Predicting Student Performance from Their Behavior in Learning Management Systems

Parisa Shayan and Menno van Zaanen

***Abstract*—Nowadays, Information and Communication Technology (ICT) provides an opportunity to discover new knowledge and create a desirable learning environment. That is why the influence of ICT on education is irrefutable. Technology has changed the learning styles: the way people prefer to learn and improve the quality of their learning. Physical and online classes can be held concurrently so that lecturers and students can interact via learning management systems. A Learning Management System (LMS) is an application software that plays a significant role in educational technology. Such software can be designed to augment and facilitate instructional activities including registration and management of education courses, analyzing skill gaps, reporting, and delivery of electronic courses concurrently. Since all information and corresponding data are recorded and monitored in the LMS, it can provide an accurate insight into student's online behavior. In general, measuring student performance is an important part of the education system. The fields of learning analytics and educational data mining both emphasize the analysis of educational data in order to improve teaching and learning styles as well as to predict student performance. In the current study, we use data from the Moodle LMS from a collection of courses from a single institution to identify weak/strong students during the course. The result has to be interpretable and understandable as the aim is to give this information to lecturers, who may use the information to improve their course and identify students who may need special attention.**

***Index Terms*—Data mining, clustering, decision tree, rule-based, student.**

I. Introduction

Large datasets of students' properties including academic and educational backgrounds in most educational institutions are available. Finding a predictive model in this information can help educational institutions improve learning processes such as assessment, recognition of academic status and counseling. Learning analytics and educational data mining extract understandable, useful, unknowable, valid and exquisite patterns from large datasets. In addition, hidden patterns can help educational institutions in better decision making and have a more advanced plan for students learning process.

In this research, we are going to predict students' performance from their online behavior. If we want to know which students need extra help or an extra challenge during a course, we need to be able to identify them. This is possible when the lecturer knows all students but even then, it may be difficult in large groups. That is why we look at LMS data. However, in this study, we consider the LMS as a platform that provides online courses for educational institutions using communication and administrative tools [1]. As an LMS forms a rich source of data including all stored and recorded actions, this data can be used in tools that analyze and predict students' behavior and performance [2]. One of the main reasons for attempting to predict student performance is that it allows lecturers to take immediate action when needed: weak students may be identified and given additional training, whereas high performing students may be challenged or helped more if needed. The ultimate aim is to accomplish a higher level of quality in education and more personalized education, even in large groups.

Here, we present research that aims to provide a predictive model of student performance based on their online behavior in order to help lecturers to identify which students may require special attention. To be useful in a more generic way, the predictive model should not only be accurate, it should also provide information that indicates which properties of student behavior have an impact on their performance. So, the lecturer understands certain students who need either more help or need to be challenged more. For this purpose, we investigate data mining models to evaluate students' performance and gain insight into decision making. We investigate the prediction of students' performance using decision tree J48 and ID3 algorithms to be able to interpret the results.

Our main contributions in this study are as follows:

- Develop systems that can identify groups of students based on their online behavior during the course (formative assessment).
- Learning analytics and educational data mining can be used to extract understandable and useful patterns of predicting students' performance and decision making about educational courses.
- Results of the systems can identify weak/strong students in order to provide extra training for weak students, and challenging tasks for strong students.

II. Background

A. Learning Analytics and Educational Data Mining

Educational Data Mining (EDM) pertained to Developing methods to search for unique data from educational settings

Manuscript received August 10, 2018; revised September 1, 2018.This work was supported in part by the School of Humanities and Digital Sciences, Tilburg University, Netherlands and Organizing Committee of ICFL 2018, University of Barcelona, Spain.

Parisa Shayan is with School of Humanities and Digital Sciences, Tilburg University, Netherlands (e-mail: P.Shayan@uvt.nl).

Menno van Zaanen is with School of Humanities and Digital Sciences, Tilburg University, Netherlands (e-mail: M.M.vanZaanen@uvt.nl).

doi: 10.18178/ijiet.2019.9.5.1223

and use those methods to better understand students and settings they learn [3]. Learning Analytics and Knowledge (LAK) refers to measuring, collecting, analyzing, and reporting data on student progress in areas where learning occurs [4]. In EDM, the advanced data mining techniques are used to automatically explore learner models and adapt the learning environment. In contrast, the LAK often use statistical analyzes, which are the result of models that mainly inform teachers about their learning progress [3], [4]. Despite some differences, both EDM and LAK focus on improving teaching process and learning style [3].

In general, Student performance prediction is the major focus of LAK and EDM [5]. However, since in formative assessment, intervention requires knowing student performance performance and also not all features are available, most LAK and EDM research only used LMS data for summative assessment [5]-[9]. Whereas some other research shows that student characteristics and past performance have higher predictive value than LMS data [10], [11]. In general, the previous studies typically focused on the pass/fail prediction in summative assessment, while in this research we concentrate on student performance in formative assessment. In addition, we include both student characteristics and past performance as well as LMS data to determine which properties have a large impact on prediction students' performance during the course.

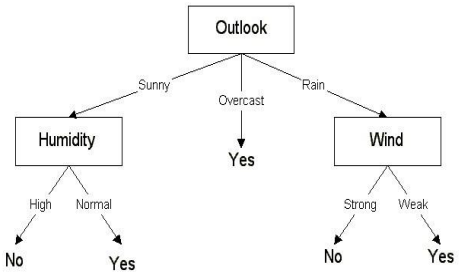

Fig. 1. Example of ID3.

B. Classification and Regression Algorithms

Classification and regression are both related to the prediction, where regression predicts a value from a continuous set, whereas classification predicts the belonging to the class. Classification models can predict a class and Regression models result in continuous values. In other words, a classifier has been made to predict definitive labels and a regressor will be created that predicts a continuous-valued-function. The machine learning classifiers and predictors try to find regularities in samples, so they can predict unseen data points [12].

In the current study, decision tree used as the preferable classifiers and entropy as a measuring tool to interpret the results. A decision tree is a graphical display of a particular decision condition that is used when a complex gap occurs in a structured decision process. Each rule in a decision tree is displayed by tracking a series of paths from root to node to the next node and so on until an action is performed. The main benefit of decision trees is that the most important properties in the data are found in the upper nodes in the tree structure, whereas the marginal properties are set aside. Yet they are either used if useful or not.

Most algorithms developed for decision tree learning are versions of a basic algorithm that uses greedy and top-down approach to search for possible decision trees. This approach is known as the ID3 algorithm and its inheritor C4.5. Our basic algorithm, or the ID3, finds a decision tree with a top-down search. As Fig. 1 shows, the ID3 creates a decision tree from a fixed set of instances. The resulting tree is used to classify future samples. The below example has multiple attributes and belongs to a class (e.g. yes or no).

The most important choice in the ID3 algorithm is the selection of features to test in each tree. Preferably, the selected feature should help most in distinguishing the classes. In other words, the ID3 algorithm is providing the most useful characteristics at each node in the tree. To do this, ID3 uses a statistical property called the information gain to provide information on the reduction of entropy and the amount of noise or uncertainty in the dataset to be classified.

Two classification algorithms: decision tree J48 and ID3 were run with WEKA and R respectively. They provide a human readable result, which in many cases display the information in a comprehensive manner [13], [14].

III. Methodology

A. Population and Sample

For this research, data including student and course characteristics, behavioral and performance data from Moodle LMS, were collected from blended courses at Eindhoven University of Technology (The Netherlands) in the first two quarters (fall and winter) of the academic year (2014– 2015). All courses were blended courses, as the course including four to six hours of face-to-face lectures per week and part of the course presented online in Moodle LMS.

The dataset contains data from a total number of 426 students. As some students participated in multiple courses (32 students followed one course, 326 followed two courses, and 68 followed three courses), this resulted in a total of 888 students in five courses. The five courses included were: Calculus A, Calculus B, Applied Physical Sciences formal, Applied Physical Sciences conceptual, and Introduction to Psychology and Technology. Data from Moodle LMS comes from a previous research Conijn *et al.* [11]. Courses data were collected in the fall of 25th of August 2014 (1 week before the beginning of the lecture) until 9th of November 2014 (end of the test week) and grouped each week, which led to 11 weeks of data. Courses data in the winter quarter were collected also from 3rd of November 2014 (1 week before the beginning of the lecture) until 1st of February 2015 (end of the test week). As the two-weekly Christmas holidays fell in the quarter, a total of 13 weeks of LMS data was obtained.

B. LMS Data

The data collected from the Moodle LMS has been used in a previous study [11], which focused on predicting student performance on a fail/pass level, whereas in this research, we are going to predict more fine-grained students' performance. Hence due to the dissimilar nature of the two studies, no direct comparison is possible.

As can be seen from Table I, The LMS variables are incorporated with prior performance data, course and student characteristics. The four collected events which often used in literature were extracted: the total number of clicks students done, the number of course page views students done, the number of online sessions students participate, and the total time students were online. The Collected events are grouped each week to show activity levels over a specific week in the course. In addition, five variables relevant to the study patterns consisted of: the irregularity of study time (S.D. of time per session), the irregularity of study interval (S.D. of time between sessions), the largest period of inactivity, the time until the first activity, and the average time per session. A more detailed information of these variables can be found in [11].

C. Data Performance and Data Analysis

In the current study, the collected data for all 888 samples contained in-between assessment grade and final exam grade. All grades are from 0 to 10, where grade above or equal to 5 indicates that a student passed the course and grade less than 5.5 represents a fail. Two classification algorithms were implemented with WEKA and R respectively: decision tree J48 and ID3 algorithms. For classification, four attribute sets were used: course and student characteristics, in-between assessment grade (Midterm), and LMS data. Since there was direct relationship between students 'Grade' and their performance 'Grade' was considered as target variable for classification: grade≤3 (bad performance), 3≤grade≤5 (particularly bad performance), 5≤grade≤7 (particularly good performance), and 7≤grade≤9 (good performance).

The final exam scores were very low (M = 5.31, S.D. = 2.10): The average students were not able to pass the course. The in-between assessment scores were significantly higher (M = 6.93, S.D. = 1.33). In-between assessment grade includes scores for the rating graded over the course (i.e., entrance exam, assignments, online and offline assignments as well as midterm exam). These evaluations can be done online via Moodle LMS or offline and manually or through other systems. In addition, according to the Fig. 2 and 3, up to two weeks to the midterm, the number of online sessions, the number of views, and the number of clicks had the highest information gain respectively (sessions= 0.030, views= 0.025, clicks= 0.020). Also, after the midterm test, the last two weeks before the final exam (at week 8) information gain for the above features has reached its peak. With the difference that the number of clicks as well as the number of views has gone up (clicks= 0.028, views= 0.027) while the number of sessions has been significantly reduced to less than 0.025. Maybe this is because, towards the end of the semester, while not session yet, students have to do their final assignments. That is why, they need to be in touch more with lecturer and classmates via LMS Forum to do their assignments and projects.

$$Entropy(S) = \sum_{i=1}^{c} -p_i \log_2 p_i$$

$$Gain(S, A) \equiv Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

Information gain works well for most cases, unless you have a few variables that have a large number of values or classes. Information gain somehow is biased towards choosing attributes with many values as root nodes. Whereas gain ratio is a modification of information gain that reduces its bias and is usually the best option. It corrects information gain by taking the intrinsic information of a split into account. As can be seen from Fig. 3, there was a similar trend for the gain ratio either with the difference that this time the number of views with more than 0.15 had the most gain ratio and then this were followed by the number of clicks and the number of sessions respectively at week 3 (two weeks before midterm). Moreover, regardless the last two weeks before final term, we see an extremely descending trend for all features. As described above, this means that there is no student view or click unless they are forced to do.

IV. Results (Interpretability)

Although the rule-base algorithm is considered as one of the best classifiers to predict student performance, in this research, the decision tree J48 and ID3 classification algorithms lead to better results to be interpreted particularly and they also can provide more insight in which variables are useful for predicting student performance. Hence, these classifications are preferred. As the decision tree J48 led to unclear results and this is not easy to read yet, we only report the ID3 algorithm (information gain and gain ratio) here.

According to the Fig. 2 and 3, the information gain and the gain ratio showed that the number of views, the number of clicks, and the number of sessions were the most important features on predicting students' performance. While these attributes are more considered at the beginning and the end of course by students than in between (during the course). In addition, as can be seen from Table I, prior GPA (gain ratio=0.1071, information gain=0.0679) and midterm grade (gain ratio=0.0949, information gain=0.0839) in the course were also found the important features for the prediction students' performance. In general, it seems that the previous performance of the students should be considered more. Furthermore, as can be seen from Fig. 3, lecturers should force on students to put more effort during the course with more participation in online sessions as well (after week 5).

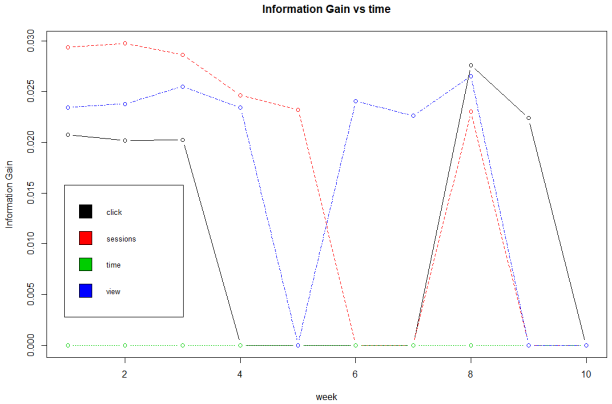

Fig. 2. Information gain vs weeks on predicting the most impressive Feature sets during the course.

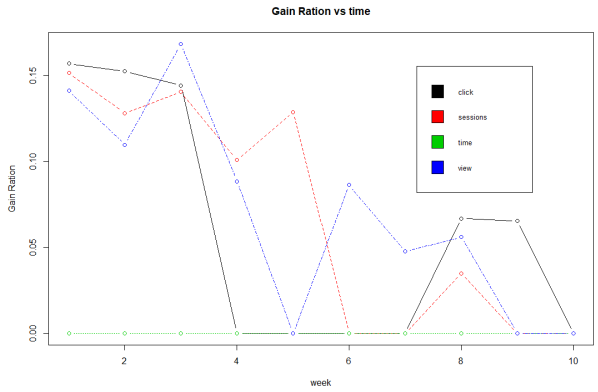

Fig. 3. Gain ratio vs weeks on predicting the most impressive Feature sets during the course.

V. Discussion

The aim of this research was to predict the students' performance during the course from their behavior using a learning management system. Online student behavior is accessible data to investigate students' performance. It means that it is attempted to track all events (clicks, views, sessions) registered in LMS for predicting students' performance in order to reinforce the positive behaviors and work on tackling the negative ones. Hence our purpose was to figure out exactly which features have a large impact on the prediction of students' performance during the course and how we can give the properties in a meaningful and understandable way to lecturers.

Whereas in previous studies [5]-[9] mostly students' performance was measured in summative assessment after the course, in the current study we focused on measuring students' performance in formative approach. In addition, the findings from some researchers [10], [11] showed that student characteristics and past performance have higher predictive value than LMS data. That is why, in the current study, we included student characteristics and past performance when predicting students' performance but with other machine learning classifiers and in formative assessment.

In short, relatively similar results were obtained with the difference that student characteristics particularly past performance (prior GPA) at the beginning of the course had a positive influence on predicting students' final grade. Whereas towards the end of the semester at the second half through the course, when more information is available, prior GPA is less important. Instead, performance data (Midterm grade) as well as the LMS data (the number of mouse clicks, sessions and views) were more important features on the student's final grade and hence predicting pass/fail.

V. Conclusion

In the current study, we analyzed the prediction models of student performance from their online behavior based on Moodle LMS data, in order to enhance learners' achievement. To do this, learning analytics and educational data mining was used to extract understandable and useful patterns. As noted above, the decision tree J48 and ID3 classification algorithms were employed on predicting students' performance. In this research, four classes of features were used: LMS data, student and course characteristics as well as performance data.

TABLE I: Gain Ratio and Information Gain on the Most Impressive Feature Sets

Attribute	Gain Ratio	Information Gain
Midterm Grade	0.0949177	0.0839575
Prior GPA	0.1071662	0.0679591
Total number of clicks	0.1569597	0.0207545
Number of online sessions	0.1515015	0.0293681
Number of course page views	0.1411736	0.0234459
Clicks_week9	0.1569597	0.0207545
Sessions_week9	0.1515015	0.0293681
Views_week9	0.1411736	0.0234459
Clicks_week8	0.1525034	0.0201652
Sessions_week8	0.1280684	0.0297457
Views_week8	0.1096977	0.0237694
Clicks_week7	0.1443519	0.0202131
Sessions_week7	0.140526	0.0286373
Views_week7	0.1683819	0.0254956
Sessions_week6	0.1008233	0.0246347
Views_week6	0.0884159	0.0234022
Sessions_week5	0.1287051	0.0231933
Views_week4	0.0862866	0.0240446
Views_week3	0.0475758	0.0226015
Clicks_week2	0.0668809	0.0275827
Sessions_week2	0.0347917	0.0230368
Views_week2	0.0560153	0.0265159
Clicks_week1	0.0653105	0.0223715
Views_week1	0.0000000	0.0000000
Sessions_week1	0.0000000	0.0000000

As a result, it was found that students' characteristic (prior GPA) had the highest gain ratio in the first halfway through the course and in the second half of the semester, performance data (midterm grade) and LMS data (the number of views, clicks, and sessions) particularly before the midterm and the final exam were the most important features .This means that all these features had an impressive impact on students' performance and lecturers have to pay more attention to this.

To conclude, LMS has created huge changes (availability anytime and anywhere, centralized information, increased communication, costs and time saving) in the education system and the learning process but there are still some challenges in what we can learn and extract from the system. Hence, it is attempted to reach a general conclusion about the online student behavior to identify students who might need additional help or additional challenges during the course. In general, we are trying to use LMS data (behavioral data) to provide lecturers with more information. Therefore, a more fine-grained analysis of the LMS data and the evaluation by lecturers of the interpretability of the results can be beneficial for further information.

References

[1] A. Pirani, "The learning management systems evolution," *Learning Management Systems Evolution*, 2014.
[2] M. Stracke, "Open learning for improving school education, lifelong learning," in *Proc. the Fifth International Conference on e-Learning*, Belgrade, Serbia, pp. 1–6, 2014.
[3] G. Siemens and R. S. J. Baker, "Learning analytics and educational data mining: Towards communication and collaboration," in *Proc. the 2nd International Conference on Learning Analytics and Knowledge*, 2012.

[4] N. P. A. Sclater and J. Mullan, "Learning analytics in higher education: A review of uk and international practice full report," *Tech. Rep., Jisc.*, 2016.
[5] C. Romero and S. Ventura, "Data mining in education," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 3*, vol. 1, pp. 12–27, 2013.
[6] B. Minaei-Bidgoli and W. F. Punch, "Using genetic algorithms for data mining optimization in an educational web-based system," in *Proc. Genetic and Evolutionary Computation Conference*, Springer, Berlin, Heidelberg, pp. 2252–2263, 2003.
[7] L. V. F. C. Morris and S.-S. Wu, "Tracking student behavior, persistence, and achievement in online courses," *The Internet and Higher Education*, no. 3, pp. 221–231, 2005.
[8] A. Zafra and S. Ventura, "Predicting student grades in learning management systems with multiple instance genetic programming," *International Working Group on Educational Data Mining*, 2009.
[9] N. Z. Zacharis, "Multivariate approach to predicting student outcomes in web-enabled blended learning courses," *The Internet and Higher*, vol. 27, pp. 44–53, 2015.
[10] D. T. R. B. Tempelaar and B. Giesbers, "In search for the most informative data for feedback generation: Learning analytics in a data-rich," *Computers in Human Behavior*, vol. 47, pp. 157–167, 2015.
[11] R. S. C. K. A. Conijn and U. Matzat, "Predicting student performance from lms data: A comparison of 17 blended courses using moodle lms," *IEEE Transactions on Learning Technologies*, Springer, Berlin, Heidelberg, pp. 2252–2263, 2017.
[12] T. S. F. G. J. B. G. M. M. Le and G. D. Fatta, "Computationally efficient rule-based classification for continuous streaming data," pp. 21–34, 2014.
[13] K. B. C. Hornik and A. Zeileis, *Open-Source Machine Learning: R Meets Weka*, Springer-Verlag, vol. 24, no. 2, pp. 225–232, 2008.
[14] T. M. Mitchell, *Machine Learning*, McGraw HillScience/Engineering/Math, 1997.

Parisa Shayan is PhD researcher in communication and information sciences, Tilburg University, the Netherlands, 2017-now; the MSc. information and communication technologies, Eastern Mediterranean University, Cyprus, 2014-2016; the BSc. computer engineering, Sheikhbahaee University of Isfahan, Iran, 1996-2002.

She has job experience in national Iranian oil company as Msc Engineer in Communication and Information Systems, 2004-2017 and as IT Assistant at Eastern Mediterranean University (EMU), Cyprus, 2015-2016. Her publications are as follows:

P arisa Shayan, Assoc. Prof. Dr. Ersun İşçioğlu, "An Assessment of Learning Management Systems Acceptancein Iran: A Case Study of Payamnoor and Farhangian Universities", Engineering, Technology & Applied Science Research, Vol. 7, No. 4, pp. 1874-1878, 2017.

Parisa Shayan, Assoc. Prof. Dr. Mustafa İlkan, Dr. Fatma Tansu Hocanın, MOOC Effectiveness and Efficiency, International Conference on New Trends in Educational Technology (INTET2016), Famagusta, North Cyprus, 0304 May 2016.

Her current research interests are assessment of users' satisfaction using novel technologies and MOOC providers along with commercial LMSs.

Menno van Zaanen got the post-graduate certificate in Higher Education, Macquarie University, Sydney, Australia, 2007; the PhD in computational linguistics, Leeds University, Leeds, UK, Bootstrapping Structure into Language: Alignment-Based Learning", 2002; the MA in computational linguistics, University of Amsterdam, Amsterdam, the Netherlands, "Publishing and Translation - Problems and Solutions", 1998; the MSc in computer science, Vrije Universiteit, Amsterdam, the Netherlands," Error Correction using DOP", 1997.

He is assistant professor at School of Humanities, Tilburg University, Tilburg, the Netherlands, September 2009-now. His previous occupations were respectively: Researcher (researching implicit structure in sequences) and lecturer; Implicit linguistics project, ILK/Computational Linguistics, Tilburg University, Tilburg, the Netherlands, January 2008-August 2009.

His research focuses on automatic learning in sequences. This includes the development and application of machine learning systems on sequential data such as natural language and music. His current research interests are computational linguistics, pattern recognition in multi-modal data, and assessment of users' satisfaction using novel technologies.