Research on Adaptive Learning Prediction Based on XAPI

Jun Xiao, Lamei Wang, Jisheng Zhao, and Aizhen Fu

Abstract—**In the field of online learning, there is a problem of high student turnover rate. How to accurately identify learners and provide targeted teaching support services is an urgent problem for education researchers. In this paper, 1306 online learners majoring in finance from Shanghai Open University were selected as the subjects, and two kinds of data sets are adopted, which are learning data of online learning platform and learning behavior data of students based on xAPI, to analyze the relationship between learners' various online learning behaviors and learning achievements, and to determine the characteristics related to learning state of learners, describe the personalized learning state portrait, and select a variety of machine learning algorithms to build prediction model based on two data sets, to explore which data is more effective for building prediction models to identify potential risk learners. It is found that data mining analysis based on xAPI data has higher prediction accuracy than traditional online learning data.**

Index Terms—**Data mining, xAPI, adaptive, learning prediction.**

I. Introduction

Now online learning groups are increasing, but there are problems such as high student turnover rate and low completion rate of courses. Distance education, e-learning, and other online learning behaviors show a very growing trend of data flow, and the ever-increasing online learning platform also stores many learning data. The application of data mining technology in the field of education provides a solution for the rational use and interpretation of these data, and the realization of personalized analysis and learning prediction of learners.

The purpose of the data mining method is to extract meaningful knowledge from data [1]. Its application in the field of education is called education data mining [2]. The ET L-EDM LA report of the U.S. Department of education defines education data mining (EDM) as the widespread use of statistics, machine learning algorithms, and data mining techniques to process and analyze education big data. Through modeling find the relationship between students' learning results and variables such as learning content, learning resources and teaching behavior, and then predict the future of students learning trends [3]. Data mining techniques relevant in education are prediction, clustering, relationship mining, discovery with models, and distillation of data for human judgment [4]. The application of data mining technology in online learning can effectively help students, teachers, curriculum developers and administrators under the network environment to establish an online learning mechanism, improve learning efficiency and teaching efficiency [5].

With the innovation of education concept, various online learning platform is gradually beginning to attach importance to the personalized teaching service. This paper starts from the application of education data mining, taking online learners majoring in finance of Shanghai Open University as the research object, based on the online learning platform learning data and xAPI data of students' learning behavior, analyzing the personalized learning behavior of learners, discusses the relationship between learning behavior and learning achievement, and describes the personalized learning state of learners by integrating various behavior characteristics. At the same time, the prediction models are built based on two kinds of data sets respectively, to explore the accuracy and effectiveness of data mining analysis based on xAPI and traditional data mining analysis. Through comparison, this paper studies which kind of data-based prediction model has higher prediction accuracy, which provides a basis for better construction of the learner model.

II. Related Work

A. Education Data Mining Technology

Romero and Ventura analyzed several applications of data mining in education from 1995 to 2005, including statistics and visualization, clustering, classification and outlier detection, association rule mining and pattern mining, and text mining [6]. At present, the research direction of education data mining mainly focuses on the construction of student models, among which a large number of education researchers have built a learning prediction model based on online learning behavior log data. For example, Talavera and Gaudioso proposed using clustering to mine student data to find patterns that reflect user behavior [7]. Lykourentzou *et al.*, used three different data mining techniques, namely neural network, support vector machine, and probability integration simplified fuzzy ARTMAP (fuzzy logic and adaptive response theory map) to predict dropout in their online learning courses [8]. Macfadyen *et al.*, conducted regression analysis based on the online learning data of the BlackBoard platform to study the impact of different learning process data on the final learning performance, including online time, online link access, number of posts, etc., they established a prediction model [9]. Huseyin *et al.*, used decision tree algorithm to predict the factors that affected students' academic success and

Manuscript received January 20, 2020; revised April 12, 2020.

Jun Xiao and Lamei Wang are with the Shanghai Engineering Research Center of Open Distance Education, Shanghai Open University, China (e-mail: ecnuxj2003@163.com, wlmyisa@163.com).

Jisheng Zhao is with the Foda intelligence Inc., China (e-mail: zhaojisheng@foda.ai).

Aizhen Fu is with the East China Normal University, China (e-mail: aienfu@163.com).

doi: 10.18178/ijiet.2020.10.9.1443

constructed an education data mining system using a variety of model views [10].

B. Machine Learning Algorithm

Machine learning has widely used in different analytical applications, including the scenarios in the education area. The standard machine learning techniques mainly divided into two categories based on the data labeling: supervised and unsupervised. Based on the recent progress of transfer learning, semi-supervised learning is also getting more importance regarding the restriction of the size of the data set.

For education applications, researchers used to employ classification algorithms to distinguish the users with various properties. Standard classification algorithms include the support vector machine (SVM), naive Bayesian network, Bayesian belief network, random forest (RF), k-nearest neighbor classification (KNN).

C. xAPI

The American organization "ADL" (Advanced Distributed Learning) has issued Experience API (xAPI) version 1.0.0 [11]. It is a technical specification for storing and recording learning behavior data [12]. xAPI enables learning records to get rid of the association with devices or learning platforms, collecting and recording learners' online or offline learning experiences in different learning activities. These multi-modal learning data come from a variety of learning environments, such as formal learning, social media, and web-based or video-based informal learning [13].

xAPI uses "Activity Stream" to describe the learning experience [14], which mainly includes three elements: actor, verbs and activities related to the learning experience. xAPI specification can used to represent the interaction sequence, so it is widely used in Education [15], and very suitable to be integrated into the virtual learning environment (LMS) model as a component of log learning analysis. In the LMS, xAPI can separate or add learning event data from LMS so that learning content can be analyzed and identified in different systems [16]. Some educational researchers use xAPI specification to record student behavior logs, especially informal learning behaviors, such as Yee-King M J, Grierson M, etc. according to xAPI specification, record programmer learning behavior log, for example, to view all the actions of the student in a course, to analyze the learning behavior [17].

The learning data provided by xAPI can mine and analyze the learning behaviors of learners in multiple dimensions, and then draw the learning portraits of learners in multiple dimensions. Therefore, compared with the traditional mining form of learning platform log data, this paper also attempts to use the xAPI specification as a scheme to achieve the acquisition of education big data, and on this basis, explore the relationship between learning behavior and learning achievement.

III. Methods

A. Participants

Shanghai Open University, supported by modern information technology, is a modern open university that can provide not only academic education, but also vocational training and leisure culture education. At present, there are 80000 students in the school. The research object in this paper is 1306 online learners of a finance major at Shanghai Open University, including 494 male students and 812 female students. The learning behavior records adopted attendance records, learning duration, forum comments, learning materials downloading, video viewing documents, learning score records, etc. There are two different ways to obtain the data: on the one hand, based on the learning data of the online learning platform, on the other hand, based on the learning behavior data provided by the xAPI.

B. Dataset Description

The learning data of the online learning platform comes from the database system of Shanghai Open University, which mainly includes the basic information of students, the data of students' access to teaching resources, and some examination information of students. The specific analysis of students includes age, gender, course selection, resource utilization, access activity of teaching resources, and test score grading as the criteria of students' learning status Quasi equal. The specific sample data table includes student basic information data table, resource access log table, student online learning time table, test score table, and other data table items. Among them, there are 1306 non- repetitive students, more than 270000 access and browsing log data, more than 60000 learning resource browsing records, and more than 30000 student score records. After removing the data of students who not related or in the database, the primary data table, including browsing records, resource records, student information, student scores and so on sorted out. The table shows the classification and relationship between teaching resources and student users, as well as the relationship between them. At the same time, the single student record statistical information data table obtained, the student statistical information table is divided into model training data set and test data set, which provides the data source for the later prediction model construction.

The xAPI date provides more than 4W behavioral records of 8 learning activities include: registered, accessed, submitted, downloaded, watched, posted, studied and scored. According to the data statistics, each subject in the data set has different operations corresponding to different objects. Based on the specific times and frequency of these different operations, the information that only contains various behavior records of a single subject obtained. The essential data exploration and the impact of related eigenvalues are carried out.

Based on the learning behaviors obtained from the above two data sets, exploring the relationship between learning behaviors and students' learning results, and the key factors affecting learners' learning failure are determined as the input characteristics to describe learners' personalized portrait labels and model construction.

C. Data Analysis

Counting the distribution of students' academic performance based on the two data sets, as shown in Fig. 1, and Fig. 2. According to the students' academic performance carries on the primary stage judgment to the student, which is used to mark the types of students' learning status. It is an important index to predict the learning state of students and the basis of data analysis.

As shown in Fig. 1, the learning achievement distribution based on the online learning platform data meets normal distribution, indicating that the delivery of students' learning achievement is relatively ideal, and it is reasonable to use learning achievement as the status mark of students. The results divided into three stages: 90 and above is good, 70 and below is bad, others are normal. In turn, our research pays attention to the four characteristics of students' gender, age, access frequency, and online learning duration, and then get the influence of the four features on the performance distribution through statistics of students' performance information.

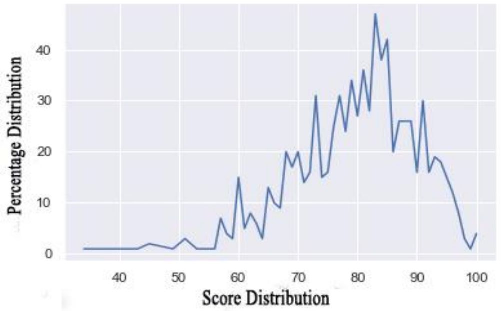

Fig. 1. Distribution of learning achievement based on online learning platform data.

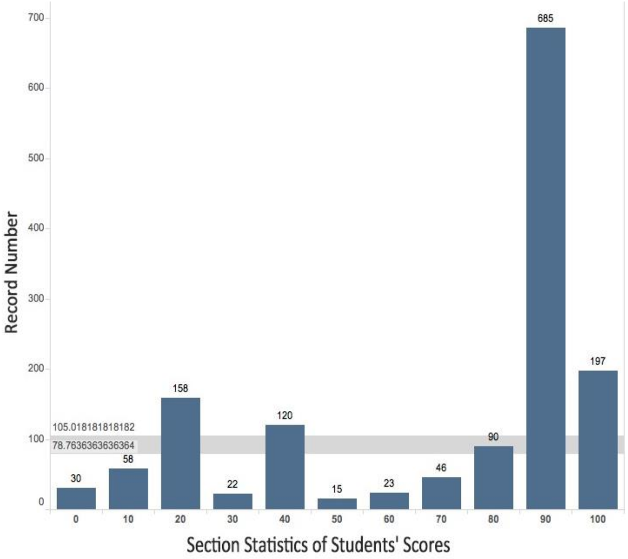

Fig. 2. Distribution of learning achievement based on xAPI data.

It is classifying and defining the actual state of students' learning by score. From Fig. 2, it is found that the distribution of students' learning performance based on xAPI data is not uniform. Because 686 invalid students with an average score of -1 are excluded, the delivery of students' performance is a structure with more at both ends and less in the middle. For students with a grade of -1, it is reserved as a prediction set. The real data set used for data analysis is the student achievement distribution shown in Fig. 2. According to the more detailed score distribution, score less than 60 points, defined as "dangerous" state, marked as 0, score higher than 60, and less than 85 points is defined as "good" state, marked as 1, score greater than 85 points is defined as "positive" state, marked as 2.

Then our research calculates the influence distribution between each learning behavior and learning achievement in the two data sets to determine which learning behavior has an essential impact on the learning state of learners as a feature label to describe the personalized learning state of learners.

IV. Result

A. Data Analysis Results

The first step of data mining and algorithm research is to find out the potential association between data and determine which factors affect the state of students' learning results based on the statistics of the influence distribution relationship between the learning behaviors and students' scores in the two data sets. The exploration results of learning data sampling data set based on student online learning platform are shown as follows.

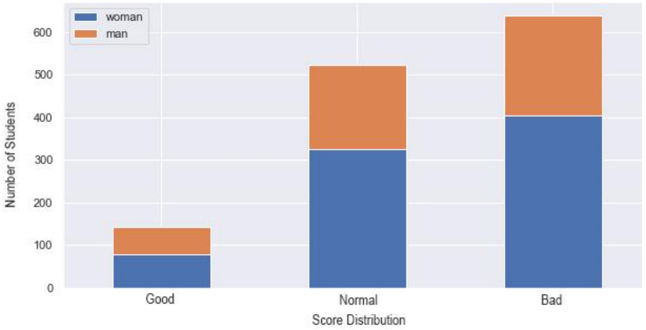

Fig. 3. The influence of students' gender on their performance.

It can be seen from Fig. 3 that in different stages of student performance, the percentage of male and female students is the same and the percentage of male and female students in the learning stage is almost the same, indicating that gender has little impact on learning performance.

The age distribution information of the students is shown in Fig. 4. The age distribution is divided into three stages: marked as 2 for 45 years old and above, 0 for 28 years old and below, and 1 for others. The influence of the age of the students on their academic performance is calculated, as shown in Fig. 5.

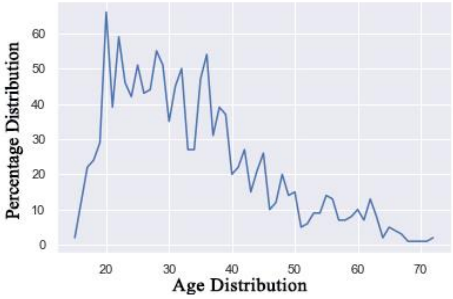

Fig. 4. The age distribution of students.

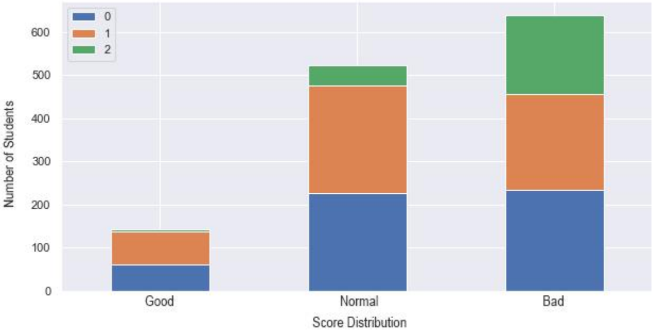

Fig. 5. The influence of students' age on their academic performance.

From Fig. 5, it can be seen that the age of students has a particular impact on the performance distribution of students.

The older the students are, the less likely they are to achieve good results. In feature selection, age is an important feature.

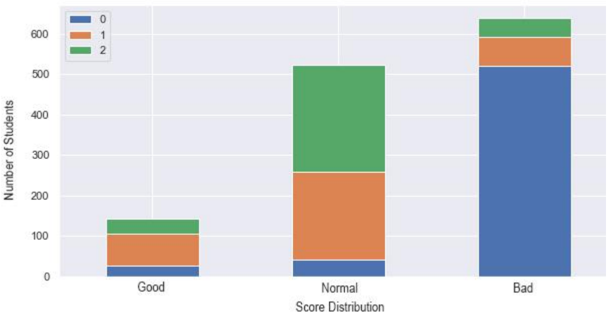

Fig. 6. Impact of interview records on score distribution.

The students divided into three categories according to the number of interview records. 0 means no visit, 1 means average frequency of visit, and 2 means the high rate of visit. It found from Fig. 6 that in bad students, the percentage of students who are not visited is more than 70%, while in good students, the percentage of students who are basically not visited is less than 10%. Therefore, the frequency of website visit has a positive impact on the distribution of students' performance, and the visit record will become an essential feature of students' learning result status portrait.

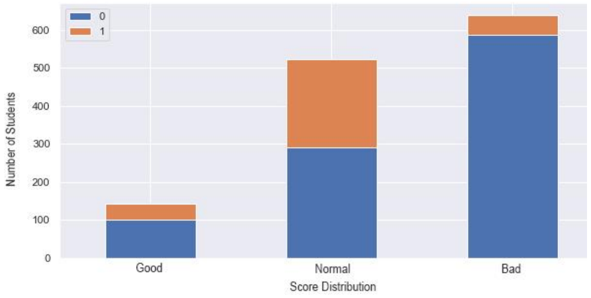

Fig. 7. The distribution of online learning hours to scores.

The students divided into two categories according to the length of study. 0 indicates that the study time is average, 1 indicates that the duration is longer. As shown in Fig. 7, in bad students, the percentage of students in general learning time is more than 80%, while in good students, the percentage of students in general learning time is less than 70%. It concluded that the length of learning time has a particular impact on the distribution of students' performance.

To sum up, in the learning data collection based on the online learning platform for students, age, website access frequency and learning duration can be used as the input characteristics of the model.

The exploration results of the data set provided by the xAPI shown in Fig. 8, it can be seen that the average value of students and scores in each segment is constant, which shows that these two behaviors have no impact on the ratings. Secondly, the distribution of accessed, downloaded and watched is chaotic and irregular, and the order of the maximum and minimum values is not consistent with the direction of the abscissa axis, and there is no apparent correlation. Finally, there is a positive linear correlation between submitted, registered and posted behaviors, and students' academic performance to a certain extent, indicating that the number of these three behaviors will affect their academic performance, which can be used as the input characteristics of the model.

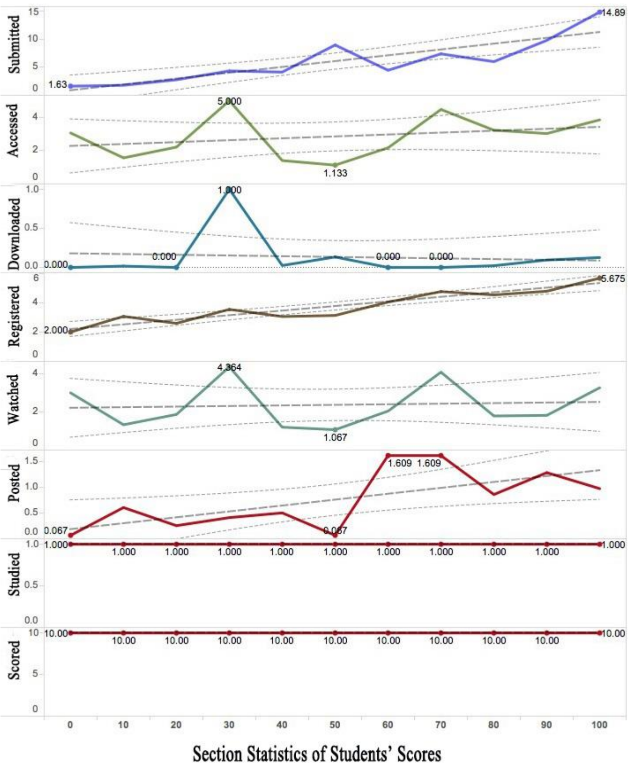

Fig. 8. Distribution of learning behaviors and achievements of students.

B. Algorithm Model Construction and Prediction Results

Through the exploration of the data, our research can roughly understand the influence relationship between the data. Through the understanding of the design goal, this paper determines the research goal of classifying the learning state of learners based on the test data set, that is, to achieve it through the learners' learning state classifier. Because the data uses the data with result mark, and the algorithm needs to be implemented by the classifier, it needs to use a supervised learning classification algorithm. Considering that there is an absolute correlation between the dimensions of the data set, such as login registration behavior will affect the occurrence of other practices such as browsing, submission, etc., so the Bayesian algorithm is not effective in this test set and will not use. The two datasets have different features that need two separate machine learning models to make the best utilization of data. Therefore, the critical features obtained from the above analysis results used as the input data of the model. A variety of classification prediction algorithms implemented on two test sets. The prediction results of the algorithm model are as follows:

According to the data characteristics of the test data set, our research uses the KNN, SVM, and RF algorithm to predict the learning state of students in the sample data set of the learning platform. The data set is randomly divided into 8:2 training data set and test data set, and the model training is conducted on the training data set, and the prediction test is conducted on the test data set. Then compared the predicted learning state results with the actual learning state results of the test data set, and the classification prediction accuracy of the three algorithms shown in Table I:

TABLE I. CLASSIFICATION PREDICTION ACCURACY BASED ON PLATFORM SAMPLE DATA SET

Algorithm Model	KNN	SVM	RF
Accuracy	72.8%	73.6%	76.2%

From the Table I, it can be seen that the test accuracy of the three models for the test results is more than 70%, and the effect is very close, among which the RF random forest algorithm has the highest accuracy.

For the test data set based on xAPI, decision tree (DT) and random forest (RF) algorithm models used. The model also divides the data into 8:2 training data set and test data set, and does model training on the training set, then uses the training model to predict， and test the test data set and compares the real learning state to get the prediction accuracy as shown in Table II:

TABLE II. CLASSIFICATION PREDICTION ACCURACY BASED ON XAPI DATA SET

Algorithm Model	RF	DT
Accuracy	81.9%	82.5%

It can be seen from Table II that the accuracy of the test results of these two models for the xAPI data set is more than 80%, and the effect difference is very close, among which the accuracy of the DT algorithm model is higher than RF algorithm model.

Comparing the model accuracy of learning platform data set with that of xAPI data set, it is found that the classification accuracy of xAPI data set is at least 5 percentage points higher than that of learning platform data set, which shows that students' behavior data based on xAPI data set is more accurate for the description of students' learning results.

V. DISCUSSION AND CONCLUSION

The traditional learning prediction model mainly based on online learning data, which has limitations for analyzing learners' learning behavior in all aspects. Compared with the conventional online learning data, the online learning represented by xAPI learning data has the characteristics of flexibility. It can share the follow-up data between different systems. Using this standard data can track all aspects of learning behavior [18], and analysis learners' learning behaviors in multiple dimensions, to obtain learners' learning portraits in various dimensions. Therefore, in order to more accurately identify the learning state of learners and meet the personalized learning characteristics and needs of learners. This paper adopts two kinds of data sets: learning data from online learning platform and learning behavior data based on xAPI, and compares the prediction accuracy of the prediction model based on the two kinds of data sets.

First, our research explored the influence distribution of learning behaviors on students' learning achievement, and the characteristics of learning portraits used to build the prediction model are preliminarily determined. Among them, it found that the age, learning time and access frequency of learners have a direct impact on learning achievement in the data based on the online learning platform, and the behavior of submitted, posted and registered has a more direct impact on learning achievement in the data set based on xAPI. These characteristics are used as the data input of the model to build a learning risk prediction model.

Through the comparison of the results of two data sets, our research found the accuracy of data mining analysis based on xAPI data is higher than that of traditional data mining, which shows that the students' behavior data based on xAPI data set is more accurate for the description of students' learning state, this provides new research ideas for researchers who explore the relationship between learners' learning behavior and learning outcome state.

Based on the above research content, this paper also developed a set of graphical learning state prediction platform. Through the platform, our research can realize the personalized learning state prediction of learners. The student portrait provides specific portrait content of a single student, and include the prediction results of the algorithm model. By presenting similar traffic with different states for students according to the warning signal of the signal light, students divided into three groups: red, yellow, and green. Green means that students are likely to achieve their goals if they continue to maintain their current learning state. Yellow indicates that the student has potential hazards in a course. Red means that the student union will be suspended. Through this function, students' status recognition and risk warning realized.

According to the prediction results, the adaptive early warning intervention system provides information push for the students who need to intervene, combined with learning habits. Push content includes personalized learning material recommendation, personalized learning method recommendation, customized chemical industry status assessment report, guiding opinions, etc.

Despite the promising results, there are some limitations. Firstly, In terms of data feature collection, only the influence of frequency on the results is considered, the influence of different behavior durations, and results on the prediction of learning state is not considered. It is necessary to comprehensively examine the addition of this part of data, continuously explore the influence of these data on the prediction of learning results, to make the performance of various features more characteristic and rich. Secondly, in terms of the collection of xAPI-based data, there are only eight kinds of behavior data, compared with xAPI's nearly 40 types of learning behavior data, there is only about one-fifth of the coverage. There is a one-sided and insufficient expression of the characteristics of the learners, so it is necessary to strengthen the coverage of more learning behavior data.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

Jun Xiao did the system design and system analysis, Lamei Wang did the data analysis, Jisheng Zhao did the algorithm support, Aizhen Fu did the literature analysis.; all authors had approved the final version.

ACKNOWLEDGMENT

This paper is supported by China's National General Project granted by China National Office for Education Sciences Planning (GRANT NO. BCA160053): The construction and application of online learners' persona based on big data analysis.

REFERENCES

[1] H. Jiawei. and K. Micheline, "Data mining: Concepts and techniques," *Data Mining Concepts Models Methods & Algorithms Second Edition*, vol. 5, no. 4, pp. 1-18, 2006.

[2] R. S. J. D. Baker and K. Yacef, "The state of educational data mining in 2009: A review and future visions," *Journal of Educational Data Mining*, vol. 1, no. 1, 2009.

[3] M. Bienkowski, M. Feng, and B. Means, "Enhancing teaching and learning through educational data mining and learning analytics: An issue brief," Washington: US Department of Education, pp. 1-57, 2012.

[4] P. Nithya, B. Umamaheswari. and A. Umadevi, "A survey on educational data mining in feld of education," *International Journal of Advanced Research in Computer Engineering & Technology*, vol. 5, no. 1, pp. 69-78, 2016.

[5] L. Huang. and G. Liu, "Application of web data mining in on-line education," in *Proc. the 2012 Second International Conference on Electric Technology and Civil Engineering*, 2012.

[6] C. Romero and S. Ventura, "Educational data mining: A survey from 1995 to 2005," *Expert Systems with Applications*, vol. 33, no. 1, pp. 135-146, 2007.

[7] L. Talavera and E. Gaudioso, "Mining student data to characterize similar behavior groups in unstructured collaboration spaces," presented at Workshop on Artificial Intelligence in Computer Supported Collaborative Learning at European Conference on Artificial Intelligence, 2004.

[8] I. Lykourentzou, I. Giannoukos, V. Nikolopoulos, G. Mpardis. and V. Loumos, "Dropout prediction in e-learning courses through the combination of machine learning techniques," *Computers & Education*, vol. 53, no. 3, pp. 1-965, 2009.

[9] L. P. Macfadyen and S. Dawson, "Mining LMS data to develop an 'early warning system' for educators: a proof of concept," *Computers & Education*, vol. 54, no. 2, pp. 1-599, 2010.

[10] H. Guruler and A. Istanbullu, *Modeling Student Performance in Higher Education Using Data Mining*, 2014.

[11] Advanced Distributed Learning. (2013). Webinar on the Experience API Version 1.0.0. [EB/OL]. [Online]. Available: http://www.adlnet.gov/webinar-on-the-experience-api-version-1-0-0/

[12] Advanced Distributed Learning. (2014). xAPI Background and History [EB/OL]. [Online]. Available: http://www.adlnet.gov/tla/experience-api/background-and-history/

[13] T. Rabelo, M. Lama, R. R. Amorim *et al.*, "SmartLAK: A big data architecture for supporting learning analytics services," in *Proc. Frontiers in Education Conference (FIE)*, 2010, pp. 1-5.

[14] N. Hruska, (2012). From ADL Team Member Nikolaus Hruska: Using Activity Streams in Next Generation SCORM [EB/OL]. [Online]. Available: http://www.adlnet.a-using-activity-streams-in-next-generation-scorm/

[15] A. R. Cano, M. B. Fernández, J. Álvaro, and T. García, "Using game learning analytics for validating the design of a learning game for adults with intellectual disabilities," *British Journal of Educational Technology*, vol. 49, no. 4, pp. 659-672, 2018.

[16] J. M. Kevan and P. R. Ryan, "Experience API: Flexible, decentralized and activity-centric data collection," *Technology Knowledge & Learning*, vol. 21, no. 1, pp. 143-149, 2016.

[17] M. J. Yee-King, M. Grierson *et al.*, "Evidencing the value of inquiry based, constructionist learning for student coders," *International Journal of Engineering Pedagogy*, vol. 7, no. 3, pp. 109-129, 2017.

[18] C. S. González and A. César, "Collazos, Roberto García. Desafíoen el diseñode MOOCs: Incorporación de aspectos para la colaboración y lagamificación," *Red Revista De Educación A Distancia*, vol. 48, no. 7, pp. 1-23, 2016.

Jun Xiao is a professor at Shanghai Engineering Research Center of Open Distance Education, Shanghai Open University, China. He is the member of China E-learning Technology Standardization Committee and Shanghai Dawn Scholar. His research focused on learning analytics, learner persona, augmented reality, virtual reality, artificial intelligence, etc. He is committed to using new technologies to improve the teaching effect of teachers and the learning effect of learners.

Lamei Wang is currently a research associate at Shanghai Engineering Research Center of Open Distance Education, Shanghai Open University, China. She is Hong Kong Croucher Foundation Visitorship for PRC Scholars 2017/2018. She got her master degree of computer applied technology at East China Normal Univercity in 2010. Her research interests gravitate towards learning analytics and learning system design.

Jisheng Zhao is the chief technical officer in Foda intelligence Inc., where he builds artificial intelligence based applications for education area. He earned his doctor degree from University of Manchester in 2007. His research work focuses on machine learning techniques in natural language process and large scale data analysis. He has published more than 50 research papers on various venus.

Aizhen Fu is currently studying in the first year of master's degree in the Department of Educational Information Technology at East China Normal University. Her research direction is learning analytics.