ADV MATH SCI JOURNAL
Advances in Mathematics: Scientific Journal **9** (2020), no.6, 4039–4046
ISSN: 1857-8365 (printed); 1857-8438 (electronic)
https://doi.org/10.37418/amsj.9.6.85 Spec Issiue on ICAML-2020

PREDICTIVE ANALYTICS IN MOOCS: A-REVIEW

JEEVAN BHATIA[1], HARJEET SINGH, AND AKSHAY GIRDHAR

ABSTRACT. To make the MOOCs (Massive Open Online Courses) more efficacious, automated systems that can support predictive evaluation mechanisms are required. The key significant issue is to address the problem of identifying lagging areas of students early enough, so that some measurable remedial actions can be taken on time. Technology-based teaching methodologies not only improve students' learning outcomes, however also proved as rich source of data capable to address this issue. The existing assessment methodologies inculcated in MOOCs merely focus on maximizing the accuracy of prediction tools, rather than considering the timely and personalized predictions. This paper throws the light on importance of incorporating grade prediction algorithms in MOOCs to predict final grades, whereby addressing the issue of personalized and timely prediction, when the algorithms can acquire maximum expected accuracy. In these scenarios, algorithms learn by themselves regarding the optimal prediction and relevant time to do this accurately. Here, the confidence estimate plays a major role in judging the prediction accuracy. Researches validate that it is advisable to make students participate in early assessment tasks to ensure timely performance prediction, thereby aiding necessary interventions through recommender system.

[1]*corresponding author*

2010 *Mathematics Subject Classification.* 68T05.

Key words and phrases. grade prediction algorithms, online learning, learning analytics, forecasting algorithms, outcome-based model.

1. Introduction: the MOOCs and predictive analytics

Nowadays, there is a paradigm shift in online education. According to Chen et al. [1], knowledge is becoming readily available with the advent of MOOCs (Massive Open Online Courses). There are different active-learning practices available that provide rich data, such as Peer Instructions, Activities, Assignments, Video Streaming and Objective Evaluation. Some researches associate's performance on these pedagogies with student learning outcomes and thereby, predict students who are at the risk. These new technologies help in giving personalized learning, and optimal support system to teachers. With rapid increase in MOOCs phenomenon, it is impossible for teachers to keep performance track of individual student. As a result, there can exist some students, who failed but would have passed if timely predictions were made and remedial actions were suggested to such students. These could consist of additional study material, personalized lectures and resources. Therefore, it is pivotal to predict student performance before the course finishes. This ensures the great need to design an efficient prediction algorithm that can find the best time to predict student's grade, by virtue of which timely interventions can be made for low performing students as emphasised by Moreno-Marcos et al. in [2, 3].

2. Key essential concepts

2.1. **Prediction variables.** According to Ruiprez-Valiente et al. [4], independent variables have been used in designing of prediction models. These are:

(i) Presage variables: The parameters that are available or can be determined before the commencement of algorithm. These are:
    (1) Performance Variables: high-school GPA, age, gender
    (2) Attitude Variables: motivation/engagement level, study habits.

(ii) In-Progress Variables: These are measured using results gathered from ongoing assessments, such as assessment variables of tests.

2.2. **Challenges in Time- Based Grade Prediction Algorithm (TGPA).** There are various challenges involved in incorporating TGPA. These are:

(i) As stated by Lemay and Doleck [5], due to inconsistent motivation level in students during the whole course timeframe, it is difficult to correlate the test scores with performance.

(ii) Though syllabus remains same, marking rubrics of assessment tools change over the years, thereby necessitating the need to redesign assessment pedagogy and algorithm parameters.

(iii) Lemay and Doleck [6] argued that student's predictability is different due to background diversifications. For instance, it is advisable to make predictions based on initial performances, but some students demand over the time performance analysis to get equally relevant results.

These challenges demand that time for doing the predictions should be analyzed for each student individually rather than altogether [7].

3. Recent trends in educational activities

Distinct studies [8–10] analyzed the importance of standardized tests and grade prediction algorithms (GPA) for investigating academic success in online learning programs. They advocate correlation between predictors and success measures. Researches [11, 12] show that there exist other factors that result in strong correlation among GPA predictors, like immutable elements that student possess at the beginning. Additionally, other factors have been added in recent researches such as behavior, attitudes, online study time, self-efficacy, video repitition and assessment tools' data. However, some of the data is difficult to be modelled at multiple MOOCs. Researchers [13, 14] reciprocates that non-linear complex patterns have not been found in these streams and even advocated that with variety, accuracy suffers while the validity increases. Different models are being used to do predictions from general statistical models (simple correlation) to specific machine learning models (random forests, nearest neighbor, clustering, classification, neural networks, decision trees, regression and support vector machines) [15].

Qu and Chen [16] developed VisMOOC, a comprehensive data analytical tool providing detailed insights. Among all the click events, a seek event appeared most interesting as it represents the skipped or re-watched part. Here, course and video level provides insights about video popularity and demographic student distribution, thereby providing concerned temporal information. Dashboard view provides option to do social network and sentiment analysis. However, despite of varied features to analyze learning behaviors, there are some limitations. Firstly, there is a support lag for on-the-fly analysis of streaming

web log data. Secondly, particular user group analysis helps instructors to design tailor-made content according to user needs. However, currently there is no scope of differential analysis rather it is done on collective user group. Lastly, there is a need of predictive analytics to deal with the issue of low retention rate and high dropout probability.

Fu et al. [17] conducted a design study on iForum (interactive visual analytics system) to discover temporal and structural patterns using massive heterogenous data extracted from MOOCs. It reveals all the thread information by minimizing the screen real estate, along with overcoming the limitation of lengthy thread discussions called Thread River. Additionally, to provide ease in exploration, all visualization views are interactively coordinated. User profile level data is also accumulated and processed. Data can be explored using distinct multiple levels namely, macroscopic, mesoscopic and microscopic. Complete temporal dynamics are retrieved at macroscopic level. This includes trends related to topic posts, threads and users' volume/lifespans. Mesoscopic level supports matrix-based visualization of distinct user group and is equipped with interactive filtering mechanisms. At microscopic level, interested data subsets are detailed in three views. An in-depth analysis is done to check the performance on real world datasets. Visual clutter is reduced by aggregating responses under same post, that is certainly presented in Thread Arcs. For extending the design scalability, focus+context approach is implemented to visualize lengthy threads effectively. Furthermore, this system is also helpful in understanding the between and within group user social interactions.

Yang et al. [18] used time series neural networks to do behavioural prediction analysis with clickstream and grade datasets. These learn both from prior performance and data and are processed to compute input features (overcoming data sparsity). Instructors used feature distribution, model quality and predictive values to take remedial actions using GUI dashboard. To check the prediction accuracy and the relative gain, Root mean square error (RMSE) is used, and is compared with average historical performance and linear regression. The designed algorithms namely, FTSNN (assessment features only) and IFTSNN (behavioral and assessment features both) are also compared with naive baseline (NB) and lasso regression (LR). The proposed non-linear algorithm outperform NB and LR by the percentage of 60 and 15 respectively. Moreover, the insights reflect higher gains in beginning and increases the probability for instructors to

handle student challenges in early phases by feeding these algorithms in recommender system. The research throws light on collaboration of behavior features with other parameters and advocated that they are not correlated solely with performance.

Meier et al. [19] designed a system to optimally do prediction based on the past history of performances. The research work demonstrated that for 85 percent students, algorithm has given 76 percent accuracy in predicting after fourth week. Algorithm can be used in two modes, i.e., in regression settings, where overall grade can be predicted, and in the classification settings, where performance can be predicted in two groups (well performed/poorly). Algorithm is being continuously trained from the past years, hence, as more data is retrieved, it gives more accurate performance. It has been demonstrated that robustness increases when course is taught by different instructors. Algorithm has been compared with the benchmark prediction methods (Linear/Logistic regression and k-Nearest Neighbors) over the same datasets. At last, preferred way of designing the course is elaborated. The simulations of this research claims that simple linear methods provide similar accuracy as the complex methods. As this algorithm run on students who have already completed the course, therefore the impact of timely interventions is not analyzed. This algorithm can be extended to perform multiple predictions per student.

Liao et al. [20] proposed a robust machine learning technique to identify low performing students in advance. The proposed work also defines a modelling methodology that helps in predicting final grades using the clickstream data collected by instructors from peer instruction pedagogy. To accomplish this task successfully, model uses support vector machine binary classifier to predict next terms outcomes. Here, binary classification is performed in context to whether or not student is going to give expected performance. To train the dataset, prior term's student clicker data and final exam grades are used, which is further applied on next terminal data in the first three weeks. This model has been implemented on five distinct computer science courses, taught at different institutions by three instructors. The major research strengths are doing timely predictions and using lightweight student data. The results shown 62 percent accuracy in predicting low performing students out of instructor-determined driven 40 percent class bottom cut-off.

TABLE 1. Comparative analysis of related work

	[13, 14]	[15, 16]	[17–19]	[20–22]
Aim of Paper	Find feature set	Grade Prediction	Grade Prediction	Answer Accuracy
Predictors (Course-C, Other-O)	O	C, O	C, O	C
Previous year data	NA	No	Yes	Yes
Early Performance variables	NA	No	Yes	Yes
Technique (Regression-R Classification-C)	NA	R	Both	C
Modelling (Real-world-R Cross-institution-CI)	NA	R	CI	R

Table 1 summarises the comparative analysis of related work.

The major limitation of previously implemented GPAs is that they don't fit in all educational settings, especially in online learning environment. Some variables are not readily available due to inaccessibility and privacy concerns. Some researchers do descriptive modelling, whereby studying the relationships between independent and dependent variables on the training datasets. This is not fruitful for future predictions rather can be used to identify past relationships. Thus, it is the necessity to use predictive modelling, wherein the test data is merely used for checking the model accuracy. Accuracies of such models falls under two broad categories, i.e., single accuracy and area under the curve ROC (Receiver Operating Curve). Single accuracy depicts the percentage of students at risk. On the contrary, ROC curve quantifies the student percentage who are at the risk and are correctly classified [21–23].

4. CONCLUSION

Instructors and students both are at benefit, if prediction models identify students at risks early enough. As via this, instructors are able to recognize such students and learn the areas where they need more help, so that they can improve their course outcomes in time. Although extensive research is carried out

on MOOCs and the learning behavior of students, yet limited literature is present on timeliness and accuracy both in regression and classification prediction settings. Usually, robustness of predictions decreases when the MOOCs courses are taught by different instructors in online learning environment. Hence, it can be concluded that to articulate detailed insights, it would be beneficial to do multiple performance predictions after needful interventions. These insights would be helpful in analyzing the trend of predicted grades.

REFERENCES

[1] X. CHE, H. YANG, C. MEINEL: *Automatic online lecture highlighting based on multimedia analysis*, IEEE T. Learn. Technol, **11**(1) (2018), 27–40.

[2] P. M. M. MARCOS, P. J. M. MERINO, J. M. MAHAUAD, M. P. SANAGUSTÍN, C. A. HOYOS, C. D. KLOOS: *Temporal analysis for dropout prediction using self-regulated learning strategies in self-paced moocs*, Comput. Educ., **145**(2020).

[3] P. M. M. MARCOS, T. C. PONG, P. J. M. MERINO, C. D. KLOOS: *Analysis of the factors influencing learners' performance prediction with learning analytics*, IEEE Access, **8**(2020), 5264–5282.

[4] J. A. R. VALIENTE, S. HALAWA, R. SLAMA, J. REICH: *Using multi-platform learning analytics to compare regional and global mooc learning in the arab world*, Comput. Educ., **146**(2020).

[5] D. J. LEMAY, T. DOLECK: *Grade prediction of weekly assignments in moocs: mining video-viewing behavior*, Educ. Inf. Technol., **25**(2019), 1333–1342.

[6] D. J. LEMAY, T. DOLECK: *Predicting completion of massive open online course (mooc) assignments from video viewing behavior*, Interact. Learn. Environ., (2020), 1–12.

[7] R. S. NISHA, R. RADHA: *A systematic analysis of data-intensive moocs and their key challenges*, 3rd Int. Conf. ICCCT, IEEE, (2019), 245–252.

[8] J. GARDNER, C. BROOKS, R. BAKER: *Evaluating the fairness of predictive student models through slicing analysis*, Proc. 9th Int. Conf. LAK, (2019), 225–234.

[9] G. ALEXANDRON, L. Y. YOO, J. A. R. VALIENTE, S. LEE, D. E. PRITCHARD: *Are mooc learning analytics results trustworthy? With fake learners, they might not be!*, Int. J. Artif. Intell. Educ., **29**(4) (2019), 484–506.

[10] K. MONGKHONVANIT, K. KANOPKA, D. LANG: *Deep knowledge tracing and engagement with moocs*, Proc. 9th Int. Conf. LAK, (2019), 340–342.

[11] K. SHARMA, P. DILLENBOURG, M. GIANNAKOS: *Stimuli-based gaze analytics to enhance motivation and learning in moocs*, IEEE ICALT 2019, **2161**(2019), 199–203.

[12] M. E. A. MENCÍA, C. A. HOYOS, C. D. KLOOS: *Chrome plug-in to support srl in moocs*, Springer EMOOCs, (2019), 3–12.

[13] T. DOLECK, D. J. LEMAY, R. B. BASNET, P. BAZELAIS: *Predictive analytics in education: a comparison of deep learning frameworks.*, Educ. Inf. Technol., (2019), 1–13.

[14] P. LIN, A. WOODERS, J. T. WANG, W. M. YUAN: *Artificial intelligence, the missing piece of online education?*, IEEE Eng. Manage. Rev., **46**(3) (2018), 25–28.

[15] M. YOUSSEF, S. MOHAMMED, B. F. WAFAA, ET AL.: *A predictive approach based on efficient feature selection and learning algorithms' competition: Case of learners' dropout in moocs*, Educ. Inf. Technol., **24**(6) (2019), 3591–3618.

[16] H. QU, Q. CHEN: *Visual analytics for mooc data*, IEEE Comput. Graph. Appl., **35**(6) (2015), 69–75.

[17] S. FU, J. ZHAO, W. CUI, H. QU: *Visual analysis of mooc forums with iforum*, IEEE Trans. Vis. Comput. Graph., **23**(1) (2017), 201–210.

[18] T. YANG, C. G. BRINTON, C. J. WONG, M. CHIANG: *Behavior-based grade prediction for moocs via time series neural networks*, IEEE J-STSP, **11**(5) (2017), 716–728.

[19] Y. MEIER, J. XU, O. ATAN, M. VAN DER SCHAAR: *Predicting grades*, IEEE T. SIGNAL. PROCES., **64**(4) (2016), 959–972.

[20] S. N. LIAO, D. ZINGARO, K. THAI, C. ALVARADO, W. G. GRISWOLD, L. PORTER: *A robust machine learning technique to predict low-performing students*, ACM Trans. Comput. Educ., **19**(3) (2019), 1–19.

[21] J. BHATIA, A. GIRDHAR, I. SINGH: *An automated survey designing tool for indirect assessment in outcome based education using data mining*, 5th IEEE Int. Conf. MITE., (2017), 95–100.

[22] K. KIILI, H. KETAMO: *Evaluating cognitive and affective outcomes of a digital game-based math test*, IEEE T. Learn. Technol., **11**(2) (2018), 255–263.

[23] J. BHATIA, H. SINGH: *Indirect assessment of outcomes in education a-review*, Int. J. Comput. Sci. Eng., **5**(2017), 273–278.

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
PUNJABI UNIVERSITY
PATIALA (PUNJAB), INDIA
Email address: jeevankaur20@gmail.com

PG. DEPARTMENT OF COMPUTER SCIENCE
MATA GUJRI COLLEGE
FATEHGARH SAHIB (PUNJAB), INDIA
Email address: zrjeet@gmail.com

DEPARTMENT OF INFORMATION TECHNOLOGY
GURU NANAK DEV ENGINEERING COLLEGE
LUDHIANA (PUNJAB), INDIA
Email address: akshay1975@gmail.com