Contents

Prediction of student's performance using data mining techniques

Rahila Umer Sumalani

A proposal submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Computer Science

Computer Science Department
Massey University, Auckland
New Zealand
June 2017

Prediction of student's performance using data mining techniques

Rahila Umer Sumalani

Summary

0.1 Chapter 1

0.2 Introduction

0.2.1 Motivation

The emergence of new technologies in the field of education is one of the major innovations in past decade which holds promise of enhancing the learning achievement[Cha(2012)].In 2011, the Horizon report refereed Learning Analytics as the future trend in teaching and learning[Nor(2010)]. The use of INTERNET in education, development of educational software, e-learning resources and student information systems is providing large repositories of data. Exponential growth of such educational data provides an opportunity for stakeholders to understand the learning process which ultimately help to take managerial decisions.
Educational Data mining (EDM) and Learning Analytics(LA) are two different research communities that are exploring capabilities of Big data in field of education.EDM refers to "developing, researching and applying computerized methods to detect patterns in large collections of educational data that would otherwise be hard or impossible to analyze due to the enormous volume of data within which they exist"[Rom(2013)]. The Society of Learning Analytic Research(SoLAR) defines LA is "Learning analytics is the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environments in which it occurs".
LA and EDM share similar goals i.e. to understand and to optimize learning process but are different in their approach to achieve the said goal[Sie(2010)]. Regardless of the difference in approach, the combination of LA and EDM on education data can lead to understand the learner's behavior, interactions and learning path in a data-driven way and offer opportunities to optimize learning process through educational system.
In this study we are intended to explore the Machine learning, Data mining and Process mining methods in the field of education and investigate their feasibility. The abundance of educational data increased opportunities in field of LA and EDM. These data sets provide a benchmarks through which new algorithms can be developed and can be compared to existing algorithms[Ver(2012)]. In this study we will be exploring different datasets that are available on-line and also use Massey University student's data.

0.3 Objectives of the study

0.4 Assumptions and Research Questions

0.4.1 Expected Contributions

The main contributions of the research are as follows:

- Prediction of student's performance
- Machine learning methods
- Investigate Process mining applications

0.4.2 Report Outline

This report comprises of 4 chapters. The chapter 2 and 3 provide an introduction to the topics and related work. Next 2 chapters include empirical studies and the results. Last chapter provides the conclusion of this work.The report chapters are briefly described below.

- Chapter 2 describes the background of the topic. It starts with introduction of Learning Analytics(LA), Educational Data Mining(EDM) and Process Mining(PM). Methods of Machine learning, data mining and process mining are explained further.
- Chapter 3 describes the current state of the art in the fields of LA, EDM and PM. It starts with the introduction of the fields and their general application and then narrowed down to the application of classification, prediction and process mining.
- Chapter 4 presents our first empirical study. This chapter starts with introduction of ML methods used in prediction of student's performance and ends in the results of analysis.
- Chapter 5 prsents our second study where we investiage the application of process mining methods on educational data set. It starts with introduction of Process mining methods and in the end the results of the analysis are presented.
- Chapter 6 provides a discussion over the results of the empiral studies, summarizes the results and conclusion of the report, and present the future plan of remaining two years of research.

0.5 Chapter 2

0.6 Introduction

This chapter is dedicated to illustrate the fundamental concepts of this thesis. It provides a background on concepts and methods used in our study.

0.7 Learning Analytics and Educational Data mining

Learning Analytics (LA) is an emerging field in the Technology Enhanced Learning(TEL). It exploits the use of big data in education field to find patterns and extract knowledge from unstructured data to improve the overall learning process[Siemens and Long(2011)]. LA concepts and methods are evolved from multiple disciplines like Machine Learning(ML),artificial intelligence(AI),information retrieval, visualization, statistics. LA is also related to research areas of Technology Enhanced Learning(TEL) e.g. academic analytics,action research,educational data mining, personalized adaptive learning. [Cha(2012)]. However,the main focus of learning analytics is the conversion of education data into intelligent actions to improve learning process which may also include the process of analyzing the relationship between learner, content, institution and instructor.
Educational Data Mining(EDM) is concerned with" developing, researching and applying computerized methods to detect patters in large collections of educational data[Romero C(2010)].EDM is considered as an application of data mining(DM) techniques that apply on educational data sets to resolve educational research issues[Rom(2007)] "

0.7.1 Learning Analytics Process

A typical LA process starts with the data gathering from learner's activities when they interact with learning environment like Virtual Learning Environment or Learning Management System. Submitting assignments, taking on-line exams, reading materials or writing a forum posts are few examples of learner's on-line activities, however real challenge is to integrate the off-line activities to get optimal results. In data collection process learner's privacy is to be addressed and is considered as important as functional requirements[Pap(2014)]. The second step is to apply different mining techniques such as clustering, classification, association rules and social net-

work analysis on preprocessed data. Last step is reporting the results to the stakeholders about the overall performance and makes a ground for further decision making. Results of the analysis are mostly graphically visualized and can be integrated as widget into a VLE or a dashboard. It is easy to interpret the visualized results, however graphical visualization does not guarantee that the results will be interpreted correctly which makes it very important step for taking appropriate decision.

Regardless of the approach, primary research objective is detection, identification and modeling student's learning behavior. Data sources VLE as main source of LA produces 'big data' which are too large to be handled. Process Mining Process mining provides a set of algorithms, techniques to analyze event data. Process mining includes process discover, conformance checking and enhancement. Discovery techniques helps to get process models through log data. Conformance checking compares process models with the event log and checks the actual business process model and event log follow same process or can find the deviations. While enhancement allows to modify or improve the models through results of process discovery and conformance checking [van der Aalst(2011)]

0.7.2 Analytics Methods and Data Features

0.7.3 Classification

0.7.4 Clustering

0.7.5 Process Mining

0.7.6 Summary

0.8 State of the Art

0.8.1 Introduction

Following four stages define review protocol to conduct the literature review. 1. Data collection 2. reviewing and aces

The emergence of Massive On-line Open Courses offered by selective higher education institutes has attracted many researchers, policy maker and stake holders. A number of studies have been conducted in this topic.

0.8.2 Classification in Education

One of the key research areas is prediction of dropout and retention of students. Research conducted in Thai university [Boongoen(2017)] link-based cluster ensemble method is used as a data transformation framework for prediction. Results are compared with several state-of-the art dimensionality reduction techniques.
In this paper[Ye2(2014)] authors have used and extended *traditional* features for MOOC analysis with higher granularity to make more accurate predictions for dropout and performance.Analysis was made using video lectures, weekly quizzes, and peer assessments from the ten-week course. Traditional features were extended using some detailed temporal features like when assessment was started during the week or when first lecture was viewed. Results compared with existing studies show that these features improved the prediction accuracy. The time when student's start the peer assessment assignment was found to be the good predictor. Once the peer assessment score was available prediction performance improved. Analysis show that the student's who watched video and did not take quiz was the ones who mostly drop out. Overall results show that more precise temporal features and more quantitative information improved early prediction accuracies and false alarm rates as compared to using only assessment score features.

0.8.3 Process mining in Education

Data mining techniques have been applied to education data to find patterns or to build descriptive and predictive models.These analysis help us to understand the underlying process. However it does not provide visual representation of the process for analysis. Process mining is a new area of data mining that provides such kind of analysis. Process mining tools extract process-related knowledge from event logs and provide information about the process. In [Pec(2009)] process mining techniques were applied to on-line as-

sessment data where students were given different options to go through the questions. Results show that how students followed the process and what path majority of the students followed.

0.8.4 Clustering in Education

0.8.5 Machine learning Methods

0.8.6 Summary

Student's Data Analysis for Prediction future in MOOC environment

0.9 Summary

The motivation for the research is the popularity of MOOC courses and early drop out and low completion rates in MOOC courses. Early predictions can provide individual guidance and support students that are struggling through the course, which ultimately can help to decrease the drop out rate. Generally, student's who drop out the embedded quizzes, can be guided and explained the importance of such modules in their learning.Dropout rate in MOOC environment is more as compare to the traditional settings. The motivation for the research is the popularity of MOOC courses and early drop out and low completion rates in MOOC courses. Early predictions can provide individual guidance and support students that are struggling through the course, which ultimately can help to decrease the drop out rate. Generally, student's who drop out the embedded quizzes, can be guided and explained the importance of such modules in their learning.Dropout rate in MOOC environment is more as compare to the traditional settings. The motivation for the research is the popularity of MOOC courses and early drop out and low completion rates in MOOC courses. Early predictions can provide individual guidance and support students that are struggling through the course, which ultimately can help to decrease the drop out rate. Generally, student's who drop out the embedded quizzes, can be guided and explained the importance of such modules in their learning.Dropout rate in MOOC environment is more as compare to the traditional settings.

0.10 Chapter 3

0.10.1 Introduction

Massive open on-line courses have become very popular among student's community providing them an opportunity to register for courses offered by prestigious universities around the world. MOOC provides a learning environment which attracts large number of learners with different goals and motivation. Courseera, edX and Udacity are the three pioneers of MOOC platform which then followed by several around the world like Miriada and Spanish MOOC in Spain, Khan Academy in north America, University in German, FutureLearn in England, Open2Study in Australia, Fun in France, Veduca in Brazil, Schoo in Japan, and xuetangX in China.
MOOC's brought revolution in education by centralizing the global resources and restructuring the learning environment. Unlike traditional higher education learning environment, MOOC provides an open access to the course to anyone who got access to the INTERNET. It provides free and open access to high quality advanced courses that comprises of video lectures, reading materials, quizzes, problem sets and forums for productive discussions to foster learning process and communities. MOOC provides large number of courses from selected higher education institutes on an unprecedented scale and at a negligible marginal cost.
In MOOC student's all activities are recorded and increased availability of such data from such environments provide opportunities to investigate student's learning behavior closely and improve the learning process. Although, there are massive enrollments in courses offered in MOOC, but the completion rate and the persistence students, is quite low, often less than 20% [Kiz(2013)].
One of the challenges in MOOC is the low retention rate of the students which is heavily criticized. Therefore predicting the likelihood of dropout is necessary to maintain and encouraging student's in their learning activities. In this report, the focus in on predicting student's performance through the traces they leave. The aim is to apply data mining/machine learning algorithms to student data, as the students are progressing through a course, in order to predict which students are at risk of not satisfying course requirements, or are likely to withdraw. The identification of such students would then enable educators to carry out various forms of early intervention, or provide additional and more tailored support as mitigation measures.

0.10.2 Study Objectives

The study will strive to determine:

- which variables/factors are the most meaningful for achieving high predictive accuracy,
- which algorithm or ensembles of algorithms are most suitable for generating classification models of high accuracy.
- Can we derive accurate and reliable early predictors of student dropout and performance in MOOC environments?

0.11 Experimental Design

0.11.1 Event Data

0.11.2 Feature Selection

0.12 Results

0.13 Discussion

0.14 Summary

0.15 Application of Data Mining in Education

0.15.1 Introduction

0.15.2 Prediction of Performance

0.15.3 Data Set

0.15.4 Results

0.15.5 Discussion

0.15.6 Summary

0.16 Appendix A MOOC Data Set

0.16.1 Data Set Description

0.16.2 Features

Bibliography

[Rom(2007)] Educational data mining : A survey from 1995 to 2005. 33: 135–146, 2007. doi: 10.1016/j.eswa.2006.04.005.

[Pec(2009)] Process Mining Online Assessment Data. *Proceedings of the 2nd International Conference on Educational Data Mining*, pages 279–288, 2009.

[Nor(2010)] *Investigation of lewis acid versus lewis base catalysis in asymmetric cyanohydrin synthesis*, volume 16. 2010. ISBN 9780982829059. doi: 10.1002/chem.201001078.

[Sie(2010)] Learning Analytics and Educational Data Mining : Towards Communication and Collaboration. 2010.

[Cha(2012)] A Reference Model for Learning Analytics. *International Journal of Technology Enhanced Learning*, 4(5/6):318–331, 2012. ISSN 1753-5255. doi: DOI: 10.1504/IJTEL.2012.051815.

[Ver(2012)] Dataset-Driven Research to Support Learning and Knowledge Analytics. *Educational Technology & Society*, 15:133–148, 2012. ISSN 1436-4522.

[Kiz(2013)] Deconstructing Disengagement : Analyzing Learner Subpopulations in Massive Open Online Courses. *Lak '13*, page 10, 2013. ISSN 9781450317856. doi: 10.1145/2460296.2460330.

[Rom(2013)] Data mining in education. 3(February):12–27, 2013. doi: 10.1002/widm.1075.

[Pap(2014)] Learning Analytics and Educational Data Mining in Practice : A Systematic Literature Review of Empirical Evidence The research questions. 17:49–64, 2014.

[Ye2(2014)] Early Prediction of Student Dropout and Performance in MOOCs using Higher Granularity Temporal Information. *Journal of Learning Analytics*, 1(3):169–172, 2014.

[Boongoen(2017)] Natthakan Iam-on Tossapon Boongoen. Improved student dropout prediction in Thai University using ensemble of mixed-type data clusterings. *International Journal of Machine Learning and Cybernetics*, pages 497–510, 2017. ISSN 1868-8071. doi: 10.1007/s13042-015-0341-x. URL `http://dx.doi.org/10.1007/s13042-015-0341-x`.

[Romero C(2010)] Pechenizky M Baker R. Romero C, Ventura S. *Handbook of Educational Data Mining*. CRC Press, 2010.

[Siemens and Long(2011)] George Siemens and Phil Long. Penetrating the Fog: Analytics in Learning and Education. *EDUCAUSE Review*, 46: 30–32, 2011. ISSN 1527-6619. doi: 10.1145/2330601.2330605. URL `http://search.proquest.com.proxy.library.vanderbilt.edu/docview/964183308`

[van der Aalst(2011)] Wil M.P. van der Aalst. *Process Mining*. Springer-Verlag Berlin Heidelberg, 1 edition, 2011. doi: 10.1007/978-3-642-19345-3.

0.17 Glossary