Baltic J. Modern Computing, Vol. 8 (2020), No. 1, 154-163
https://doi.org/10.22364/bjmc.2020.8.1.08

Input Determination for Models Used in Predicting Student Performance

Kārlis KRŪMIŅŠ[1], Sarma CAKULA[2]

[1]Riga Technical University, Kaļķu iela 1, Rīga, LV-1658, Latvia,
[2]Vidzeme University of Applied Sciences, Cēsu iela 4, Valmiera, LV - 4201, Latvia,

karlis.krumins_3@edu.rtu.lv, sarma.cakula@va.lv

Abstract. To capture and code as much as possible of student behavior and environment to apply learning analytics in conventional classrooms, patterns among successful inputs of existing online learning / learning management systems can be identified, to find existing but uncaptured classroom data. The goal of this review is to suggest proposals on expanded use of learning analytics in traditional classrooms. Predictors from learning models used in online learning can be applied to the traditional classroom and analogues may be found for unavailable predictors. Approaches used in developing these predictors can be used to develop predictors for conventional classrooms. Existing data can be used, or data that is convenient may be captured, with emphasis on approaches that work on smaller groups, where training of individual models should be attempted. The data collected should be simple to obtain, support the users otherwise, or have provable benefits.

Keywords: online learning, blended learning, input selection, prediction model, learning analytics.

1. Introduction

Student performance prediction has become a viable means to improving academic performance and course content in online learning. Predictive models such as artificial neural networks, decision trees and linear regression are used to transform inputs (e.g. past performance, social background, learning system usage patterns, test results) into outputs (course completion, expected grade, difficulties encountered, personalized learning suggestions). Often, existing quantitative data (e.g. from Massive Open Online Courses – MOOCs) drive model design, especially when applying such models to the conventional classroom (using a Learning Management System – LMS), and the person delivering the course is a passive participant in designing models and delivering data. This produces a "streetlight effect", "looking for solutions where it is easy to search, not where the real solution might be" (Ochoa et al., 2016), as only data from online learning is used, and even there only data that can be acquired easily. Therefore, one current approach to expanding data available is multimodal learning analytics, where automated systems are used to capture multimodal sensory data from collaborative, real-world, non-computer mediated environments (Ochoa et al., 2016).

Another approach may be to persuade educators and learners to directly input previously unused data into learning analytics systems, by either producing a more effective learning analytics system (one that improves learning) or by providing direct benefits from this data capture (for example checklists like by recording the time when feedback is provided / works are returned to students, pedagogical personnel are assured they will never forget to return and comment on a student's work). In seeking to capture and code as much as possible of student behavior and environment to apply learning analytics to a mostly conventional classroom, the most successful inputs (predictors) among existing models can be identified, categorized and their common characteristics determined. Together with a study of formative and summative assessment methods (e.g. types of feedback and how it can be captured) and factors affecting student performance in classroom (e.g. environmental factors), this allows identifying existing data in classrooms that are not captured by current learning management systems, thus allowing expanded use of learning analytics and student performance prediction in traditional classrooms, with a focus on personalized suggestions.

The goal of the paper is to identify patterns among inputs used in existing models of student learning (based on online learning and learning management system data mining) that can then be applied also to the traditional classroom within its existing workflow, by looking at the features used by existing and proposed learning analytics approaches.

Research question: how characteristics common to effective predictors of student performance can be used to identify predictors among data produced in the traditional classroom?

2. Methodology

A literature review is performed where inputs captured and features discovered in existing learning analytics systems are characterized, along with the methods used to identify features and the modelling approaches employed. The learning setting and the stakeholders who are expected to benefit from the studied approach are recorded. Studies that describe their methodology in an exhaustive manner are highlighted.

Searches were performed using Google Scholar, Ex Libris Primo, and using the search features of academic databases such as ACM Digital Library, EBSCO Academic Search Ultimate JSTOR, ScienceDirect, Wiley Online Library, etc.

Search terms used were either broad ("learning analytics dataset ", "classroom data", "multimodal learning analytics") or more narrow, prejudicial guesses ("classroom comfort performance", "student age learning analytics"). A significant limitation of the study is that the narrow guesses did produce some articles that were not found otherwise, therefore suggesting that the search results using broad terms were not exhaustive. Also, conceptual articles are excluded from the study, even though some did identify promising approaches. The focus of this study is on articles about specific methodologies used in learning analytics or educational data mining, therefore excluding review articles.

An attempt is made to identify measures in online learning that may have analogues in the traditional classroom (e.g., seating patterns and communication in chatrooms) or for which proxies may be found (e.g., screen size and lighting quality, where the proxy is the classroom number).

The corresponding outputs are recorded where possible, with a focus on those that allow providing feedback for individual students or for course/curriculum deliverers/designers (to allow improving the success of future students in this course).

3. Results

A total of 250 articles responding to the search queries were selected based on their titles. Of these, 53 could be broadly classified as review articles, and 114 were either conceptual, or dealt with further aspects of learning analytics such as data presentation or providing feedback. Of the 85 articles that supported the effort to identify features that could be used in developing an LMS, 68 deal with analytics in MOOCs or datasets of existing LMSs, where at least 15 of these provide significant methodological detail on the inputs used, and are described in Table 1.

Table 1. Lessons learned from applying Learning Analytics (LA) to online learning

	Setting	**Inputs / identified features / feature extraction approach**	**Ideas gained from article**
Yang et al., 2017	MOOC	Video clickstreams, quiz results	Detailed features improve prediction. Models can be trained per individual student – cannot anonymize data too early.
Veeramachaneni et al., 2014	MOOC	Clickstreams, forum and wiki activity, student state	Crowdsourced and derived features can be better predictors.
Jo et al., 2015	LMS	Time management strategy through login data	Use ideas from theoretical studies of learning theory to construct proxy variables.
Fernandes et al., 2019	EDM	Demographics, attendances, grades	Demographic data should be used to identify concerns before course start.
Hone and Said, 2016	MOOC	Survey data post-MOOC (completion/dropout)	Course structures enforced by MOOCs may improve success. Surveys provide important feedback, especially from early dropouts where other data not available in quantity.
Lonn et al., 2015	LMS	Survey data, LMS data	Feedback provided by an LMS to students may harm learning.

Rienties and Toetenel, 2016	VLE	Aggregate statistics per week	Learning design (student activity types) can be identified and linked to engagement, satisfaction and retention
Maldonado-Mahauad et al., 2018	MOOC	Process mining	Learning strategies can be identified to support learners
Ma et al., 2015	LMS	Instructor activity	Instructor activity during the course has significant impact on completion
Zacharis, 2015	Blended	Message activity, contribution to output, quiz attempts	Availability of optional content may provide more data while improving motivation
Gašević et al., 2015	LMS	LMS trace data, data from the institutional student information system	Instructional conditions significantly affect the effect of an LMS on students and prediction models may not be generalizable across courses
Ramesh et al., 2014;	MOOC	Forum post content (using seeded LDA), course results	Forum post content (tone) can be used to predict possible dropouts
Wang et al., 2015	MOOC	Forum post content (using a manually coded training set)	Forum behavior (on/off topic, active/interactive/ constructive contribution) predicts student learning gains
Hernández-García et al., 2018	LMS	Statistics (rates, means, proportions) of both active and passive interactions in forums	Team-specific indicators (teamwork assessment) is possible using LMS log data . The user submitting collaborative work may not be the main author (one team member may be designated to perform submissions).
Blikstein et al., 2014	LMS	Code submission patterns and code characteristics	Changes in behavior pattern and high frequency data can be powerful predictors but require intensive computing resources

An attempt to broaden the data available to learning analytics practitioners was found to be covered in 11 articles, as described in Table 2.

Table 2. Lessons learned from attempting to widen the data available to learning analytics

	Idea/tool	Information monitored	Ideas gained from article
Prasad et al., 2016	Apply LA to existing textbooks using a customized epub viewer	Page views / chapter views, reading sequence across time/student	Adapt existing digital, but not instrumented/online, resources to LA approaches
Pardo, Kloos, 2011	Apply LA to any computer-based activity	Use a VM to monitor student activity	Possible to monitor how much students use tools that were not preselected to be instrumented
Ferguson, Shum, 2011	LA on realtime customized conference chat environment	Monitoring text chat during a face-to-face activity	It may be possible to extract analytics from an online chat that is in parallel to an activity
Tabuenca et al., 2015	Self-reported monitoring of time spent learning combined with repeated questionnaires	Patterns of time management, learning strategy	Self-reporting of time management and affective aspects of learning can be combined
Leeuwen et al., 2015	Monitor teacher involvement with a LA tool	What support teachers provide to who	Accept LA tools may provide general improvements to study even if not reaching the initially intended objectives
Fulantelli et al., 2014	Create a tool that supports mobile learning activities while collecting analytics	Monitor interactions with the learning environment, context, communication	LA can be applied to mobile learning, by simultaneously providing support for learners and monitoring them
Donia et al., 2018	Improve teamwork by requiring and reporting peer feedback	Monitor performance of students against feedback received	Relatively low effort additions to an existing offline course can provide guaranteed improvements
Jovanović et al., 2017	Monitor students who use an LMS in a flipped classroom	Learning state distribution across time (i.e. distribution of learning actions)	Student learning pattern clusters allow identifying successful strategies that could then be suggested to others
Worsley et al., 2011	Analysis of student speech audio	Transcribed speech and sound file analysis	Techniques used for analysing text can be applied to speech to identify competency
Ezen-Can et al., 2015	Posture and gesture recognition	Using a Kinect sensor to improve dialogue analysis	Improve understanding of student interactions in learning environments; use better monitored example interactions to improve models
Schneider and Blikstein 2015	Combine an interactive learning environment with physical monitoring	Monitor interactions with a tangible user interface using a Kinect sensor	Multimodal analytics is very promising even when using unsupervised learning algorithms

Successful learning analytics approaches use fine-grained longitudal data, where it is often impossible to predict which specific measures will be the best predictors. Therefore, at least during the prototyping stages, more inputs than would be used in production should be requested, while in production it should be easy for users, educators and learners alike, to volunteer more data than requested by default. A dashboard could be provided for monitoring the effectiveness of features at predicting student performance, with the ability to suggest features in addition to the built in ones, as Veeramachaneni et al. (2014) have shown the success of user-invented inputs to student models.

If data are fine-grained enough, it becomes possible to train models for individual students, that Yang et al. (2017) have shown to be more effective than general models; in addition, to improve such models, or adapt models to a specific cohort, it may be possible to undertake sessions where additional sensors are used to collect additional physical data for model training that Ezen-Can et al. (2015) have shown to improve performance of models. If resources permit, multimodal data have been shown to be promising (Schneider and Blikstein, 2015) and consideration should be given even to simply recording speech (Worsley et al., 2011), if privacy issues are accounted for. In addition, model performance during the beginnings of a course may be improved using directed surveys, which should also be given to dropouts (Hone, Said, 2016).

If possible, textual, not quantitative, data should be collected, as it has been shown to provide information on student state, often early, though this does seem to require a coded training set, both of low effort to predict dropouts (Ramesh et al., 2014) or if the model developer puts in more effort, to monitor student learning gains (Wang et al., 2015). In addition, this would permit applying LA techniques to classes where some of the most popular LA outputs are unavailable, as for example in classes where dropping out is not possible, motivation does drop and this may become visible in student textual output.

When constructing features, theoretical studies of learner behaviour from other pedagogic research fields can be used to inform model development (e.g. Jo et al., 2015) or to develop tools to monitor student learning strategies (e.g. Tabuenca et al., 2015).

Process mining that identifies activity patterns has been shown to be a powerful predictor (Maldonado-Mahauad et al., 2018), therefore as much as possible information on event timing should be collected. In addition to monitoring students, instructor activity and interaction with learning analytics and students should be monitored, as multiple studies have shown these to have a significant impact on learner performance (e.g. Gašević et al., 2015, Ma et al., 2015, Leeuwen et al., 2015).

Interventions are possible early by using demographic data that is already known about a learner at the beginning combined with models trained on large datasets (Fernandes et al., 2019), though this may require access to datasets relevant to the specific region or educational system, which presents privacy issues and requires governmental stakeholder involvement.

If possible, adaptations can be developed for specific types of courses. For example, programming courses benefit from instrumented integrated development environments (Blikstein et al., 2014) and it may be possible to develop generalizable data collection techniques for any computer based activity using an instrumented virtual machine (Pardo and Kloos, 2011), though both approaches may require unfeasible amounts of data processing even for small cohorts.

4. Discussion

Recently, there has been more focus on increasing the visibility into models of learning and on involving learning personnel in designing, modifying and running those models. Providing inputs and recognizing the features they represent determines the success of such models. Therefore, recognizing existing successes and applying them to formative assessment methods may be a means of recognizing additional inputs to and features used in models, while involving educators. Applying learning models to the traditional classroom as an integrated part of the learning management (school record keeping/grading) systems may allow to expand their use, while simultaneously increasing the predictive power and effectiveness of (personalized) suggestions, both by using existing data, and by providing tools for educators to transform the existing feedback they provide into data than can be used as inputs for models.

It is evident that for a learning analytics platform to be successful, it needs to either provide the ability to collect data close to effortlessly or to provide benefits to those providing the data independent of any learning analytics output. The ways a learning analytics platform can be valuable to teachers include, for example, giving them access to tools that reduce their effort or improve their confidence. The course structures enforced by MOOCs have been shown to be successful (Hone, Said, 2016) so if there is the desire to provide blended learning or online learning opportunities, providing such structures in the LMS can attract both educators and students as users, as Zacharis (2015) has shown optional content to be capable of improving motivation and as this would improve educator confidence that their courses are structured according to best practices.

Another approach to providing value with relatively low effort is instrumenting already existing resources, as Prasad et al. (2016) have shown by developing an epub (an electronic book format) viewer to monitor interactions similarly to how a MOOC does; many courses in schools use books that are available as PDFs, therefore instrumenting such books may provide clickstream-like information on reading habits without the requirement to redevelop course resources. In addition, for those subjects that perform field studies, techniques from mobile learning may be adapted, with success shown if the developed applications actually support the student in their learning goals required by the existing course content while providing LA data (Fulantelli et al., 2014).

As activity patterns are a powerful predictor of learner performance, an LMS should integrate as much of the "bookkeeping" functions of a school (or be integrated with existing systems), as this would permit collecting data without additional effort from the teacher (e.g., absences, lesson times, students that study together, etc.) or to collect additional data while providing value to the teacher (e.g., when a teacher provides feedback/hands back work he/she records this interaction in the LMS using a checklist; this provides event timing data while permitting the teacher certainty that he has spoken with the specific student about the specific work and provided required feedback).

Rienties and Toetenel (2016) have shown that it is possible to identify successful course design patterns, again providing value to those educators who maintain their (often legally required) course (design) documentation in an LMS. An LMS may be capable of identifying learning patterns that are effective (Maldonado-Mahauad et al., 2018) and these patterns can then be suggested to other students, which applies even to flipped classrooms (Jovanović et al., 2017), again providing value.

Collaborative learning is gaining more and more providence, hence tools that support it may also gain acceptance more easily. Donia et al. (2018) have shown that it is possible to improve teamwork through relatively low effort, by using peer feedback tools that already exist; if attempting to more closely monitor team performance, Hernández-García et al. (2018) have shown this to be possible, with caveats such as the need to verify that LMS/wiki users correspond to the actual students performing work. Another approach may be to transfer existing students chats used for collaborative activities to a monitored environment, as Ferguson and Shum (2011) have shown the possibility of analysing real time chat, though students may not be willing considering privacy issues or the need to discuss among themselves.

It has been shown that providing feedback, at least in the form of a learning analytics dashboard available to students, can be detrimental to their performance (Lonn et al., 2015); therefore, availability of LA output should be carefully evaluated, and there should be monitoring, if possible, of when a teacher provides feedback so that effective types of feedback can be found.

5. Conclusion

Predictors used in learning models in online learning can be applied to the traditional classroom. Analogues may be found for predictors that are not available in the conventional classroom. Common characteristics and categorisation of predictors may be used to identify predictors among existing data, including data provided to students (e.g. formative feedback) that is not captured by existing learning management systems used. As a conventional classroom may have small cohorts, approaches that work on smaller groups should be preferred and personalized training of models should be attempted. An LMS applied to a conventional classroom would ideally require minimum additional effort, therefore the data collected should either be simple to obtain, support the users in other ways, or be provable as providing beneficial analytics.

List of abbreviations

EDM	Educational Data Mining
LA	Learning Analytics
LMS	Learning Management System
MOOC	Massive Open Online Courses

References

Blikstein, P., Worsley, M., Piech, C., Sahami, M., Cooper, S., Koller, D. (2014). Programming Pluralism: Using Learning Analytics to Detect Patterns in the Learning of Computer Programming. Journal of the Learning Sciences, 23, 561-599. doi: https://doi.org/10.1080/10508406.2014.954750

Donia, M. B., O'Neill, T. A., Brutus, S. (2018). The longitudinal effects of peer feedback in the development and transfer of student teamwork skills. Learning and Individual Differences, 61, 87-98. doi:https://doi.org/10.1016/j.lindif.2017.11.012

Ezen-Can, A., Grafsgaard, J. F., Lester, J. C., Boyer, K. E. (2015). Classifying Student Dialogue Acts with Multimodal Learning Analytics. Proceedings of the Fifth International Conference on Learning Analytics And Knowledge (pp. 280–289). New York, NY, USA: Association for Computing Machinery. doi: https://doi.org/10.1145/2723576.2723588

Ferguson, R., Shum, S. B. (2011). Learning Analytics to Identify Exploratory Dialogue within Synchronous Text Chat. Proceedings of the 1st International Conference on Learning Analytics and Knowledge (pp. 99–103). New York, NY, USA: Association for Computing Machinery. doi: https://doi.org/10.1145/2090116.2090130

Fernandes, E., Holanda, M., Victorino, M., Borges, V., Carvalho, R., Erven, G. V. (2019). Educational data mining: Predictive analysis of academic performance of public school students in the capital of Brazil. Journal of Business Research, 94, 335-343. doi:https://doi.org/10.1016/j.jbusres.2018.02.012

Fulantelli, G., Taibi, D., Arrigo, M. (2015). A framework to support educational decision making in mobile learning. Computers in Human Behavior, 47, 50-59. doi:https://doi.org/10.1016/j.chb.2014.05.045

Gašević, D., Dawson, S., Rogers, T., Gasevic, D. (2016). Learning analytics should not promote one size fits all: The effects of instructional conditions in predicting academic success. The Internet and Higher Education, 28, 68-84. doi:https://doi.org/10.1016/j.iheduc.2015.10.002

Hernández-García, Á., Acquila-Natale, E., Chaparro-Peláez, J., Conde, M. Á. (2018). Predicting teamwork group assessment using log data-based learning analytics. Computers in Human Behavior, 89, 373-384. doi:https://doi.org/10.1016/j.chb.2018.07.016

Hone, K. S., Said, G. R. (2016). Exploring the factors affecting MOOC retention: A survey study. Computers & Education, 98, 157-168. doi:https://doi.org/10.1016/j.compedu.2016.03.016

Jo, I.-H., Kim, D., Yoon, M. (2015). Constructing Proxy Variables to Measure Adult Learners' Time Management Strategies in LMS. Journal of Educational Technology & Society, 18, 214-225. Retrieved from http://www.jstor.org/stable/jeductechsoci.18.3.214

Jovanović, J., Gašević, D., Dawson, S., Pardo, A., Mirriahi, N. (2017). Learning analytics to unveil learning strategies in a flipped classroom. The Internet and Higher Education, 33, 74-85. doi:https://doi.org/10.1016/j.iheduc.2017.02.001

Leeuwen, A., Janssen, J., Erkens, G., Brekelmans, M. (2015). Teacher regulation of cognitive activities during student collaboration: Effects of learning analytics. Computers & Education, 90, 80-94. doi:https://doi.org/10.1016/j.compedu.2015.09.006

Lonn, S., Aguilar, S. J., Teasley, S. D. (2015). Investigating student motivation in the context of a learning analytics intervention during a summer bridge program. Computers in Human Behavior, 47, 90-97. doi:https://doi.org/10.1016/j.chb.2014.07.013

Ma, J., Han, X., Yang, J., Cheng, J. (2015). Examining the necessary condition for engagement in an online learning environment based on learning analytics approach: The role of the instructor. The Internet and Higher Education, 24, 26-34. doi:https://doi.org/10.1016/j.iheduc.2014.09.005

Maldonado-Mahauad, J., Pérez-Sanagustín, M., Kizilcec, R. F., Morales, N., Munoz-Gama, J. (2018). Mining theory-based patterns from Big data: Identifying self-regulated learning strategies in Massive Open Online Courses. Computers in Human Behavior, 80, 179-196. doi:https://doi.org/10.1016/j.chb.2017.11.011

Ochoa, X., Worsley, M. (2016). Editorial: Augmenting Learning Analytics with Multimodal Sensory Data. Journal of Learning Analytics, 3(2), 213-219. https://doi.org/10.18608/jla.2016.32.10

Pardo, A., Kloos, C. D. (2011). Stepping out of the Box: Towards Analytics Outside the Learning Management System. Proceedings of the 1st International Conference on Learning Analytics and Knowledge (pp. 163–167). New York, NY, USA: Association for Computing Machinery. doi: https://doi.org/10.1145/2090116.2090142

Prasad, D., Totaram, R., Usagawa, T. (2016, 9). Development of Open Textbooks Learning Analytics System. The International Review of Research in Open and Distributed Learning, 17. doi: https://doi.org/10.19173/irrodl.v17i5.2541

Ramesh, A., Goldwasser, D., Huang, B., Daumé, H., Getoor, L. (2014, 6). Understanding MOOC Discussion Forums using Seeded LDA. Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications (pp. 28-33). Baltimore: Association for Computational Linguistics. doi: https://doi.org/10.3115/v1/W14-1804

Rienties, B., Toetenel, L. (2016). The impact of learning design on student behaviour, satisfaction and performance: A cross-institutional comparison across 151 modules. Computers in Human Behavior, 60, 333-341. doi:https://doi.org/10.1016/j.chb.2016.02.074

Schneider, B., Blikstein, P. (2015). Unraveling Students' Interaction Around a Tangible Interface using Multimodal Learning Analytics. JEDM | Journal of Educational Data Mining, 7(3), 89-116. https://doi.org/10.5281/zenodo.3554729

Tabuenca, B., Kalz, M., Drachsler, H., Specht, M. (2015). Time will tell: The role of mobile learning analytics in self-regulated learning. Computers & Education, 89, 53-74. doi:https://doi.org/10.1016/j.compedu.2015.08.004

Veeramachaneni, K., O'Reilly, U.-M., Taylor, C. (2014). Towards Feature Engineering at Scale for Data from Massive Open Online Courses. arXiv preprint https://arxiv.org/abs/1407.5238

Wang, X., Yang, D., Wen, M., Koedinger, K. R., Rosé, C. P. (2015). Investigating How Student's Cognitive Behavior in MOOC Discussion Forum Affect Learning Gains. In O. C. Santos, J. Boticario, C. Romero, M. Pechenizkiy, A. Merceron, P. Mitros, J. M. Luna, M. C. Mihaescu, P. Moreno, A. Hershkovitz, S. Ventura & M. C. Desmarais (eds.), EDM (p./pp. 226-233), : International Educational Data Mining Society (IEDMS). https://eric.ed.gov/?id=ED560568

Worsley, M., Blikstein, P. (2010). Towards the development of learning analytics: Student speech as an automatic and natural form of assessment. In Annual Meeting of the American Education Research Association (AERA). http://www.marceloworsley.com/papers/aera_2011.pdf

Yang, T., Brinton, C. G., Joe-Wong, C., Chiang, M. (2017, 8). Behavior-Based Grade Prediction for MOOCs Via Time Series Neural Networks. IEEE Journal of Selected Topics in Signal Processing, 11, 716-728. doi: https://doi.org/10.1109/JSTSP.2017.2700227

Zacharis, N. Z. (2015). A multivariate approach to predicting student outcomes in web-enabled blended learning courses. The Internet and Higher Education, 27, 44-53. doi:https://doi.org/10.1016/j.iheduc.2015.05.002

Received February 20, 2020, accepted March 14, 2020