learning index and online learning behavior of learners to predict their learning results, among which the self-regulated learning index is obtained through a survey [10]. In addition, most of the researches are based on the data of one course to develop predicting model for a specific course. Still, few researches also use the data of multiple courses to explore the cross-course predicting model. For example, Gašević et al. constructed a cross-course predicting model based on the data of nine courses, and compared it with the predicting models of each specific course [11]. The results show that it should be prudent to integrate the data of multiple course data to develop a cross-course predicting model because learners' online learning behaviors are quite different in different courses [11].

In the aspect of predicting indicators, researchers have explored the impact of many indicators on the effect of learning result predicting from different perspectives. Recently Fan and Wang summarized three kinds of indicators used in learning result predicting through the in-depth analysis of 83 kinds of literature: human-computer interaction indicators, human-human interaction indicators and individual tendency indicators [12]. Human–computer interaction indicators reflect the interaction between learners and learning platform, such as the frequency of login, the total time of online learning, number of browsed resources, number of completed assignments, scores of the tests and so on. Human-human interaction indicators reflect the interaction between the learner and leaner, learner and teacher, mainly include the number of posts, replies, social network location and so on. Individual tendency indicators mainly include background and psychological characteristics reflecting individual differences of learners, such as gender, age, education level, prior knowledge, learning motivation, the level of self-regulated learning and so on. The early research of online learning result predicting mainly used human-computer interaction indicators and human-human interaction indicators, but in recent years more and more researches began to introduce some advanced psychological characteristics into learning result predicting model to further improve its accuracy and interpretability. Although researchers have conducted in-depth research on the predicting indicators, due to different research scenarios and research data, the results of

these studies are not consist of. Recently, Conijn et al. extracted 23 predicting indicators which were commonly used from the log data of 17 courses, and compared the effect of each indicator on the predicting of learning result of different courses [13]. They found that in addition to the mid-term test score is significantly related to the final result in all courses, other indicators are only significantly related to the final result in some courses, and the correlation between the same indicator and the final result shows different effect in different courses [13]. This shows that it is difficult to find a set of general predicting indicators, so we should select appropriate predicting indicators for specific situations.

In the aspect of predicting model development, the predicting variable defined by most of the researches is a binary classification variable. That means the predicting result is whether the learner passed the course or not. However, some researches also defined prediction variable as a continuous numerical variable. That means the prediction variable is a continuous grade of a student. According to the different predicting variable defined, the researchers adopt different predicting algorithms to develop a predicting model. When the predicting variable is the final grade of the student, the most used predicting algorithm is Mmultiple Linear Regression (MLR) [13]. When the predicting variable is whether a student will pass the course or not, the predicting algorithms used by the researchers mainly include Logic Regression (LR), Decision Tree (DT), Naïve Bayes (NB), K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Random Forest (RF) and so on [8, 9]. For example, Marbouti et al. defined the predicting variable as to whether a student will pass the course or not, and they developed different predicting models using LG, DT, NB, KNN, MLP, SVM respectively. However, the experimental results showed that no one model can achieve satisfactory results in all aspects [8]. Therefore, they further used the ensemble learning to develop a prediction model and optimized the model through feature selection and increasing training set. Finally, they found the ensemble model is the best one [8]. Howard et al. defined the predicting variable as the final score, and developed predicting models using RF, BART, SGBoost, PCR, SVM, NN, KNN, respectively [9]. The experimental results show that the sixth week is the best time to identify at-risk students, which not only has enough time to intervene the students, but also can ensure the accuracy of the predicting model. Furthermore, at this time, the model developed by BART gets the best performance [9]. From the existing researches, we can know that although the researchers have compared the effect of a variety of predicting algorithms in learning results predicting, which algorithm was best for learning result predicting has not reached a consistent conclusion. Also, the existing predicting algorithms are the traditional classic machine learning algorithms, and few kinds of research have explored the effect of the latest advanced machine learning algorithms on learning result predicting, such as deep learning algorithms [14].

Based on the research on predicting methods, some institutions have also developed early warning systems for online learning, such as "Course Signals" of Purdue University in the United States, and "OU Analysis" of Open University in the United Kingdom. Course Signals is an early warning system for online courses developed by Purdue University in 2007. It was originally developed for the freshmen of Purdue University to predict the academic performance of students and improve the success rate and retention rate [15]. Course Signals mainly uses four kinds of predicting indicators: test scores, effort levels, previous academic achievements and background information [15]. Based

on the above indicators, Course Signals uses a specific student success algorithm (SSA) to predict the learning results of learners. According to the predicted results, students' learning states are divided into three states: red light (high risk), yellow light (early warning) and green light (good) [15]. The results of a three-year study show that the academic achievement of students using Course Signals is significantly higher than that of students not using the system, and the corresponding retention rate of students is significantly higher than that of students not using the system [15]. OU Analysis is an early warning system for online courses developed by UK Open University. Its goal is to identify at-risk learners as early as possible, to give effective intervention to improve the retention rate of learners. To achieve this goal, OU Analysis selects some background information and online learning behavior of learners as predicting indicators, trains four predicting models using NB, KNN and CART respectively, and finally determines whether students are at risk or not using voting mechanism [16]. OU Analysis provides two views: course view and learner view. The course view shows an overview of all learners' online learning behavior, the predicted results of whether each learner will participate in the next test, and the predicted results of each learner's final score. The learner view shows an overview of a learner's online learning behavior, actual scores and predicted results of each test, as well as recommended learning activities and learning resources [16]. As of the summer of 2016, OU Analysis has been widely used in more than 40 courses of UK Open University.

3 Proposed LSTM-Based Framework

In order to integrate the dynamic pattern of learning behavior into the learning result predicting model, this paper proposes a learning result predicting method based on LSTM. The framework of this method is shown in Fig. 1, which includes two parts: predicting model development and learning result predicting. The basic process of the predicting model development is as follows: 1) aggregate each learner's scores according to the defined schema to generate the final scores, and further divide the learners into two or three categories, such as success, fail and withdraw; 2) select the appropriate predicting indicators based on the existing researches and the learning behavior data

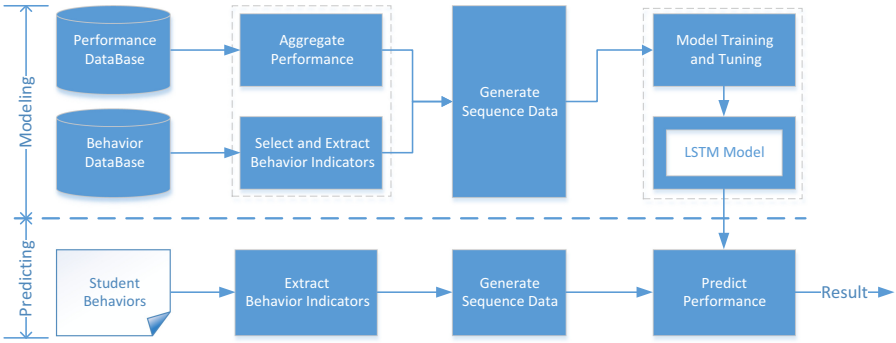

Fig. 1. The proposed framework for the detection of at-risk students

collected by the learning platform; 3) take one week as the period, extract the values of predicting indicators in each week from the raw learning behavior data, and generate a sequence data set; 4) train the LSTM-based predicting model using the back propagation algorithm and gradient descent algorithm. The basic process of learning result predicting is as follows: 1) extract the values of predicting indicators in each week from learners' raw learning behavior data to generate a sequence data; 2) input the sequence data into LSTM model to predict a learner's learning result.

3.1 Behavior Indicator Selection

From the previous literature review, we can see that researchers have explored many predicting indicators from many aspects. These predicting indicators mainly involve three aspects: first, the personal characteristics of learners, such as the gender, age, race, learning motivation, prior knowledge and so on; second, the results of process assessments, such as assignment scores, test scores, mid-term test scores and so on; third, the learning behaviors of learners, such as the frequency of login, number of browsed resources, number of posts and so on. As the goal of this study is to explore whether the dynamic pattern of learning behavior can be mined and improve the accuracy of learning result predicting, this study only considers the learning behavior indicators. It does not consider the personal characteristics and process assessment results.

In addition, because the data used in this study is from the Open University's Learning Analysis Dataset (OULAD) [8] when selecting the learning behavior indicators, we can only choose from the learning behaviors recorded in the dataset. OULAD is an open dataset produced by Kuzilek et al. of the Institute of Knowledge Media of the UK Open University, which records the detailed click behaviors and assessment scores of 22 courses [8]. All the 22 courses are deployed on the Moodle platform. However, the data in OULAD is not the raw Moodle log data, but the aggregate data. Kuzilek et al. divide the raw click behaviors of learners into 20 kinds of learning behaviors according

Table 1. Description of the nine behavior indicators

Activity type	Description
Resource	Usually contains pdf resources such as books
Oucontent	Represents content of assigments, which students should pass during presentation
Forumng	Disscussion forum
Url	Contains links to external or internal resources or for example video/audio content
Glossary	Consist of basic glossary related to content of course
Homepage	Course homepage
Subpage	Points to other sites in the course together with basic instructions
Oucollaborate	Online video discussion rooms (tutor - students)
Dataplus	Additional information/videos/audios/pdf

to the characteristics of the clicked objects [8]. OULAD recorded each learner's daily clicked objects, their frequency and the type of learning behaviors. Although OULAD contains 20 different types of learning behavior, not all courses contain these 20 types of learning behaviors. Because the course selected in this study only involves nine types of learning behaviors, these nine types of learning behaviors are selected as the predicting indicators. These nine learning behavior indicators mainly involve the use of learning resources, forums, assignments, glossary, homepage and other objects. See Table 1 for a detailed description.

3.2 Sequence Data Generation and Preprocessing

After selecting the predicting indicators, the value of each predicting indicator in a period can be calculated for training the predicting model. As mentioned above, most of the researches obtain values of predicting indicators from the accumulated data to train the predicting model. Different from these researches, this study calculates the value of each predicting indicator in different time periods, respectively, to generate the sequence data to train the LSTM model. Although every object clicked by each learner every day and its clicking frequency and learning behavior category are recorded in the OULAD, these data can't be directly used to train the LSTM model. They need to be transformed to generate the sequence data of each predicting indicator. The process of sequence data generation is shown in Fig. 2 below:

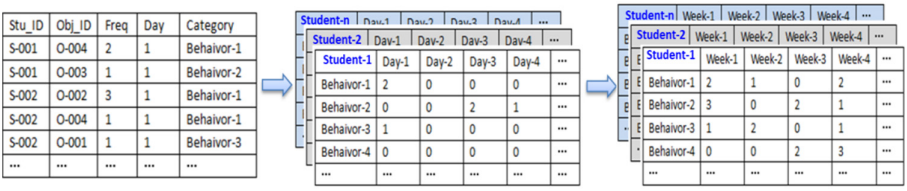

Fig. 2. Process of sequence data generation

Frist, according to the identification of learners, the behavior category, and the time when the clicking behavior occurs, we can calculate the frequency of every behavior indicator of every learner in every day. Second, these sequential data is further aggregated into the frequency of each behavior indicator of every learner in every week. The reason why the frequency of each behavior indicator is calculated by week is that there is a problem of data sparsity when calculating each behavior indicator by day. Some learning behaviors do not occur for several consecutive days while calculating by week can solve this problem to a certain extent.

After generating the sequence data of each behavior indicator, pre-processing is implemented. In OULAD, the learning results of learners are divided into four categories: pass, fail, withdraw and distinction. Because there are very few samples whose learning results are distinction, these samples are eliminated in the pre-processing stage. In this study, the final learning results of learners are divided into three categories: pass, fail and withdraw.

3.3 Prediction Modeling Based on LSTM

In order to use the dynamic characteristic of learning behavior to improve the accuracy of learning result predicting, we adopt the LSTM network to develop learning result predicting model. LSTM network is a special kind of RNN, which can make full use of not only the useful information close to the current position, but also the useful information far from the current position. The basic structure of the LSTM network is the same as that of the simple RNN, and the main difference is the internal structure of the recurrent unit. Different from the structure of the recurrent unit in simple RNN, the LSTM recurrent unit has a special structure with three "gates", which are usually called the input gate, forget gate and output gate. By these three gates, the LSTM selectively influences the state of the recurrent neural network in every moment. The structure of the recurrent unit in LSTM network is shown in Fig. 3:

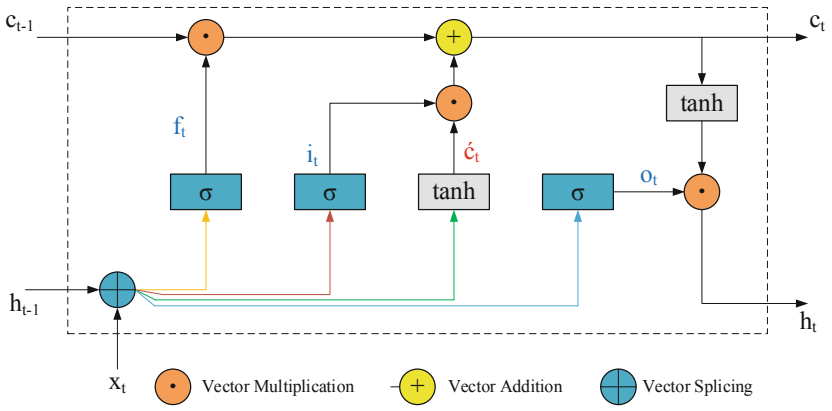

Fig. 3. Long Short Term Memory Network

In the above figure, c_t is the current state of the neural network, c_{t-1} is the state of the neural network last time, h_{t-1} is the output of neural network last time, $\acute{c}_t$ is the candidate state obtained by nonlinear function, x_t is the input of the current time, i_t is the input gate of the recurrent unit, $i_t \in [0, 1]^D$; f_t is the forget gate of the recurrent unit, $f_t \in [0, 1]^D$; o_t is the output gate of the recurrent unit, $o_t \in [0, 1]^D$. Forget gate f_t decides how much information of the state of last time c_{t-1} needs to be forgot. Input gate i_t decides how much information about the candidate state at the current time $\acute{c}_t$ needs to be saved. Output gate o_t decides how much information of the current state c_t needs to be passed to the output of current time h_t. When $f_t = 0$, $i_t = 1$, the recurrent unit clears the history information, and the candidate state vector $\acute{c}_t$ is written, the state of the neural network c_t is still related to the historical information of the previous moment. When $f_t = 1$, $i_t = 0$, the recurrent unit will copy the contents of the previous time without writing any new information.

LSTM calculates the state at the current time c_t and output h_t as follows:

1) Using the output of previous time h_{t-1} and the input at the current time x_t, three gates i_t, f_t, o_t are calculated. The calculation methods are shown in Formula 1, 2 and 3 respectively:

$$i_t = \sigma(W_i x_i + U_i h_{t-1} + b_i) \tag{1}$$

$$f_t = \sigma(W_f x_i + U_f h_{t-1} + b_f) \tag{2}$$

$$o_t = \sigma(W_o x_i + U_o h_{t-1} + b_o) \tag{3}$$

2) Using forget gate f_t and input gate i_t to update the current state c_t. The update method is shown in formula 4:

$$c_t = f_t \odot c_{t-1} + i_t \odot \acute{c}_t \tag{4}$$

By substituting Formula 1 and formula 2 into formula 4, we can further replace the calculation method of c_t, which is shown as formula 5:

$$\mathrm{c_t} = \sigma(\mathrm{W_f x_i} + \mathrm{U_f h_{t-1}} + \mathrm{b_f}) \odot \mathrm{c_{t-1}} + \sigma(\mathrm{W_i x_i} + \mathrm{U_i h_{t-1}} + \mathrm{b_i}) \odot \acute{\mathrm{c}}_\mathrm{t} \tag{5}$$

3) Combined output gate o_t, pass information of internal state to external state h_t, h_t is calculated as follows:

$$\mathrm{h_t} = \mathrm{o_t} \odot \tanh\ \mathrm{c_t} \tag{6}$$

Substituting formula 3 into formula 6, h_t can be further expressed as:

$$\mathrm{h_t} = \sigma(\mathrm{W_o x_i} + \mathrm{U_o h_{t-1}} + \mathrm{b_o}) \odot \tanh\ \mathrm{c_t} \tag{7}$$

According to the forward propagation process of LSTM network, this study takes the sequence data of behavior indicators from the first week to the n-th week as the input of the LSTM network. It uses the back propagation algorithm and the gradient descent algorithm to train the LSTM network. To dynamically predict learners' learning results and identify at-risk leaners, we should train an LSTM model for every week.

4 Experiment and Result

4.1 Dataset and Data Preprocessing

The data of this study comes from Open University (OU), which is one of the largest distance learning institutions in Europe. The OU modules are increasingly using the Virtual Learning Environment, Moodle, to provide learning materials, rather than the previous paper materials provided in the past. In 2017, Open University Learning Analytics Dataset (OULAD) was released. OULAD contains a subset of the OU student data from 2013 and 2014, including the information about 22 courses, 32,593 students, their assessment results, and logs of their interactions with the VLE represented by daily summaries of student clicks (10,655,280 entries). At present, there are two public datasets commonly used for learning behavior analysis and learning result prediction: KDD Cup

2010 dataset and KDD cup 2015 dataset, Compared with these two datasets, OULAD is quite different, which includes demographic data of learners and interaction data with the university's VLE.

In the experimental stage, this study selected the "AAA" course (code_module = "AAA") from October 2014 (code_presentation = "2014J"). The course lasts 269 days from the official start to the end (from date = 0 to date = 269), taking seven days as a week, 38 weeks and three days, plus four days(all kinds of behavior data are expressed as 0), a total of 39 weeks. During this period, the number of learners who chose to study this course was 365. Learning outcomes are divided into four categories, among which 299 are "Pass", 46 are "Fail", 66 are "Withdraw", and 24 are "Distinction". Because the number of "Distinction" is too small, the whole experimental data may be unbalanced, leading to the prediction effect. Excluding the category of "Distinction", the number of learners in experiment is 341, learning results are divided into three categories: Pass, Fail and Withdraw. There are 147653 learning records for 341 learners in the experiment. There are nine main behavior operations: dataplus, forumng, glossary, homepage, oucollaborate, oucontent, resource, subpage, url, the number of each operation is shown in Fig. 4.

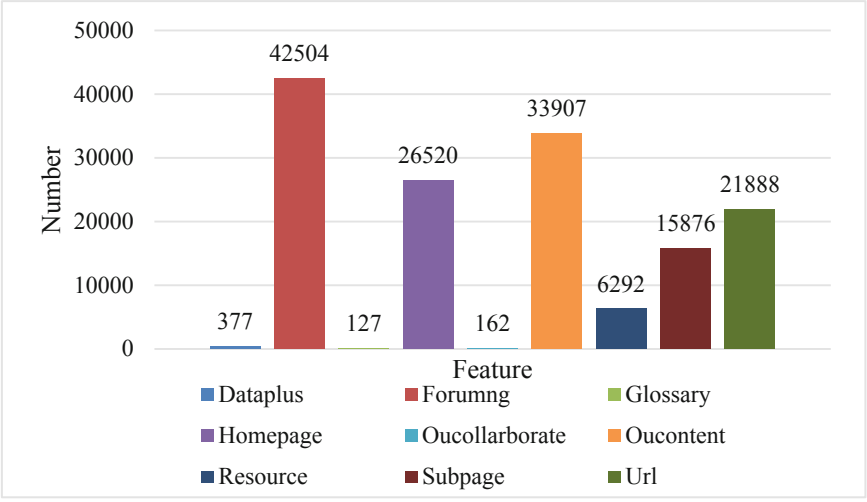

Fig. 4. Composition of experimental data

During the experiment, in order to ensure the validity of the experiment, the training set and test set are randomly assigned, according to 60% of the training set and 40% of the test set. Behavior data is processed in two ways: aggregate data and sequence data.

1) Aggregate data. The course contains 39 weeks, with five weeks as the time node, the first 5 weeks, the first 10 weeks, the first 15 weeks, the first 20 weeks, the first 25 weeks, the first 30 weeks, the first 35 weeks and the first 39 weeks as the units. The nine behavior categories in the time node segment are aggregated for statistics, and the data preprocessing results of each time node segment are 9 columns (categories).

Aggregate data is the aggregation statistics of 9 kinds of behavior data in a specified time period, reflecting the total number of each behavior operation in this time period.

2) Sequence data. Taking one week as the unit, the aggregation data of nine behaviors were counted. Then, the first 5 weeks, the first 10 weeks, the first 15 weeks, the first 20 weeks, the first 25 weeks, the first 30 weeks, the first 35 weeks and the first 39 weeks were taken as the time node, and the nine behaviors in the time node period were spliced and summarized by week. The data preprocessing results of each time node period were n * 9 columns (categories) (n = 5, 10, 15, …, 35, 39). Sequence data not only reflects the total amount of each behavior in a specified time period, but also can compare the number of behavior changes in different time periods after splicing the behavior data of adjacent time periods.

4.2 Implementation Details

This paper mainly solves two problems: Which is the greater influence of on learning result prediction, sequence data or aggregate data? Which model is the best model to predict the learning results of sequence data? Aiming at these two problems, the following two experiments are designed.

4.2.1 Comparison of Prediction Models on Aggregate Data and Sequence Data

According to the prediction models of learning result used in related research, Logistic Regression (LR), Navie Bayes (NB), Radom Forest (RF), K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Iterative Dichotomiser 3 (ID3), Classification and Regression Tress (CART), and Neural Network (NN) eight classic machine learning algorithms were selected, aggregate data and sequence data of the first 5 weeks, the first 10 weeks, the first 15 weeks, the first 20 weeks, the first 25 weeks, the first 30 weeks, the first 35 weeks and the first 39 weeks are respectively input into the models, the AUC of various prediction models were compared. LR model is generally used to solve the binary classification problem, because the learning results of this research are divided into pass, fail and withdraw, it belongs to multi classification problem, and one vs one (OVO) method is adopted, two categories are selected for comparison from the three categories, three comparisons are made, given a new sample, the probability corresponding to each category of the sample is calculated, and the prediction result of the new sample is the category with the highest probability; newton-cg algorithm is used to iteratively optimize the loss function by using the second derivative matrix of the loss function, i.e. Hessian matrix. NB model is based on GaussianNB classification algorithm, that is, the prior is Gaussian distribution of naive Bayes, the main parameter is prior probability, in the experiment, a priori probability P = mk/m, where m is the total number of training set samples, mk is the number of training set samples of the k class. RF model is a Meta estimator, which is composed of multiple decision trees, and each decision tree has no correlation. In the experiment, the number of decision trees in the forest is 10, and the entropy function of information gain is used to measure the performance of splitting quality. KNN model is a commonly used classification algorithm. If a sample is the most similar to K samples in the dataset, and most of the K samples belong to a certain category, the sample also belongs to a certain category. This model is related to the

initial K. In our research, we set K = 3, 4, 5, 6, respectively, to compare the AUC of the model. Experimental results show that the AUC of the learning result prediction model is the highest when k = 5. And according to the sample data, it can automatically get the appropriate algorithm from the ball_tree、kd_tree and brute algorithm. The SVM (SVC) classifier is selected, when the penalty parameter is set to 1.0, the penalty for misclassification increases, the kernel function is rbf, radial basis function determines the classification boundary according to the distance from each support vector, which can be mapped to infinite dimensions. ID3 model and CART model are classic algorithms of DT. In the experiment, ID3 model uses information entropy as the standard of feature selection, CART model uses gini coefficient as the standard of feature selection, and both models set splitter as best, which require to find the optimal dividing point in all the dividing points of features. NN is a kind of artificial intelligence machine learning technology, which simulates the human brain. This experiment uses the most classic three-layer neural network, including input layer, hidden layer and output layer. When using aggregate data to predict learning results, the input layer is 9, the bath_size is 30, the activation function is relu; the hidden layer is 6, the activation function is relu; the output layer is 3, and the activation function is softmax. When using sequence data to predict learning results, the input layer is 9 * n (n = 5, 10, 15, 20, 25, 30, 35, 39), the bath_size is 30, the activation function is relu; when the hidden layer is 6, the activation function is relu; when the output layer is 3, the activation function is softmax. The optimizer selects Adam, which is an adaptive learning rate method. It dynamically adjusts the learning rate of each parameter by using the first-order moment estimation and the second-order moment estimation of gradient. Each iterative learning rate has a clear range, which makes the parameter change very stable. The loss function was categorical_crossentropy, and the evaluation standard was accuracy. The number of iterations is determined by the current experimental model. According to experience, the number of iterations may be different when the input data changes weekly. Since the test set and training set are randomly assigned, the values of AUC predicted by each model may be different. Therefore, each model on the aggregate data and sequence data in different weeks are experimented for ten times, and the average value of predicted AUC is taken as the final prediction result on aggregate data or sequence data in this period of the model.

4.2.2 Prediction Model of Learning Results Based on Sequence Data

The best prediction models of learning results selected from the last experiment are compared with RNN model and LSTM model on sequence data, and the best prediction model of learning results is selected.

RNN is mainly used for the prediction on sequence data. The experimental data in this research is sequence data. Through experiments, RNN is compared with the best model in the previous experiment. The Keras framework is used in the experiment. The RNN model is constructed in three layers. The input layer is a three-dimensional vector: input_size × time_steps × cell_size, input_size is the length of data in each time period, that is, the number of features. In our research, input_size is the nine features extracted in the earlier stage; time_steps is the number of weeks, i.e. time_steps = 5, 10, 15, 20, 25, 30, 35, 39; cell_size is the number of neurons, which is set as 351 in the experiment; If the

data input model of the first 15 weeks is used for prediction, the input three-dimensional vector of the input layer is: $9 \times 15 \times 351$, the activation is relu; the units of the hidden layer are 351, the activation is relu; the output layer is output three classification, and the activation function is softmax. The model optimizer is Adam, corresponding to softmax classifier, and the model loss function is set to categorical_crossentropy, which is the logarithmic loss function of the multi classification. The criteria for model evaluation is accuracy. The same to the NN model, the number of iterations is determined by the current experimental model. The number of iterations may be different when the number of input data changes. According to the sequence data of different weeks, the model also tests ten times in each time period, and the average value of prediction AUC value is the ten times AUC values predicted by the learning results of the model.

RNN has a great advantage in processing sequence data. It can use the previous information to carry out corresponding operations on the current task, but if the location is far away, it can't be directly operated. LSTM is a special RNN model, which can solve the problem of "long dependence". In the experiment, a three-layer LSTM model is built by using Keras framework. The input layer of LSTM model is also a three-dimensional vector: input_size $\times$ time_steps $\times$ cell_size, the meaning and set of parameters in each dimension are the same as RNN model. Input_size is the length of data in each time period, that is, the number of features. Input_size is the nine features extracted in the earlier stage; time_steps is the number of weeks, including several weeks. In our research, five weeks is a time node, so time_steps = 5, 10, 15, 20, 25, 30, 35, 39; cell_size is the number of neurons, which is set as 351; the units in the hidden layer are 351, and activation is relu; the output layer is output three classification, and the activation function is softmax. The model optimizer is Adam, corresponding to softmax classifier, and the model loss function is set to categorical_crossentropy, which is the logarithmic loss function of the multi classification. The criteria for model evaluation is accuracy. Like RNN and NN models, the AUC value of LSTM model is also the average value of 10 times prediction results on sequence data in each time period.

4.3 Result and Discussion

4.3.1 Prediction of Learning Results on Accumulated Data and Sequence Data

Eight classic machine learning models, LR, NB, RF, KNN, SVM, ID3, CART and NN, are used to predict the learning results on aggregate data and sequence data, respectively. The prediction results are shown in Fig. 5. Where (a) represents aggregate data and (s) represents sequence data. For example, the prediction effect of LR model on aggregate data and sequence data are LR (a) and LR (s). It can be seen from the AUC of each learning result prediction model, the prediction AUC of LR model on aggregate data is 0.74 at the lowest and 0.80 at the highest, while that on sequence data is 0.68 at the lowest and 0.74 at the highest. KNN, SVM and LR are the same, the prediction results on aggregate data are better than that on sequence data. The prediction AUC of RF model on aggregate data is 0.76 at the lowest and 0.78 at the highest, while that on sequence data is 0.77 at the lowest and 0.83 at the highest. The prediction effect of RF on sequence data is better than that on aggregate data. The prediction AUC of NB, CART, ID3 and NN models on sequence data is higher than that on aggregate data. The experimental

results show that the prediction effect on sequence data is better than that on aggregate data, the RF model and CART model are better than other models.

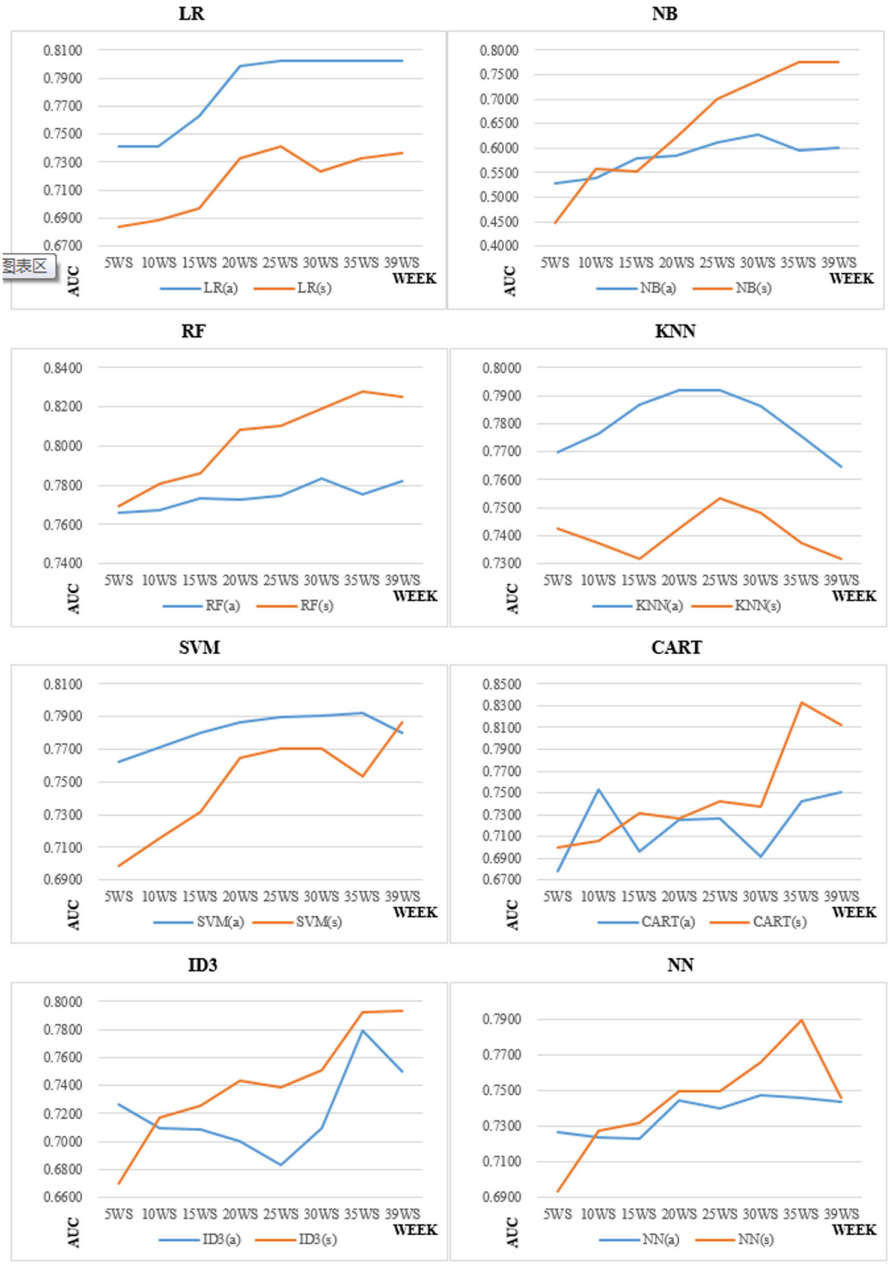

Fig. 5. Comparison of models in prediction of learning results

4.3.2 Prediction of Learning Results on Sequence Data

The experiment in the previous section proves that the prediction results on sequence data are much better than those on aggregate data. RF model and CART model with the highest average AUC value in each time period is selected as the representatives of machine learning models, and compared with RNN model and LSTM model; the results are shown in Fig. 6. The prediction results show that CART model is the worst of the four models. RF model based on the behavior data of the first ten weeks and the first 20 weeks is better than RNN model in the same time period, in other time periods, the RNN model is better. The prediction result of RNN model based on the behavior data of the first 35 weeks reaches the highest value of 0.85 of the four models. LSTM model has the best prediction effect of the four models, with the lowest AUC of 0.78 and the highest of 0.84. Generally speaking, the AUC values of the four models show a stable growth trend with the increase of the number of sequence data weeks until the first 35 weeks, and the predicted AUC based on the behavior data of the first 39 weeks shows a flat or even decline.

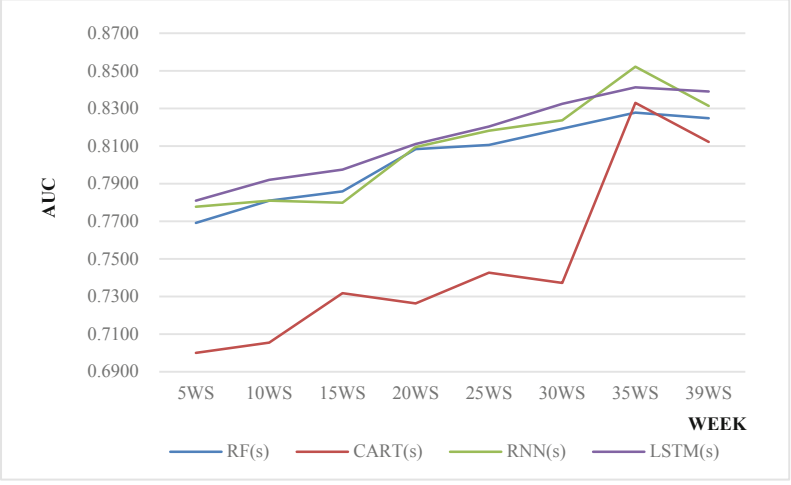

Fig. 6. Prediction of learning results of sequence data

5 Conclusion and Discussion

According to the related research of learning result prediction and the module of the learning platform, this paper divides the learning behaviors that affect learning results into three categories: human-computer interaction indicators, human-human interaction indicators and individual tendency indicators. A course of the Moodle platform is selected, and the nine most relevant learning behaviors are extracted. Taking one week as the basic unit and five weeks as the time node of learning behavior, the aggregate data and sequence data of the first 5 weeks, the first 10 weeks, the first 15 weeks, the

first 20 weeks, the first 25 weeks, the first 30 weeks, the first 35 weeks and the first 39 weeks are formed. Eight classic machine learning methods, i.e. LR, NB, KNN, RF, SVC, CART, ID3 and NN, are selected to predict the learning results in different time nodes based on aggregate data and sequence data. The experimental results show that the prediction effect of the NB model, RF model, CART model, ID3 model and NN model on sequence data is better than that on aggregate data on the whole, and the prediction effect of the LR model, KNN model and SVM model on aggregate data are better than that on sequence data. Generally speaking, sequence data is more effective for the prediction of learning results than aggregate data. Sequence data is not only the aggregation of behavior data in a fixed time period, but also the splicing of behavior data between adjacent time periods. It not only reflects the amount of behavior data between different time periods, but also reflects the change of the amount of row data, which is also the implicit indicator of learning results. The prediction AUC of RF model on sequence data is 0.77 at the lowest and 0.83 at the highest; the AUC of the CART model on sequence data is 0.70 at the lowest and 0.83 at the highest. RF model and CART model are the best models of the eight prediction models on sequence data.

RF model and CART model are representative models of eight classic machine learning methods, they and RNN model, LSTM model are used to predict learning results on sequence data, respectively. The experimental results show that the AUC of each model is the lowest in the first five weeks. Until the first 35 weeks, the prediction effects of four models have been steadily increasing, and the prediction effects of four models in the first 39 weeks are the same as before, or even decline. It is estimated that the number of learning behavior in the last four days of the 39th week are all 0, which is artificially added, as a complete week, resulting in changes between the data of the first 39 weeks and the data of the first 35 weeks affect the prediction effect of learning results. LSTM model is good at processing sequence data and solving "long dependence" well. The AUC of LSTM model is 0.78 at the lowest and 0.84 at the highest, and the AUC of LSTM model is the highest in all time periods, and the growth is very stable. The AUC of RNN model is 0.78 at the lowest, and 0.85 at the highest, which reaches the peak value of the four models. In the first 20 weeks of sequence data, the AUC of RF model is higher than that of RNN model, but the AUC of RNN model is higher than that of RF model in the following time periods. Comparing the predictive effects of the four learning results, LSTM is a model with the highest value of AUC and stable growth based on sequence data, and it is also the best model for predicting learning results in existing experiments.

In future research, we will mine the sequence of learning behavior types based on sequence data, and constantly improve the deep learning model of learning result prediction, to predict the learning result early and accurately.

Acknowledgements. This work is supported by the Fundamental Research Funds for Central Universi-ties (CCNU18JCK05), the National Science Foundation of China (No. 61532008; No. 61572223), the National Key Research and Development Program of China (No. 2017YFC0909502), and the Ministry of Education of Humanities and Social Science project (No. 20YJCZH046).

References

1. Deng, R.Q., Benckendorff, P., Gannaway, D.: Progress and new directions for teaching and learning in MOOCs. Comput. Educ. **129**, 48–60 (2019)
2. Jayaprakash, S.M., Moody, E.W., Lauría, J.M., Regan, R., Baron, J.D.: Early alert of academically at-risk students: an open source analytics initiative. J. Learn. Anal. **1**(1), 6–47 (2014)
3. Hu, Y.H., Lo, C.L., Shih, S.P.: Developing early warning systems to predict students' online learning performance. Comput. Hum. Behav. **36**, 469–478 (2014)
4. Chen, W.Y., Brinton, C.G., Cao, D., Mason-Singh, A., Lu, C., Chiang, M.: Early detection prediction of learning outcomes in online short-courses via learning behaviors. IEEE Trans. Learn. Technol. **12**(1), 44–58 (2018)
5. Tempelaara, D.T., Rienties, B., Giesbers, B.: In search for the most informative data for feedback generation: learning analytics in a data-rich context. Comput. Hum. Behav. **47**, 157–167 (2015)
6. Zacharis, N.Z.: A multivariate approach to predicting student outcomes in web-enabled blended learning courses. Internet High. Educ. **27**, 44–53 (2015)
7. You, J.W.: Identifying significant indicators using LMS data to predict course achievement in online learning. Internet High. Educ. **29**, 23–30 (2016)
8. Marbouti, F., Diefes-Dux, H.A., Madhavan, K.: Models for early prediction of at-risk students in a course using standards-based grading. Comput. Educ. **103**, 1–15 (2016)
9. Howarda, E., Meehana, M., Parnell, A.: Contrasting prediction methods for early warning systems at undergraduate level. Internet High. Educ. **37**, 66–75 (2018)
10. Pardo, A., Han, F., Ellis, R.A.: Combining university student self-regulated learning indicators and engagement with online learning events to predict academic performance. IEEE Trans. Learn. Technol. **10**(1), 82–92 (2017)
11. Gašević, D., Dawson, S., Rogersb, T., Gasevic, D.: Learning analytics should not promote one size fits all: the effects of instructional conditions in predicting academic success. Internet High. Educ. **28**, 68–84 (2016)
12. Fan, Y.Z., Wang, Q.: Prediction of academic performance and risk: a review of literature on predicative indicators in learning analytics. Distance Educ. China **1**, 05–15+44+79 (2018). (in Chinese)
13. Conijn, R., Snijders, C., Kleingeld, A., Matzat, U.: Predicting student performance from LMS data: a comparison of 17 blended courses using Moodle LMS. IEEE Trans. Learn. Technol. **10**(1), 17–29 (2017)
14. Moreno-Marcos, P.M., Alario-Hoyos, C., Muñoz-Merino, P.J., Kloos, C.D.: Prediction in MOOCs: a review and future research directions. IEEE Trans. Learn. Technol. **12**(3), 384–401 (2019)
15. Arnold, K.E., Hall, Y., Street, S.G., Lafayette, W., Pistilli, M.D.: Course signals at Purdue: using learning analytics to increase student success. In: Proceedings of the 2nd International Conference on Learning Analytics and Knowledge, pp. 267–270 (2012)
16. Kuzilek, J., Hlosta, M., Herrmannova, D., Zdrahal, Z., Wolf, A.: OU analyse: analysing at-risk students at The Open University. Learning Analytics Review, LAK15–1, 1–16 (2015)