Enhance E-Learning through Data Mining for Personalized Intervention

Lingma Lu Acheson and Xia Ning
Department of Computer and Information Science, Indiana University-Purdue University Indianapolis, 723 West Michigan Street, SL280, Indianapolis, IN 46202, U.S.A.

Keywords: Education Data Mining, E-Learning, Analytics in Education, Assessment.

Abstract: E-Learning has become an integral part of college education. Due to the lack of face-to-face interactions in online courses, it is difficult to track student involvement and early detecting their performance decline via direct communications as we typically practice in a classroom setting. Hence there is a critical need to significantly improve the learning outcomes of online courses through advanced, non-traditional approaches. University courses are often conducted through a web learning management system, which captures large amount of course data, including students' online footprints such as quiz scores, logged entries and frequency of log-ins. Patterns discerned from this data can greatly help instructors gain insights over students learning behaviours. This positioning paper argues potential approaches of using Data Mining and Machine Learning techniques to analyse students' online footprints. Software tools could be created to profile students, identifying those with declining performance, and make corrective recommendations to instructors. This timely and personalized instructor intervention would ultimately improve students' learning experience and enhance their learning outcome.

1 RESEARCH PROBLEM

Data Mining (DM) and Machine Learning(ML) tools have been widely used in education to assist instructors in understanding students and improving learning outcomes, particularly in online education. Due to the flexibility and often student-centred curriculum design, online courses have become an integral part of college education, hence there is a critical need to significantly improve the learning outcomes of online courses through advanced, non-traditional approaches. We believe the automated analysis of student performance and behaviour though advanced DM and ML methods, together with the intelligent personalized interventions, has a great potential to achieve the goal.

University courses are often conducted through a web learning management system such as Blackboard or Canvas. These systems capture large amount of course data, including students' online footprints such as quiz scores, logged entries and frequency of log-ins. Patterns discerned from this data can greatly help instructors gain insights over student learning behaviours. We project that it would introduce a significant amount of benefits if data mining and machine learning techniques are used to analyse students' online footprint. From this, software tools would be created to profile students, identifying those with declining performance, and make corrective recommendations to instructors.

2 OUTLINE OF OBJECTIVES

The objectives are designed to apply advanced DM and ML tools in online courses to understand student performance and behaviours to detect performance decline and risk of dropout, and to design and deliver personalized interventions to students, so as to increase their engagement and improve their learning outcomes. We believe tool development is necessary to answer the following two questions:

Question 1: How to understand and forecast student learning experience and performance based on their online footprints?

Acheson, L. and Ning, X.
Enhance E-Learning through Data Mining for Personalized Intervention.
DOI: 10.5220/0006793304610465
In *Proceedings of the 10th International Conference on Computer Supported Education (CSEDU 2018)*, pages 461-465
ISBN: 978-989-758-291-2

Question 2: How to construct and deliver personalized interventions via peer-to-peer off-line communications?

Due to the lack of face-to-face interactions in online courses, it is difficult to track student involvement and early detecting their performance decline via direct communications as we typically practice in a classroom setting. Fortunately, students usually leave a lot of digital footprints whenever they take the courses, participate the online forum discussions, submit homework, read online slides, etc. Such digital footprints are very valuable information for the instructors to understand student behaviours, and make meaningful interpretations and predictions therefrom. However, given the fact that online courses are normally large classes, it will be a huge workload if instructors manually analyse such footprint data. In addition, the highly heterogeneous student body makes any manual analysis highly nontrivial. This is because any conclusions for an individual student may or may not apply to others, for example, from a different major.

3 STATE OF THE ART

Schools offering fully online, hybrid and web-enhanced degree programs have seen substantial growth over the past ten years and all signs show that growth will continue at this rapid rate ("How Prevalent is Online Learning", 2017). In addition, Massive Open Online Courses offer a wide range of online educational programs from leading universities (Combs & Mesko, 2015). One clear advantage of an online course is that logs can provide clues about learner experiences in relation to ease of course navigation and perceived value of content (Robyn, 2013). On the other hand, the flaw of MOOCs were eagerly dissected – high dropout rates, limited social interaction, heavy reliance on instructivist teaching, poor results for underrepresented student populations, and so on (Bunk et al., 2015). For example, a program introduced by San Jose State University and Udacity to run remedial courses in popular subjects ended in a failure rate of up to 71% percent (Devlin, 2013). Despite of this, the amount of data generated from online courses are skyrocketing. Researchers and developers of online learning systems have begun to explore analogous techniques for gaining insights from learners' activities online (U.S. Department of Education, 2012).

EDM has been emerging into an individual research area in recent years (Baker et al., 2010). Several main research focuses are developed in EDM, including student behaviour modelling, student performance modelling, assessment, et. al. Bayes theorem, Hidden Markov Model, decision trees et. al. are among the most popular methods applied in these researches (Pena-Ayala, 2014).

Methods such as Collaborative Filtering (CF) (Ning, Desrosiers, & Karypis, 2015) and Matrix Factorization (MF) (Koren, Bell, & Volinsky, 2009), have attracted increasing attention in EDM applications, due to their strong ability to deal with sparse data for ranking, prediction or classification, which is particularly common in EDM. For example, Sweeney et. al. (2015, 2016) adopted developed methods including SVD, SVD-kNN and Factorization Machine (FM) to predict next-term performance. Polyzou and Karypis (2013) addressed the future course grade prediction problem with three approaches: course-specific regression, student-specific regression and course-specific matrix factorization. Moreover, neighborhood-based CF is one of the most popular methods in EDM. Many existing approaches (Ray & Sharma, 2011; Bydzovska, 2015; Denley, 2013) predict grades based on the student similarities, that is, they first identify similar students and use their grades to estimate the grades of the students of interest.

In order to capture the change of student dynamics over time, various dynamic models have been developed in EDM. Sun et. al. (2012, 2014) modelled student preference change using a state space model on latent student factors, and estimated student factors over time using noncausal Kalman filters. Similarly, Chua et.al. (2013) applied Linear Dynamical Systems (LDS) on Non-negative Matrix Factorization (NMF) to model student dynamics. Zhang et. al. (2014) learned an explicit transition matrix over the latent factor for each student, and solved for the student and course latent factors and the transition matrices within a Bayesian framework.

4 METHODOLOGY

To answer question 1, we argue that applying DL and ML tools to analyse the digital footprints of a carefully chosen online course would be a good pilot. We believe particular focus on the following information is necessary: 1) time students spend on slide reading and course video watching, 2) the frequency that students log into the learning system, 3) the frequency that students participate in online forum discussion and time they spend, 4) their interactions with other students on the forum through

asking/answering others' questions, and 5) homework scores, quiz scores and student performance on each of the single questions, etc.

To answer question 2, we could identify students of declining performance. Customized learning materials and interventions will be delivered, for example, through homework assignment, email communications, etc. Students would be characterized by different traits, for example, short of necessary reading, weak at concept understanding, according to our analysis outcomes. We would design and maintain a pool of homework questions, and assign the questions that are helpful in improving the corresponding weakness to respective students. Matching algorithms could be applied so as to guarantee coverage and fairness in the personalized assignment for all students. Email communications would also take place so that we could get feedback on the learning experience, suggestions, demands, etc., from identified students. Such feedback would be further integrated into our DM and ML tools to consistently improve the analysis accuracy and sensitivity.

Typically, the mentioned intervention could be divided into three stages – data collection, model building, and model testing. At the early stage, digital footprints would be collected and properly formatted. Once sufficient data are in place, we could perform some initial data analysis to gain general understanding of student behaviours and their association with learning outcomes, then apply DM and ML model on such data. It is expected that due to the high heterogeneity of the student body, we might need to adjust our previous models (e.g., via parameter configurations, including additional components) to have it more adapted to the current student body. We would then apply the model to the students. We would also get feedbacks from the students and get feedbacks on the model predictions. Based on such feedbacks, we would continuously improve and adjust the model for better predictions.

The course to be chosen to participate in this study is one of the institution's general education core, thus is a course of significant importance. It routinely enrols over 600 students per academic year, with approximately 500 of those being online and with a diverse student body majoring in science, technology, business, education, liberal arts, philanthropy, communication etc.. Around 50% of the students are first-year and second-year students, who generally face significant change, growth and challenge after stepping out of high school. They have unique needs and require additional support to nurture their cognitive learning and emotional development. If more instructor intervention and individual attention is given, it will greatly build their confidence in continuing with their college education. In particular, we will choose an online format because online courses can have large enrolments, and this approach requires sufficient amount of data. Also, online courses need special attention due to lack of face-to-face time with students. Online courses require more time-management skills and more self-discipline. Lack of face-to- face time makes it hard for instructors to interact with students, identify problems and disperse timely feedback. Thus an online session will be a good candidate to measure the effectiveness of the proposed approach.

In the chosen session, all study materials (syllabus, slides, videos, and supplementary documents), instructions and assignments would be published on the learning management tool, namely Canvas, before the semester starts. Deadlines for assignments would be announced at the beginning of the semester so students could set pace for themselves. Assignments will include projects, quizzes, reading materials and exams. Weekly emails would be sent out to students providing summary for previous weeks(s), detailed guidance for the following week, useful tips or due date reminders.

For this study, we would be mostly interested in discerning learning patterns from students' online footprints, thus data extracted from Canvas will be focused on time logged onto Canvas, time spent on each quiz and quiz scores, performance on each quiz problem, number and time of assignments submitted, assignment scores, time spent on exams and each exam problem, performance on each exam problem and overall exam scores, and course grades. Figure 1 is a snap shot of this data set that shows the number of page views, length of login time, grades for each assignment etc., for each student. Submission status

Page Views	Log-time in Minutes	Project 3	Project 4	Project 2	Project 5
679	1776	90	85	90	100
356	873	100	85	97	100
415	1325	89	100	90	100
435	763	93	80	100	100
583	685	60	65	95	100
718	759	75	75	92	100
627	939	100	100	94	100
462	442	85	100	32	100
367	702	85	80	100	90
396	999	96	100	100	100
505	568	96	90	46	100
356	1532	81	95	82	100
603	958	82	100	100	0
355	464	87	85	88	95
506	520	68.8	80	63	0
2099	698	88	85	69	90

Figure 1: Number of times student accessing course web pages, total login time in minutes and assignment grades.

for each assignment will be captured as shown in Figure 2, including due time, submission time, whether it is "On Time" or "Late", grade for each submission and so on. Item by item analysis for each quiz and exam questions as illustrated in Figure 3, is also available on Canvas. This will provide significant insight on areas of weakness in concept understanding.

Assignment Name	Status	Due At	Submitted At	Score	Performance	Low	Median	High
Orientation Quiz，due August 25, 11:50pm	On Time	2017-08-25	2017-08-23	95	Good	0	85	100
Quiz-1-slides1&2, due September 1st, 11:50pm	On Time	2017-09-01	2017-09-01	80	Good	0	73.33333	100
Project 1 Slides 1 & 2	On Time	2017-09-08	2017-09-06	90	Good	0	89	100
Quiz-2-slides3, due September 15th, 11:50pm	On Time	2017-09-15	2017-09-14	85.8333	Good	0	78.33333	100
Project 2 Slides 3	On Time	2017-09-22	2017-09-22	100	Good	0	92	100
Exam 1, Open Sep 28								

Figure 2: Submission status for each assignment.

Correct Student Count	Wrong Student Count	Correct Student Ratio	Wrong Student Ratio
41	31	0.569444444	0.430555556
63	9	0.875	0.125
57	14	0.802816901	0.197183099
64	7	0.901408451	0.098591549
49	22	0.690140845	0.309859155
34	37	0.478873239	0.521126761
57	14	0.802816901	0.197183099
51	20	0.718309859	0.281690141
68	3	0.957746479	0.042253521
26	45	0.366197183	0.633802817
65	6	0.915492958	0.084507042
65	6	0.915492958	0.084507042

Figure 3: Quiz item-by-item result analysis.

5 EXPECTED OUTCOMES

The predicted learning outcomes are expected to be better learning experience and better final grades overall. The students would be better involved in the course through personalized interventions and communications, and would better master the course materials through customized homework assignment. Eventually, the students are expected to have better homework grades and final grades.

Course evaluation data would be monitored both prior to and after implementation of this approach. Beginning and End-of-semester surveys would be given to students. We would understand the student expectations and experience regarding the DM and ML analysis, and get the feedback as to whether they feel more effective and involving in the learning experience, whether they prefer the personalized intervention and communication, and what comments they have, etc. We would also compare the performance of students with DM and ML applied with that of students from previous years without DM and ML applied. We should make sure the comparison is fair (e.g., only students of a same major or similar background will be compared) so as to get unbiased conclusions.

We would do periodical surveys on students to get their feedbacks. We would adjust our analysis strategies and models according to the feedbacks. We would also communicate with students via emails or forum posts to get their personalized comments and suggestions. We would correspondingly tailor our model with respect to certain comments or requirements.

Direct evidence could be their final grades. We expect that with DM and ML analysis in place, the students will have better grades by the end of the semester. Another evidence could be their increasing performance during the course of the learning experience. With personal interventions, we expect students be more and more involved, and their performance will be continuously improved, which can be measured by their grades on homework assignments. Other indirect evidence could include active participation in the online forum, which can be measured by the time they spend and the number of posts they post; their communications with instructors, which can be measured by the frequency of email exchanges and question/answering interactions, etc.

REFERENCES

How Prevalent is Online Learning at the Collegiate Level. (2017, November 19). Retrieved from http://www.online-psychology-degrees.org/faq/how-prevalent-is-online-learning-at-the-college-level/

Combs, C & Mesko, B. (2015). Disruptive Technologies Affecting Education and Their Implications for Curricular Redesign. *The Transformation of Academic Health Centers: Meeting the Challenges of Healthcare's Changing Landscape*. 57-68. 10.1016/B978-0-12-800762-4.00007-4.

Robyn P. (2013). Redesigning Courses for Online Delivery, *Cutting Edge Technologies in Higher Education*, Volume 8

Bunk, C. et al. (2015). *MOOCs and Open Education around the World*, Routledge

Devlin, K. (2013). *MOOC Mania Meets the Sober Reality of Education*, Huffington Post

U.S. Department of Education, Office of Educational Technology (2012). Enhancing Teaching and Learning Through Educational Data Mining and Learning Analytics: *An Issue Brief*, Washington, D.C.

Baker, R.S.J.d. et al. (2010). Data mining for education. *International encyclopedia of education*, 7:112–118

Pena-Ayala, A. (2014). Educational data mining: A survey and a data mining-based analysis of recent works. *Expert systems with applications*, 41(4):1432–1462

Ning, X., Desrosiers, C. & Karypis, G. (2015). A comprehensive survey of neighborhood-based recommendation methods. In Francesco Ricci, Lior Rokach, and Bracha Shapira, editors, *Recommender Systems Handbook*, pages 37–76. Springer

Koren, Y., Bell, R. & Volinsky, C. (2009). Matrix factorization techniques for recom- mender systems. *Computer*, 42(8):30–37

Sweeney, M., Rangwala, H., Lester, J., & Johri, A., (2016). Next-term student performance prediction: A recommender systems approach. *preprint arXiv*: 1604.01840

Sweeney, M., Lester, J., & Rangwala, H., (2015). Next-term student grade prediction. *In Big Data (Big Data), 2015 IEEE International Conference on*, pages 970–975. IEEE

Polyzou, A. & Karypis, G. (2016). Grade prediction with models specific to students and courses. *International Journal of Data Science and Analytics*, pages 1–13

Ray, S. & Sharma, A. (2011). A collaborative filtering based approach for Recommending elective courses. *In International Conference on Information Intelligence, Systems, Technology and Management*, pages 330–339. Springer

Bydzovska, H. (2015). Are collaborative filtering methods suitable for student performance prediction? *In Portuguese Conference on Artificial Intelligence*, pages 425–430. Springer

Denley, T. (2013). Course recommendation system and method. *US Patent App*. 13/441,063.

Sun, J., Parthasarathy, D. & Varshney, K. (2014). Collaborative kalman filtering for dynamic matrix factorization. *IEEE Transactions on Signal Processing*, 62(14):3499–3509

Sun, J., Varshney, K., & Subbian, K. (2012). Dynamic matrix factorization: A state space approach. *In 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1897–1900. IEEE

Chua, FCT., Oentaryo, R. & Lim, E. (2013). Modeling temporal adoptions using dynamic matrix factorization. *In 2013 IEEE 13th International Conference on Data Mining*, pages 91–100. IEEE

Zhang, C., Wang, K., Yu, H., Sun, J., & Lim, E. (2014). Latent factor transition for dynamic collaborative filtering. *In SDM*, pages 452–460. SIAM