Educational Innovations and Applications- Tijus, Meen, Chang
ISBN: 978-981-14-2064-1

Predicting Learning Outcomes with MOOCs Clickstreams

Chen-Hsiang Yu[1], Jungpin Wu[2], Aa-Chi Liu[1]

[1]Department of Information Engineering and Computer Science
[2]Department of Statictics
Feng Chia University
#100 Wenhwa Road, Taichung, Taiwan
Phone: 886-4-2451-7250 Fax: 886-4-2451-6101
chyu@fcu.edu.tw, jungpinwu@cob.fcu.edu.tw, acliu@fcu.edu.tw

Abstract

Massive Open Online Courses (MOOCs) have gradually become one of the dominant trends in education. Since 2014, the Ministry of Education in Taiwan has been promoting MOOCs programs with successful results. Due to its self-paced mode, however, the low completion rate of MOOCs has recently become the focus of attention. The mechanism to effectively improve the course completion rate continues to be of great interest to both teachers and researchers.

In this study, we generated a sequence of learning behaviors of learners by using their video clickstream records on the MOOCs platform to find patterns in the learners' cognitive participation. Then, we built practical machine learning models using K-Nearest Neighbor, Support Vector Machine, and Artificial Neural Network algorithms to predict learning performance through student learning behavior. Using these models, we were able to determine the relevance of video viewing behavior to learning outcomes in order to assist teachers in helping learners who need additional support to pass the course.

Key words: MOOCs, Clickstream, Behavior Pattern, Machine Learning

Introduction

The rapid development of information technology has brought a huge influence on education, and how to apply technology to maximize learning outcomes has always been a topic for scholars to discuss. Massive Open Online Courses (MOOCs) [1] has aroused a boom in educational circles, and currently, there are many mature MOOCs learning platforms, including the US Coursera [2], Udacity, edX [3], Australia's Open2Study, UK's Futurelearn, etc. MOOC refers to the establishment of learning communities through unrestricted participation and readily available online courses. Its characteristics include open sharing, expandability, open authorization of content, open architecture, and learning objectives

Some students are easily distracted in traditional classrooms, which leads to a lot of time spent on review and homework after returning home. MOOCs are different from traditional teaching courses. Students can play back content if they do not understand the course. MOOCs provide online peer assistance for learners and opportunities to interact online with course teachers. Compared with the previous form of online education, MOOCs are closer to personalized learning; there is no teacher supervision, no entry threshold, and no need to pay expensive fees. As MOOCs facilitate self-regulated and individualized learning, in order to enable learners to achieve better learning results, many studies now focus on analyzing the learning history records left by users of MOOCs [4], in order to predict students' possible achievements through analytical methods, and then, provide early guidance to students who need help.

In actual operation, MOOCs courses are mainly based on video viewing, which takes most of the time for learners; however, many problems have been gradually found. First, many students neither continue to participate in learning after enrolling in a course nor meet the standards for passing the course after the course ends. This behavior of "students have not completed the course"[5] prompts the question of how to reduce the "low completion rate" of courses, which is a problem that every MOOCs platform wants to solve. The reason for the low completion rate may be the student's own problems, and they must be properly supervised. It may also be the video material problem, which needs to be properly adjusted or supplemented. There is no clear answer at present, thus, how to reduce the low completion rate is a major challenge for MOOCs [6][7].

Moreover, as the number of students taking MOOCs is much higher than that of traditional courses, it is difficult for teachers to take care of each student's learning situation personally. At present, the simplest way is to arrange more teaching assistants to solve the learning problems of a large number of students; however, as the number of students continues to increase, this is obviously not cost-effective. Therefore, how to help students with poor participation and low motivation becomes an important issue.

This study uses the course of the OpenEdu [8] platform, which is a MOOCs platform based on edX open source, as the research data. As the platform provides a complete teaching environment, including course details and learning history records, we can learn about the students' behavior patterns when browsing videos by processing and analyzing the collected data, and provide possible links between the videos viewing behavior and learning outcomes. We hope to learn about the characteristics of students' learning behaviors with good and poor learning performances to provide a reference for teachers, which will allow them to implement tutoring measures in a timely fashion for students with poor learning performance.

https://doi.org/10.35745/ecei2019v2.079

Related Research

The experimental environment of this study adapts the OpenEdu platform, as established by the Chinese Open Education Consortium, and based on the edX open source software. The platform aims to continuously expand the promotion of open courses, expand the influence level of teaching innovation and change, integrate with the development trend of international digital learning, and take the responsibility of shortening the gap between urban and rural areas, thus, eliminating the digital gap and realizing peoples' equal rights to receive education. To this end, the Chinese Open Education Consortium has joined many schools or institutions interested in developing MOOCs to provide MOOCs construction guidance, teaching platform maintenance, promotion, and other services through the construction of the alliance system, including fund-raising and human operations of the organization.

In their discussions of the low completion rate of MOOCs course, the researchers analyzed the learners' video viewing, scores, and forum behavior records. In [9], where the students' activity behavior patterns were divided into five types: Viewers, Solvers, All-rounders, Collectors, and Bystanders; in [10], the students' activity behavior patterns were divided into seven types: Samplers, Strong Starters, Returners, Mid-way Dropouts, Nearly There, Late Completers, and Keen Completers; and in [11], the students' activity behavior patterns were divided into four types: Dropout, Perfect Students, Gaming the System, and Social. The purpose of the above discussion is to improve students' participation in learning, in order to solve the problem of the low course completion rate.

Machine learning is to classify collected data or train a prediction model through an algorithm, and when new data is obtained in the future, it can be predicted through the trained model. The data of machine learning is composed of feature data and real categories in the process of model training. For example, the first KNN (K Nearest Neighbor) algorithm in this study is generally used to classify data, where K represents a constant, and KNN takes the K points of the nearest distance to determine which category the object belongs to. The second SVM (Support Vector Machine) is an algorithm for supervised learning models, which is often used for pattern recognition, classification, and regression analysis. The third one is an ANN (Artificial Neural Network), composed of many neuron nodes, which can be divided into an input layer, an output layer, and a network model consisting of many hidden layers. The output of the result can only be in the two states of yes or no, while the traditional artificial neural network can train the model by the way of back-propagation, thereby, obtaining a neural network model to effectively solve the problem.

Research Method

The process of this study is shown in Fig.1. MOOCs platform learning records are OpenEdu MOOCs platform data stored in MySQL and MongoDB, while the Tracking Log is stored on the server end in the JSON format. The contents of MySQL data storage include user profile, course records, course basic data, etc. The contents stored in the MongoDB data include course discussion area content, course videos, course exercises, etc. The Tracking Log records the user's behavior on the website, where the records are distinguished by events and have a time stamp. The events include video playing events, discussion forum events, answering events, website browsing events, etc. This study conducts follow-up studies with the data taken from viewing videos. The play action includes six events, load_video, play_video, pause_video, seek_video, speed_change_video, and stop_video.

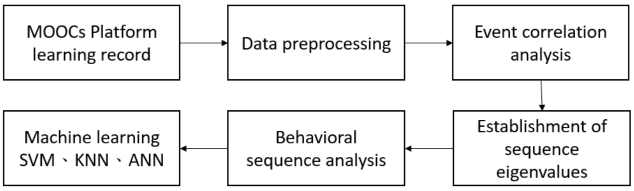
MOOCs Platform learning record
Data preprocessing
Event correlation analysis
Establishment of sequence eigenvalues
Behavioral sequence analysis
Machine learning SVM、KNN、ANN

Fig. 1 Research process

Then, the video playback events were characterized and divided into 8 kinds of feature events according to [11]. The feature event was set as Pl by the start play action of the video (play_video), Pa by the pause action of the video (pause_video), Sf by the forward skipping action of the video (seek_video), and Sb by the backward skipping action of the video (seek_video); the feature event was set as Rf by accelerating the playrate action of the video (speed_change_video) and Rs by decelerating the playrate action of the video (speed_change_video) when the video was played; when the seeking actions of these videos occur within a small time range (<1 second), these seeking events were defined as scroll actions; when the video was played, the feature events were set as Cf and Cb, respectively, by the forward scroll action and the backward scroll action.

In addition, the loading action of the video (load_video) sets the feature event as Lo; the ending action of the video (stop_video) sets the feature event as Sp; the subtitle displays action of the video (show_transcript) sets the feature event as Sh; and the subtitle hiding of the video (hide_transcript) sets the feature event as Hi. Fig.2 observes the occurrence frequency of the any two feature events in a basic course we examined. Since Sh and Hi are less relevant to learning outcomes, they are not included in the observation.

Therefore, this study observed that two feature (ngram=2) events occurred in a total of 72 combinations, of which the combinations with the top 5 highest frequencies were PlPl, PlSf, SfPl, PlPa, and PaPl, in sequence. The three feature (ngram=3) events occurred in a total of 407 combinations, of which the combinations with the top 5 highest frequencies were PlSfPl, SfPlSf, PlPlPl, PlPaPl, and PaPlPa, in sequence. In addition, four feature (ngram=4) events occurred in a total of 1,508 combinations, of which the combinations of the top 5 highest frequencies were PlPlPlPl, PlSfPlSf, SfPlSfPl, PlPaPlPa, and PaPlPaPl, in sequence.

The video sequence behaviors can be divided into seven types [12], Rewatch, Skipping, Fast Watching, Slow Watching, Clear Concept, Checkback Reference, and Playrate Transition, and the above eight feature events are used to define each type of behavior feature, provided the said behavior conforms to one of the video playback feature sequences. At present, we are fully in coincidence with data search. Due to the use of the fixed sequence mode, the frequency in full coincidence with the feature sequences is very low. Therefore, as shown in TABLE I, we use the *: don't care eigenvalue mode to redefine the feature sequence of the 7 types of video sequence behaviors.

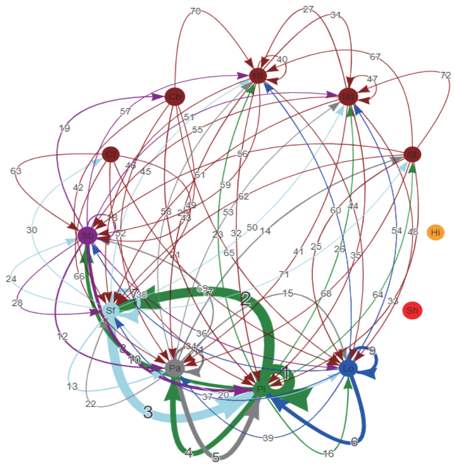

Fig. 2 Video playbck event flow

TABLE I
Video Feature Sequence

No	Behavioral feature type	Video playback feature sequence
1	Rewatch	SbPl**, *SbPl*, **SbPl, PlSb**, *PlSb*, **PlSb, Sb*Pl*, *Sb*Pl Pl*Sb*, *Pl*Sb
2	Skipping	SfSf**, *SfSf*, **SfSf, Sf*Sf*, *Sf*Sf
3	Fast Watching	PlRf**, *PlRf*, **PlRf, RfRl**, *RfRl*, **RfRl, Pl*Rf*, *Pl*Rf, Rf*Pl*, *Rf*Pl
4	Slow Watching	Pl*Rs*, *Pl*Rs, Rs*Pl*, *Rs*Pl
5	Clear Concept	SbCb**, *SbCb*, **SbCb, Sb*Cb*
6	Checkback Reference	SbSb**, *SbSb*, **SbSb, Sb*Sb*, *Sb*Sb
7	Playrate Transition	RfRf**, *RfRf*, **RfRf, Rf*Rf*, *Rf*Rf, RfRs**, *RfRs*, **RfRs, Rf*Rs*, *Rf*Rs, RsRs**, *RsRs*, **RsRs, Rs*Rs*, *Rs*Rs, RsRf**, *RsRf*, **RsRf, Rs*Rf*, *Rs*Rf

The resulting feature records of video watching statistics and test results are merged based on the test unit of the course to record their answers and scores. If a video is not followed by a test in the current learning unit, its viewing statistics will be recorded in the next test unit, which can be used as a predictive feature of learning engagement. The feature items include the number of entries to the course unit, the number of online videos played, the number of plays, load times, play times, pause times, stop times, seek times, speed_change times, Rewatch, Skipping, Fast Watching, Slow Watching, Clear Concept, Checkback Reference, Playrate Transition, the number of tests used, the number of tests answered, the number of tests tried, unit test scores, final test scores, course scores, and course assignment scores. Therefore, the generated course unit activity feature table has a total of 85 feature items. Through the feature selection function, we selected 14 feature values for machine learning model building and prediction.

The KNN, ANN, and SVM methods of machine learning were used according to their characteristics and the training data and test data ratios of KNN and ANN are 80% modeling and 20% verification.

Results and Discussion

As the course contents of the MOOCs platform are very diversified, a basic course, which had been offered at least twice, was selected to conduct analysis. The first class lasts for 6 weeks, contains 55 videos, and has a total of 532,579 learning process records. A total of 590 students took part in the course, of whom 264 obtained the certificate, while 327 failed to obtain the certificate. The second class has a total of 105,435 learning process records, a total of 346 students participated in the course, of whom 137 obtained the certificate, while 209 failed to obtain the certificate. Due to space limitation, we only analyze the second class in the following.

Sequence analysis shows that a total of 78 combinations occurred in the two features (ngram = 2) events of the second class, of which the combinations of the top 5 highest frequencies were PlSf, SfPl, PlPl, LoPl, and LoLo, in sequence; a total of 414 combinations occurred in the three features (ngram = 3) events, of which the combinations with the top 5 highest frequencies were PlSfPl, SfPlSf, PlPlPl, PlPlSf, SfPlPl, in sequence; a total of 1,391 combinations occurred in the four features (ngram = 4) events, of which the combinations with the top 5 highest frequencies were PlSfPlSf, SfPlSfPl, PlPlPlPl, SfPlPlSf, PlPlSfPl, in sequence. Therefore, if the occurrence frequencies of the fixed feature sequence were used to present the degree of learning engagement, it seems that it cannot show its significance. Hence, by adding the *:don't care eigenvalue mode to the four features (ngram=4) events, we redefined the feature sequences of the seven video sequence behavior types, including Rewatch, Skipping, Fast Watching, Slow Watching, Clear Concept, Checkback Reference, and Playrate Transition, to present learners' learning engagement behavior.

After feature selection and removal of 64 interference features, the dimension was reduced to 14 features. First, when the KNN method of R language is used, the library(ISLR) suite should be loaded beforehand using the knn() function, including 80% of the feature training data sets, 20% of the feature test data sets, and the real classification factors of the training set were course pass (1) and fail (0), where the K value (# of neighbors) was calculated as the square root of the number of the click counts, and the accuracy is 0.7948717949.

As the accuracy of KNN is poor, when the SVM method was used, the library(e1071) suite should be loaded beforehand in R language, and svm() was used to train the classification model of SVM, including 80% of the feature training data sets, 20% of the feature test data sets, and the target values of the training sets were course pass (1) and fail (0); next, the test data and the training data were used to build the prediction model, and the accuracy was0.8974358974 by using the predict() function.

To make further improvement, when the ANN method of R language was used, the library(nnet) suites should be loaded beforehand using the ann() function, including 80% of the feature training data sets, 20% of the feature test data sets, and the target values of the training sets were course pass (1) and fail (0); the variable factor was 14 feature data values, the number of units in the hidden layer was from 1 to 10, the parameter of the specific gravity attenuation was 0.001, and the maximum number of repetitions was 1000. When the numbers of units in the hidden layer were 1 to 5, the highest accuracy of the modeling was 0.902857143. Therefore, in the modeling and prediction of the first class data set, we determined that the accuracy of the best model was the highest when the number of units in the hidden layer of ANN was 2, and the results are shown in TABLE II. Under the prediction of the first class, we used the results of the three hidden layers of ANN as the prediction model.

TABLE II
ANN accuracy

Model	Size	Modeling accuracy of the first class	Prediction accuracy of the second class
ANN	1	0.897142857	0.877887789
	2	**0.902857143**	**0.884488449**
	3	0.874285714	0.858085809
	4	0.874285714	0.877887789
	5	0.845714286	0.831683168

Based on the first class feature data, the prediction model was established, and then, the second class feature data was used to predict the accuracy. Under the condition of ANN size = 2, the six-week data of the first class was used to predict the 6-week data of the first class, and the accuracy was improved from 0.8914 to 0.9485; while the six-week data of the first class was used to predict the six-week data of the second class, and the accuracy was reduced from 0.902857143 to 0.884488449, thus, using this method has indeed achieved the predicted effects.

Conclusion

This study used the click records of MOOCs videos. Firstly, the feature sequence of the viewing learning behavior is established with Ngram=4, and the feature sequence was redefined in the don't care mode as the type of learner's cognitive participation; this study used the k-Nearest Neighbor Classification (KNN) method, Support Vector Machines (SVM), and Artificial Neural Network (ANN) to predict whether or not students pass; finally, the predicted results of the first class were KNN accuracy 0.7948717949, SVM accuracy 0.8974358974, and ANN accuracy was up to 0.902857143 under two hidden layers.

In addition, the weekly tutoring list of students was provided for teachers to supervise students' learning progress. There were 313 students who needed tutoring in the first week, 313 in the second week, 311 in the third week, 305 in the fourth week, and 297 in the fifth week.

Then, the prediction accuracy of the second class was as high as 88%, and the prediction accuracy of ANN under three hidden layers was as high as 0.884488449. The weekly tutoring list of students was provided for teachers to supervise students' learning progress. There were 155 students who needed tutoring in the first week, 131 in the second week, 130 in the third week, 130 in the fourth week, and 131 in the fifth week.

Therefore, through the inference and prediction mechanism, this study analyzed the behavioral patterns and features of students' video browsing behaviors to determine the correlation between the video viewing behavior and learning outcomes, understand the features of students' learning behaviors with good or poor learning outcomes, and make predictions, which will provide a reference for teachers, in order that teachers can implement tutoring measures in a timely fashion for students with poor learning outcomes and the course completion rate can be improved.

References

[1] Kay, J., Reimann, P., Diebold, E., & Kummerfeld, B. (2013). *MOOCs: So Many Learners, So Much Potential*. IEEE Intelligent Systems, 28(3), 70-77..

[2] Severance, C. (2012). *Teaching the World: Daphne Koller and Coursera*. Computer, 45(8), 8-9.

[3] Breslow, L., Pritchard, D. E., DeBoer, J., Stump, G. S., Ho, A. D., & Seaton, D. T. (2013). *Studying Learning in the Worldwide Classroom: Research into edX's First MOOC*. Research & Practice in Assessment, 8, 13-25.

[4] Shi, C., Fu, S., Chen, Q., & Qu, H. (2015). *VisMOOC: Visualizing Video Clickstream Data from Massive Open Online Courses*. Proceedings of the IEEE Pacific Visualization Symposium (PacificVis), pp. 159-166.

[5] Liang, J., Li, C., & Zheng, L. (2016). *Machine Learning Application in MOOCs: Dropout Prediction*. Proceedings of the 11th International Conference on Computer Science & Education (ICCSE), pp. 52-57.

[6] Elbadrawy, A., Polyzou, A., Ren, Z., Sweeney, M., Karypis, G., & Rangwala, H. (2016). *Predicting Student Performance Using Personalized Analytics. Computer*, 49(4), 61-69.

[7] Brinton, C. G., & Chiang, M. (2015). *MOOC Performance Prediction via Clickstream Data and Social Learning Networks*. Proceedings of the 2015 IEEE Conference on Computer Communications (INFOCOM), pp. 2299-2307.

[8] OpenEdu, https://www.openedu.tw/.

[9] Anderson, A., Huttenlocher, D., Kleinberg, J., & Leskovec, J. (2014). *Engaging with Massive Online Courses*. Proceedings of the 23rd International Conference on World Wide Web, pp. 687–698.

[10] Ferguson, Rebecca and Clow, Doug (2015). *Examining Engagement: Analysing Learner Subpopulations in Massive Open Online Courses (MOOCs)*. Proceedings of the 5th International Learning Analytics and Knowledge Conference (LAK15), pp. 1-8.

[11] Khalil, M., & Ebner, M. (2016). *Clustering Patterns of Engagement in Massive Open Online Courses (MOOCs): the Use of Learning Analytics to Reveal Student Categories*. Journal of Computing in Higher Education, 29(1), 114–132.

[12] Sinha, T., Jermann, P., Li, N., & Dillenbourg, P. (2014). *Your Click Decides Your Fate: Inferring Information Processing and Attrition Behavior from MOOC Video Clickstream Interactions*. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 3-14.