MINING EDUCATIONAL DATA: A CASE STUDY OF KATHMANDU UNIVERSITY

[1]MANISH JOSHI, [2]SUSHIL SHRESTHA

[1,2]Digital Learning Research Lab, Department of Computer Science and Engineering, Kathmandu University, Nepal
E-mail: [1]mnzjoshi@hotmail.com, [2]sushil@ku.edu.np

Abstract - Online Learning platforms generate vast amount of data from each user interactions in the course and most of these platforms also offer sophisticated monitoring and reporting tools to keep track of these data. However, it is often incomprehensible, poorly organized and difficult to follow, because of the tabular format that most platforms use. This research aims to provide an initial step in the direction of creating a user-friendly monitoring tool that can help educators get information about the interactions in the course easily. So, data of approximately 2000 students, who were registered to Online Learning platform (Moodle) were analyzed. These logs were then used to determine the relationship between student's grades and student's participation/interactions in their respective courses. Finally, a decision tree was generated using J48 classifier which can be integrated directly in the Moodle platform.

Keywords - Educational Data Mining, Learning Analytics, Moodle

I. INTRODUCTION

Data analysis has two distinct fields in E-learning, Educational Data Mining (EDM) and Learning Analytics (LA). EDM is an emerging discipline that focuses on applying data mining tools and techniques to educationally related data [1]. While LA emerges from two converging trends: increasing use of Virtual Learning Environment (VLE) in educational institutions and application of data mining techniques to business intelligence processes. Although LA and EDM share many common goals and interests, they have distinct technological, ideological, and methodological orientations [2].

The scope of EDM includes areas that directly impact students and educators. For example, mining the course content for relevant data and developing a recommendation system that can give valuable insights to educators about the courses they are offering and how students perceive the course. Other areas within EDM include analysis of educational processes including admissions, course selections etc. Moreover, the application specific data mining techniques such as web mining, association rule mining, classification and multivariate statistics are also key techniques applied to educational-related data [3]. These data mining methods are largely exploratory techniques that can be used for prediction and forecasting of learning and institutional improvement needs. Therefore, research done within EDM primarily focuses on quantitative analyses.

The underlying idea in LA comes from a large amount of data ("big data") generated by VLEs during the learning process in Online Learning (OL) systems. Learners and educators leave a trail of data while interacting with the OL system and with the different resources offered in the course. Those interactions are essentially connected to the various learning activities within the course, allowing learners to actively interact with the content, other learners, and educators. These data are usually easy to extract and different LMSs provides easy user interfaces to extract these data in various formats. However, the volumes of these data are usually far too big to perform analysis by educators and resources needed to transform the data are often scarce [4]. Therefore, it is necessary to develop ad hoc tools which allow filtering of these data so that useful information may be extracted from them [5].

1.1. Problem Domain

Most of the OL platforms offer sophisticated monitoring and reporting tools to keep track of data generated by student interactions in the course. However, it is often incomprehensible, poorly organized and difficult to follow because of the tabular format that most OL platforms use. More than often, data generated from these tools need to be pre-processed and "mined" to extract any useful knowledge from it. This process requires knowledge of data mining tools and techniques, which are normally designed for more power and flexibility than for simplicity. Most of these data mining tools could be too complex for educators to use and the features they offer might go well beyond the scope of what an educator might require. So, there is a need to structure and develop efficient and effective monitoring tools within the OL platform, allowing better use of this knowledge generated by OL platforms.

1.2. Objectives

The primary objective of this research is to predict student performance from the data generated by student interactions in the Moodle platform. In order to fulfill the primary objective, secondary objectives of the research are to analyze different classification algorithms and to find out the suitable algorithm for the given data.

II. RELATED WORKS

The research problem of predicting student performance in OL platforms can be analyzed from different angles. Most of the recent researches analyze these data by using data mining techniques to uncover any relations and useful knowledge about the learning process [6], [7], [8], [9]. Some of the common approaches [6], [10], [11] utilize data mining techniques to provide useful knowledge from different data generated in the OL systems, while others tried to address this research problem by integrating different tools within the OL system [12] [13] [14]. In the current literature, there is a number of approaches given to provide a baseline for analyzing students' performance [15], [16], [17], [18], [6]. These approaches provide the insights to utilizing the data generated from OL systems. However, these approaches do not provide any practical implementation which can be used directly in the Moodle platform.

III. MATERIALS AND METHODS

In order to achieve the primary objective of this research, design and creation process was used with an objective to obtain a concise conceptual framework for research via clear guidelines for understanding, executing and evaluating data. The data from the OL platform at different schools of Kathmandu University was taken and processed. The process of data analysis contains implementations of data pre-processing, classification, and visualization by different algorithms.

3.1. Data Set

For this purpose, log data was collected from various OL platform used in Kathmandu University. The collected data was from the period of July 2016 to Aug 2017 and it contained more than 220,000 recorded student interactions with the OL platform.
This dataset was cleaned and pre-processed so that it will be free of any anomalies.

3.2. Data Pre-processing

Moodle stores student activity details in its database to keep track of login times, materials accessed in the course, participation in forums and submission of assignments by participants. Fig.1. shows the sample Moodle log data that was extracted from OL system of Kathmandu University. These log files usually have comprehensive information that needs to be cleaned and filtered for better understanding by a general audience (for e.g. teachers). If presented properly, these logs could help educators understand the student's participation/interactions in their respective courses. Therefore, Data pre-processing is performed on the given data to transform them into the required format used by data mining algorithms or framework.

6/06/17, 23:37	Front page	System	Course viewed	The user with id '4276' viewed the course with id '1'.
6/06/17, 23:37	Course: ENVE101 INTRODUCTION TO	System	Course viewed	The user with id '3963' viewed the course with id '76'.
6/06/17, 23:35	Course: GEOM 204: Geographical Inf	System	Course viewed	The user with id '3378' viewed the course with id '438'.
6/06/17, 23:34	Course: GEOM 204: Geographical Inf	System	Course viewed	The user with id '3378' viewed the course with id '438'.
6/06/17, 23:34	Course: GEOM 206: Cartography	System	Course viewed	The user with id '3378' viewed the course with id '437'.
6/06/17, 23:33	URL: Lecture I, II and III	URL	Course module viewed	The user with id '3378' viewed the 'url' activity with course
6/06/17, 23:33	Course: GEOM 206: Cartography	System	Course viewed	The user with id '3378' viewed the course with id '437'.
6/06/17, 23:32	Course: ENVE101 INTRODUCTION TO	System	Course viewed	The user with id '3970' viewed the course with id '76'.

Fig.1. Moodle Log data

Moodle already employs user authentication system in which logs have entries identified by users. Therefore, data pre-processing tasks such as user identification, session identification etc is generally not needed. A java application was created for reading the data and generating it in the required format. Overall, following tasks were carried out for preparing the data for further processing.

Create Summarized table: Students information and activity log of students are spread all over the database so we need to create a separate table which stores the summarized data of student's performance and activity. The datasets were cleaned of any unnecessary information (such as IP addresses, personal user information etc.) and only following attributes were used for each student:

- Student_id: Unique Id of student given by the system.
- Cs_view: Number of times student has viewed the course
- Rs_view: Number of times student accessed/viewed resource.
- Fr_view: Number of time student has read or participated in forum discussion
- Quiz_avg: Average performance if quiz.
- Performance: student's performance in their final exams

Data discretization : Discretization is the process of dividing numeric data into categorical classes in order to increase data interpretation and comprehensibility. Discretizing the data makes it easier for educators to understand the data and is more user-friendly than precise magnitudes and ranges. The numerical values such as average quiz score; final marks etc. have been discretized to have nominal values that can be easily interpreted by the classifier. The manual method of discretization was used in mark attribute to specify the two intervals and labels (FAIL if the value is <5, PASS if the value is >5). In all the other attributes

three intervals and labels were used (LOW, MEDIUM, and HIGH).

Data transformation: The data must be transformed to the format required by data mining algorithm or framework. For the purpose of this research, the processed data is exported into ARFF format for further analysis. An ARFF (Attribute-Relation File Format) file is an ASCII text file that describes a list of instances sharing a set of attributes. This file is then imported into Weka explorer to perform data mining and to discover the appropriate decision tree.

student_i	cs_view_c	assessme	quiz_attai	avg_asses	quiz_avg_	performance
45642	424	4	0	HIGH	LOW	Pass
240712	94	4	0	HIGH	LOW	Pass
360489	86	3	3	HIGH	HIGH	Withdrawn
620935	4	0	0	LOW	LOW	Withdrawn
620974	305	10	6	HIGH	HIGH	Withdrawn
620977	244	4	4	HIGH	HIGH	Withdrawn

Fig.2. Final data set required by data mining algorithm

Fig.2. shows the part of final data set after it's been cleaned and pre-processed. This data is the final data set used for the data analysis part.

3.3 Data Analysis

In order to classify the students that are having difficulty in the course, it is necessary to find out if it is possible to predict the class (output) variable using the explanatory (input) variables which are retained in the model. The choice of data mining algorithms selected in this research mostly depended upon whether the classification technique can generate a set of rules or not (for e.g. Decision Tree). This is done to ensure that the final output is easy to understand and can be easily converted to a set of production rules. The most-preferred ones in the literature, J48, JRIP, PART and Decision Table were selected and analyzed for this research using WEKA explorer. The inputs of these classifiers (J48, JRIP, PART and Decision Table) were seven different variable of the students, which were, course views, resource views, assignment submitted, quiz attended, average assignment performance, average quiz performance, and overall performance. Each classifier is applied for two testing options − cross-validation (using 10 folds and applying the algorithm 10 times – each time 9 of the folds are used for training and 1 fold is used for testing) and percentage split (2/3 of the dataset used for training and 1/3 – for testing). Cross-validation and split-percentage methods were used for testing the accuracy of each classifier.

IV. RESULTS

Data collected from different OL platform was pre-processed and analysed using four different classification algorithms; J48, JRIP, PART and Decision Table. Table 1 shows the overall performance of the classifiers with the given data set. It was found that amongst these four classifiers, J48 had the most accurate model which classified 87.9824% of instances correctly. JRip identified 87.0483% instances correctly, making it close alternative to J48 classifier. Decision Table also came close to previous two classifiers with 87.3327% correctly identified instances, while, PART identified 86.95% of instances correctly. Therefore, the model created by the J48 algorithm can be used for further deployment.

Table 1. Accuracy rates of classifiers for total data set

Classifier	J48	JRip	Decision Table	PART
Accuracy	0.879	0.870	0.873	0.869
True Positives	0.880	0.870	0.873	0.869
False Positives	0.110	0.116	0.119	0.120
Precision	0.887	0.881	0.879	0.877
Recall	0.880	0.870	0.873	0.869

In the WEKA data mining tool, J48 is an open source Java implementation of the C4.5 algorithm. The WEKA tool provides a number of options associated with tree pruning. In case of potential over fitting, pruning can be used as a tool for précising.

```
avg_assessment_performance = HIGH
|   quiz_avg_performance = HIGH: Pass (2529.0/503.0)
|   quiz_avg_performance = LOW
|   |   cs_view_count <= 42: Fail (849.0/352.0)
|   |   cs_view_count > 42: Pass (1819.0/394.0)
|   quiz_avg_performance = MEDIUM: Pass (0.0)
avg_assessment_performance = LOW
|   quiz_avg_performance = HIGH: Fail (0.0)
|   quiz_avg_performance = LOW: Fail (1292.0/1.0)
|   quiz_avg_performance = MEDIUM
|   |   cs_view_count <= 35: Fail (176.0/37.0)
|   |   cs_view_count > 35: Pass (81.0/38.0)

Number of Leaves  :  8

Size of the tree :  13
```

Fig.3. Decision tree generated by J48

Fig.3. shows the final decision tree generated by the J48 classifier. This decision tree can be converted into a set of IF-THEN-ELSE rules for integrating directly into Moodle platform.

V. DISCUSSION AND CONCLUSION

Currently, OL platforms generate vast amount of data from each user interactions in the course. These data, if utilized properly, can give valuable insights on the behavior and performance of the users. Most of the

OL platforms do offer some level of monitoring and reporting tools. However, more than often these OL systems do not provide a user-friendly interface to monitor student performance efficiently. Without any prior knowledge of data mining or analytical tools, educators often face difficult time coping up with this interface. The decision tree generated (see fig.3.) provides the initial step required to develop the monitoring tool which can be directly integrated with Moodle platform. The generated decision tree is based on set of IF-THEN-ELSE rules which can be converted into set of rules in PHP to be integrated into the Moodle platform. This tool can then provide the real-time information to the educators, without them having to use complex data mining tools and techniques. The generated decision tree is highly dependent upon the resource viewed and the average number of quizzes passed by the students. The knowledge discovered by this decision tree can prove valuable information for educators to make decisions about Moodle course activities and for classifying new students in the course.

FUTURE WORKS

The research only provides an initial step in predicting the student performance through moodle data. In the future, the decision tree generated can be integrated within the Moodle platform by developing a plug-in.

ACKNOWLEDGMENTS

This research work is partially supported by University Grants Commission (UGC), Sanothimi, Nepal and was conducted in Digital Learning Research Lab at Kathmandu University, Dulikhel Nepal.

REFERENCES

[1] R. S. Baker and K. Yacef, "The state of educational data mining in 2009: A review and future," Journal of Educational Data Mining, pp. 3-17, 2009.

[2] G. Siemens and R. S. J. d. Baker, "Learning Analytics and Educational Data Mining: Towards Communication and Collaboration," in In Proceedings of the 2Nd International Conference on Learning Analytics and Knowledge, New York, 2012.

[3] T. Calders and M. Pechenizkiy, "Introduction to the special section on educational data mining," Acm Sigkdd Explorations Newsletter, pp. 3-6, 2012.

[4] Van Barneveld, K. E. Arnold and J. P. Campbell, "Analytics in higher education: Establishing a common language," EDUCAUSE learning initiative, pp. l-ll, 2012.

[5] L. Johnson, S. Adams Becker, V. Estrada and A. Freeman, The NMC Horizon Report: 2015 Higher Education Edition., Austin: New Media Consortium, 2015.

[6] Daud, N. R. Aljohani, R. A. Abbasi, M. D. Lytras, F. Abbas and J. S. Alowibdi, "Predicting Student Performance using Advanced Learning Analytics.," in In Proceedings of the 26th International Conference on World Wide Web Companion, 2017.

[7] W. He, "Examining students' online interaction in a live video streaming environment using data mining and text mining," Computers in Human Behavior, pp. 90-102, 2013.

[8] Koprinska, J. Stretton and K. Yacef, "Students at Risk: Detection and Remediation," in Proceedings of the 8th International Conference on Educational Data Mining, Sydney, 2015.

[9] G. W. Dekker, M. Pechenizkiy and J. M. Vleeshouwers, "Predicting students drop out: A case study," Educational Data Mining, 2009.

[10] Mueen, B. Zafar and U. Manzoor, "Modeling and Predicting Student's Academic Performance Using Data Mining Techniques," I.J. Modern Eduation and Computer Science, pp. 36-42, 2016.

[11] Kuzilek, M. Hlosta, D. Herrmannova, Z. Zdrahal and A. Wolff, "OU Analyse: Analysing At-Risk Students at The Open University," Learning Analytics Review, 2015.

[12] S. Shrestha and J. Kim, "Bridging the Gap between E-Learning and Knowledge Management (KM): An enhancement of Moodle system by applying KM Functions," International Journal of Science and Technology, 2013.

[13] Z. Krajcso, "Knowledge Management and e-Learning in Higher Education," 2011.

[14] D. Tessier and K. Dalkir, "Implementing Moodle for e-learning for a successful knowledge management strategy," Knowledge Management & E-Learning, p. 414–429, 2016.

[15] R. Baker, S. Gowda and A. Corbett, "Automatically Detecting a Student's Preparation for Future Learning: Help Use is Key," in In Proceedings of the 4th International Conference on Educational Data Mining, 2011.

[16] R. Baker, A. Corbett, S. Gowda, A. Wagner, B. MacLaren, L. Kauffman, A. Mitchell and S. Giguere, "Contextual Slip and Prediction of Student Performance After Use of an Intelligent Tutor," in Proceedings of the 18th Annual Conference on User Modeling Adaptation, and Personalization, 2010.

[17] Bowers, "Analyzing the Longitudinal K-12 Grading Histories of Entire Cohorts of Students: Grades, Data Driven Decision Making, Dropping Out and Hierarchical Cluster Analysis," Practical Assessment, Research & Evaluation (PARE), pp. 1-18, 2010.

[18] S. Huang and N. Fang, "Predicting student academic performance in an engineering dynamics course: A comparison of four types of predictive mathematical models," Computers & Education, pp. 133-145, 2013.

★★★