Prediction of learning behavior characters of MOOC’s data based on time series analysis

Dezhi Sun¹, Tong Li¹, Fucheng You¹, Man Hu¹, Zhenyu Li¹ *
¹Beijing Institute of Graphic Communication
*alvis_lee@foxmail.com

Abstract. In recent years, big data is used in finance, education, biotechnology, and other fields. In the field of education, thanks to the overall characteristics of big data, it can be applied to understand the objective learning rules of students, to provide effective decision-making suggestions for educators. With the development of electronic communication and media technology, Massive Open Online Courses (MOOC), which contains a large number of user behaviors and information, has gradually become one of the mainstream ways of education. However, MOOC have also been criticized for their high dropout rate. Through the analysis of big data contained in MOOC real data set, it can help teachers master students' learning rules and learning behaviors in time, which is of great significance to improve the level of education and the dropout rate of MOOC. First, it is helpful for teachers to adjust the course content in time according to the learning status. Second, teachers can carry out benign interventions on students' learning behavior. In this paper, based on the analysis of the real MOOC data of students' learning behaviors, we find that different groups of students have the same learning behavior at the same learning period, and students' learning situations can be divided into three stages. Then ARIMA model is used to predict the learning behaviors of students, to further analyze the factors causing the difference in prediction accuracy, we propose an error analysis method based on MSE. Through the experimental observation, the method proposed in this paper can effectively find and predict the characteristics of students' learning behavior, to give suggestions for improving the quality of teaching.

1. Introduction
Massive Open Online Courses (MOOC) are free online courses that anybody can take, thanks to their openness and convenience, MOOC have gradually attracted more and more students and universities to participate in them. At the same time, MOOC has accumulated a lot of user behavior data. Online learning platforms, such as edX and Coursera, have attracted large numbers of learners, they have also been criticized for their high dropout rates. In recent years, many researchers have been committed to studying the factors that affect the learning efficiency of MOOC by analyzing user’s behaviors on MOOC. Christopher [1] predicted MOOC performance by Clickstream Data and Social Learning Networks. Jacob [2] came up with an automatic stop-out classifier that achieves higher response rates compared to traditional post-course surveys and may boost students' propensity to “come back” into the course. Tang [3] divided the user’s learning behavior into several distinct categories to help improve user retention. Xing [4] identified a small subset of at-risk students, to provide systematic insight for instructors so they may better provide targeted support for those students most in need of intervention. In the above paper, to intervene on the learning effect of MOOC, various factors
affecting the learning effect of MOOC are analyzed, but the possible influence of different time periods on students' learning behavior is neglected.

In the development process of time series analysis methods, the application of economy, finance, engineering, and other fields has always played an important role in promoting the development of time series analysis, each step of the development of time series analysis is inseparable from the application [5]. Yang [6] presents a novel method for predicting the evolution of a student’s grade in MOOC via Time Series Neural Networks. Tang [7] extracts some raw features from MOOC users' logs and applies the MOOC users' daily activities to predict dropout in MOOC using time series.

In this paper, based on the analysis of the real MOOC data of students' learning duration and the overall learning behavior, students' learning situation was divided into different stages and different groups of students have the same learning behavior at different times of the same learning period. Then Autoregressive Integrated Moving Average(ARIMA) model is used to predict the learning duration of students. Through experiments, the feasibility of ARIMA for predicting students' learning duration was proved, and the accuracy of the prediction results is related to the students' learning behaviors. Finally, the students are divided into four groups, and the corresponding suggestions are given according to their learning behavior and prediction results.

2. Data description and analysis

2.1. Data description

The results presented in this paper are based on the dataset KDDCUP 2015 [8], which uses big data to predict whether MOOCer will “Skip class”. In KDDCUP 2015, 200,905 pieces of student information were recorded, including the student's course selection information, the start and end date of the course, and the student's detailed daily learning behavior. We consider the number of days covering the beginning and end of a course as a learning period and calculate the daily learning duration of students in a learning period based on their detailed learning behaviors every day. The learning duration sequence T of a students s in a learning period is as follows: 

\[ T_s = \{ (d_0, t_0), (d_1, t_1), \ldots, (d_i, t_i) \} \]

\( d_i \) denotes student's learning duration t on day i. If no learning record of student s is detected on day i, the learning duration t of that day will be recorded as 0. Students who have no learning record in the entire learning period, that is, the array value is all 0, will be removed. In KDDCUP 2015, the course usually covers 30 days from the beginning to the end. Base on this, We use 30 days as a learning period for students to observe the effect of the increase of date on the learning duration in a learning period.

In order to observe the overall learning duration distribution of students on different days in the same learning period, we define a set of students S of size n: \( S_n = \{ s_0, s_1, \ldots, s_n \} \). Then we set up a series \( T_n \) to represent the average learning duration in a learning period for a group of students of size n: 

\[ T_n = \{ (d_0, t_{n0}), (d_1, t_{n1}), \ldots, (d_i, t_{ni}) \} \]

where \( n \) is the size of students and \( (d_i, t_{ni}) \) denotes the average learning duration \( t_{ni} \) of a group of size n on day i, and 

\[ t_i = [t_{i0}, t_{i1}, \ldots, t_{in}] \]

The average learning duration \( t_{ni} \) is defined as the standard deviation of n groups of data on day i:

\[
    t_{ni} = \sqrt{\frac{\sum_{k=0}^{n} (t_k - \bar{t})^2}{n}}
\]

The reason for using standard deviation rather than calculating the mean directly is that standard deviation is a better indicator of the degree of dispersion among individuals within the group, this helps us better understand how the data is distributed.
2.2. Data analysis

In order to further analyze the distribution of learning duration, random sampling was carried out in the data with sampling size $n = 1000$. A total of three groups of data were collected, and their distribution pattern is shown in Fig. 1 below:

Figure 1. When $n=1000$, the learning duration distribution of the three groups of random samples.

In Fig. 1, the three groups of randomly sampled data have a similar learning duration on the same day. This indicates that different groups of students have the same learning behaviors of the same learning period.

To further verify the relationship between the student groups, we extend the sample sizes to $n=2000$ and $n=5000$, their distribution pattern is shown in Fig. 2. In Fig. 2, the learning duration between different groups on the same day was closer, this means that the increase in the size of students will not change the distribution of students’ overall learning behavior. Therefore, the overall learning behavior of large-scale students can be expressed with the overall learning behavior of smaller-scale students.

Figure 2. Distribution of average learning duration in different sizes of students.

By analyzing the learning duration of different dates, students’ learning situation in the same learning period was divided into different stages. Then, we divide the learning period into three phases. In the first phase, which is generally the first week of the learning period, most students' learning duration decreases with the date, reaching a minimum around the sixth day. The second phase, in which the learning duration each day increases as the date went on, lasts about two weeks. In the final stage, the students' learning duration will produce larger mutation: First, the students' learning duration will occur explosive growth and reach the maximum, at about 20th day; Then the learning duration will gradually subside, at about the 25th day, it returns to normal; After that, until the end of the learning period, the learning duration of students has grown exponentially again.

Through the analysis of the changing trend in different stages of the learning period, it can provide some references for MOOC teachers and administrators. In general, we can take the learning duration
as a reference for students' learning enthusiasm, and a higher learning duration represents a higher learning enthusiasm. Then let us reanalyze the three stages: The first stage indicates that students are likely to be more enthusiastic about learning at the beginning of the course, but this enthusiasm gradually wanes during the first week. Therefore, if the teacher can provide better content at this stage, it may enhance the appeal of the course. And the sudden change in learning duration in the third stage suggests that students are more enthusiastic about learning at this time, perhaps because of academic pressure. Anyway, at this moment, the attention of the curriculum is the highest, and teachers can provide more important content at this moment to gain greater influence.

Since the overall learning behavior of large-scale students can be expressed with the overall learning behavior of smaller-scale students, for the convenience of calculation and presentation, we processed the data with a sample size of n=1000 based on the time series model.

3. Time series prediction

3.1. ARIMA model

The Box-Jenkins method [9] is a more perfect and accurate algorithm for analyzing and predicting time series data, the commonly used models include the Auto-regressive model (AR model), Auto-Regressive Moving Average Model (MA model), Auto-Regressive Moving Average Model (ARMA model), Auto-regressive Integrated Moving Average model (ARIMA model). In this paper, we chose ARIMA as our prediction model.

The ARIMA(p, d, q) model is an extension of the ARMA(p, q) model which can be expressed as:

\[(1 - \sum_{i=1}^{p} \phi_i L^i)(1 - L)^d X_t = (1 + \sum_{i=1}^{q} \theta_i L^i) \varepsilon_t\]  

Where L is Lag operator, \(d \in \mathbb{Z}, d > 0\).

3.2. Stationary test

The time series that can be analyzed and predicted by the ARIMA model must meet the condition of stationary non-white noise series. It is an important step for time series analysis to check the stationarity of data. If the original data does not meet the stationarity, the original data can be transformed into a stationary sequence by means of difference, log, and other data transformation methods for subsequent analysis. The stability of time series is generally tested by a time sequence diagram and correlation diagram, but this approach is highly subjective. Augmented Dickey-Fuller (ADF) test can be used to verify the stationary hypothesis of a time series. If the results of the test are statistically significant, the series can be considered to satisfy the stationary hypothesis. Therefore, we use the ADF test to judge whether the time series is stable.

For non-stationary time series, if there is an increase or decline trend, difference processing is needed and then a stationarity test is carried out until it is stationary. In theory, difference, the more the temporal information of non-stationary deterministic information extracting more fully, but in theory, the number of difference is not the more the better, every time difference operation, can cause the loss of information, so should avoid excessive difference, generally in the application, the difference of no more than two orders.

The T-test, also known as the Student's T-test, is used to test the difference between the means of two populations because the F distribution approximates a normal distribution as the degree of freedom approaches infinity, so the t-test is usually used to test the difference between the means of two normal distributions. T-statistic is the test statistic used by the t-test for hypothesis testing. P-value can be calculated through the value of t-statistic to judge whether the null hypothesis is rejected. P-value is a probability, which is the probability that the sample result will appear if the hypothesis is true. If the p-value is small, it means that under the premise that the null hypothesis is true, the probability of the sample result appearing is small or even extreme, which in turn indicates that the null hypothesis is very large and the probability is wrong. Typically, a significance level \(\alpha\) is set to
compare with the p-value, and if p-value < α, the null hypothesis is rejected under significance level α. α is normally set to 0.05 [10].

3.3. Model selection
The guiding idea of selecting the optimal model is to investigate from two aspects: one is to maximize the likelihood function, the other is to minimize the number of unknown parameters in the model. The greater the likelihood function value is, the better the fitting effect of the model is. However, we cannot simply measure the quality of the model by the fitting accuracy, which will lead to more and more unknown parameters in the model, and the model will become more and more complex, resulting in overfitting. Therefore, a good model should be a comprehensive optimization configuration of the fitting accuracy and the number of unknown parameters. Common model selection methods include AIC and BIC.

Akaike information criterion (AIC) [11], based on the concept of entropy, can balance the complexity of the estimated model with the goodness of the data fitted by the model which is a measure of Goodness of fit in a statistical model. In general, AIC can be expressed as:

$$AIC = 2K - 2\ln\left(\hat{L}\right)$$

Where K is the number of arguments and \( \hat{L} \) is the likelihood function.

Bayesian Information Criterion (BIC) [12] is to use subjective probability estimation for some unknown states under incomplete information, and then use Bayesian formula to modify the occurrence probability, and finally use the expected value and correction probability to make the optimal decision. BIC can be expressed as:

$$BIC = \ln\left(n\right)K - 2\ln\left(\hat{L}\right)$$

Where \( \hat{L} \) is the maximum likelihood function of the model.

AIC provides effective rules for model selection, but it also has some shortcomings. When the sample size is large, the fitting error of the information provided in the AIC criterion will be amplified by the sample size, the number of parameters of penalty factor but has nothing to do with sample size (2) has been, so when the sample size is large, using AIC criterion to select the model and real convergence, it is usually better than the real model of the number of unknown parameters. The proposal of AIC criterion and BIC criterion can effectively compensate the subjectivity of order determination based on autocorrelation graph and partial autocorrelation graph, and help us to find the relatively optimal fitting model within a limited range of order numbers.

4. Result and Analyse
4.1. Model application
In the purpose of the convenience of calculation and presentation, we randomly select a set of student sequences \( S_n \) of size n=1000, there is 911 valid data among them. And we calculated the mean time series \( T_n \) of \( S_n \). The autocorrelation graph of \( T_n \) is shown in Fig. 3. In Fig. 3, the X coordinate represents the order and the Y coordinate represents the autocorrelation. Generally, the order of T is judged by autocorrelation graph to select the model, through the artificial observation of the autocorrelation curve, the data contained in the confidence space [-0.5,0.5] is regarded as the effective data. But the value obtained by this method has a certain randomness, so we propose another method for model selection.
Figure 3. The auto-correlation graph of \( T_n \)

The ADF test values of \( T_n \) after difference are as follows:

| Diff | Test Statistic | p-value | Lags Used | Critical Value(5%) |
|------|---------------|---------|-----------|-------------------|
| diff 0 | -1.58626 | 0.490493 | 0         | -2.96788         |
| diff 1 | -3.21362 | 0.0192097 | 6         | -3.00543         |
| diff 2 | -3.6088 | 0.00559495 | 5         | -3.00543         |

It is easy to see from Table 1. that \( T_n \) is stable after a difference which p-value < \( \alpha \). This means that \( T_n \) reaches the stability standard after the first difference. We use the data diff1() after the first difference as the input of the model to make the next prediction.

Figure 4. Using the ARIMA model predict the learning duration of \( T_n \).

The next task is to select the model, the selection condition is to obtain the minimum loss value under AIC standard and BIC standard, and the optimal model [13]. For example, \( T_n \) obtain the minimum loss value when AIC = (0,1) and BIC = (0,1). In this case, we choose (0,1) as the (p, q) parameter of the ARIMA model, considering that \( T_n \) needs to go through a difference to become a stationary sequence, we take d = 1 as the last parameter of ARIMA. To sum up, we substituted p = 0, q = 1, and d = 1 into ARIMA (p, d, q) to get a model suitable fit \( T_n \): ARIMA (0, 1, 1), and predicted the time series for a given week, the result is shown in Fig.4.
In our example, the predicted results of $T_n$ in Date(16,23) where the prediction interval begins on day 16 and ends on day 23, are shown. From the prediction results, the ARIMA model can predict the time series and produces a relatively close result.

The feasibility of time series for the prediction of learning duration in MOOC is confirmed by the application of the ARIMA model, but the results from observations of time series alone are too subjective and not suitable for large-scale time series predictions. Therefore, MSE [14] is used to evaluate the effectiveness of the model.

4.2. Model evaluation
Mean Squared Error (MSE) is used to reflect the degree of difference between the predicted value and the actual value. Theoretically, the closer MSE is to 0, the closer the two data are. However, a smaller MSE may also represent a model overfit. MSE can be express as:

$$MSE(T) = E((T - \theta)^2)$$  

(5)

Where $(T - \theta)$ is the error which is the difference between the estimated value and the actual value. And MSE between the predicted sequence and the actual sequence can be expressed as:

$$MSE = \sum_{(i,j)} E((T - \theta)^2)$$

(6)

Where $(i, j)$ represents the starting and ending interval.

Using (6), the MSE between the predicted sequence and the real data of $T_n$ was calculated, $MSE(T_{(16, 23)}) = 97.539$, which means a good prediction result.

Figure 5. The distribution of MSE where label x represents the student label of different students and label y represents the value of MSE.

Then we take the time series $t_n$ of each student in sample $S_n$ as the input of the ARIMA model. Given prediction interval in DATE (15,30), select the applicable model, among the 911 valid data, there are 529 time series that can be matched to the appropriate model in ARIMA and we calculate the MSE for each time series, their distribution is shown in Fig. 5.

Based on the distribution of MSE, we divided the students into 3 groups: Group1 contains all students with an MSE greater than 13,000, which contains less data than the other groups but produces the greatest MSE; Group2 includes all students whose MSE is less than 3000, whose number is large and whose MSE is small; The EMS of Group3 is between Group1 and Group2, and it has a more obvious trend of change, which is the transition stage of Group1 and Group2.

For the remaining data, the ARIMA model cannot be used to predict their time series because they cannot meet the requirements of stationarity. In order to observe their distribution rule, we classify these data into a group and mark them as Group4.

The performance of different groups in a learning period is shown in the figure below:
In Fig. 6, we found that four groups of students have different learning behaviors, which means that the student’s learning behavior will have a serious impact on the predicted results.

In Group1, the learning duration of students study each day is variable, and their learning behavior tends to be more random. Learning duration and regularity of Students in Group2 are more stable, but it also meant that their study time would stabilize at a fixed level. The stability of Group3 is between Group1 and Group2. Intervention with this group of students may make more sense because they are in a transition phase. Group3 students will be inclined to change to Group2 if they receive benign intervention; if they receive bad intervention, their learning behavior may become irregular and thus change to Group1. For students in Group4, it can be found that their learning duration has an obvious upward trend, that is to say, their time series is unstable, which explains why the ARIMA model can not be used to predict their time series.

To sum up, we come to the conclusion: the accuracy of the prediction of students' time series is affected by their own learning behavior. The more regular students' learning is, the better the prediction result is; on the contrary, the more random their learning behavior is, the greater the deviation between the prediction result and the real data is.

5. Conclusions

In this paper, based on the analysis of the real MOOC data of students' learning duration and the overall learning behavior, we found that when the sample size of students is the same, different groups of students have the same learning behaviors of the same learning period. And for a sample of students of different sizes, the increase in the size of students will not change the distribution of students’ overall learning behavior, the overall learning behavior of large-scale students can be expressed with the overall learning behavior of smaller-scale students.

According to the distribution of learning duration on the date, students' learning situation was divided into different stages. In the first phase, most students' learning duration decreases with the date. In the second phase, the learning duration each day increases as the date went on. And in the third stage, the students' learning duration will produce a larger mutation. For these three different stages, we suggest that MOOC teachers adopt different strategies to intervene students.

Then Auto-regressive Integrated Moving Average (ARIMA) model is used to predict the learning duration of students. Through the judge of the stability of students’ learning duration and the choice of model, the feasibility of ARIMA for predicting students' learning duration was proved. Then through the calculation of the evaluation index MSE, we found the accuracy of the prediction results is related to the students' own learning behaviors. The more regular students' learning is, the better the prediction result is; on the contrary, the more random their learning behavior is, the greater the deviation between the prediction result and the real data is. Finally, the students are divided into four groups by prediction results of the model, and the corresponding suggestions are given according to their learning behavior and prediction results.
Through analyzing and predicting the time series of students' learning, it is helpful for MOOC teachers to master the students' learning situation in real-time and intervene with the students in a timely manner, so as to improve the learning effect.

Acknowledgments
This work was supported by Beijing Higher Education "Undergraduate Teaching Reform and Innovation Project" (202010015002), Education innovation key project of beijing institute of graphic communication (22150121002/004), and Bigc Key Teaching Reform Project (22150121034/029) initial funding for the Doctoral Program of BIGC(27170120003/022) , BIGC Project (Eb202004, Ec202007).

References
[1] Brinton, Christopher G., and Mung Chiang. "MOOC performance prediction via clickstream data and social learning networks." 2015 IEEE conference on computer communications (INFOCOM). IEEE, 2015.
[2] Whitehill, Jacob, et al. "Beyond prediction: First steps toward automatic intervention in MOOC student dropout." Available at SSRN 2611750 (2015).
[3] Feng, Wenzheng, Jie Tang, and Tracy Xiao Liu. "Understanding dropouts in MOOC." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 33. No. 01. 2019.
[4] Xing, Wanli, et al. "Temporal prediction of dropouts in MOOC: Reaching the low hanging fruit through stacking generalization." Computers in human behavior 58 (2016): 119-129.
[5] Zhou Coopers, Wang Huiqi, Lu Wangyong, et al.Time series analysis and application [M]. Higher Education Press, 2015.
[6] Yang, Tsung-Yen, et al. "Behavior-based grade prediction for MOOC via time series neural networks." IEEE Journal of Selected Topics in Signal Processing 11.5 (2017): 716-728.
[7] Tang, Cui, et al. "Time series model for predicting dropout in massive open online courses."International Conference on Artificial Intelligence in Education. Springer, Cham, 2018.
[8] http://www.kdd.org/kdd-cup
[9] Box, George EP, et al. Time series analysis: forecasting and control. John Wiley & Sons, 2015.
[10] Fisher R A. Statistical methods for research workers[J]. Statistical methods for research workers., 1934 (5th Ed).
[11] Akaike, Hirotugu. A new look at the statistical model identification: IEEE Transactions on Automatic Control , 1974
[12] Watanabe, Sumio. "A widely applicable Bayesian information criterion." Journal of Machine Learning Research 14.Mar (2013): 867-897.
[13] Burnham, Kenneth P., and David R. Anderson. "Multimodel inference: understanding AIC and BIC in model selection."Sociological methods & research 33.2 (2004): 261-304.
[14] Marmolin, Hans. "Subjective MSE measures."IEEE transactions on systems, man, and cybernetics 16.3 (1986): 486-489.