MOOC Dropout Prediction Using Machine Learning Techniques: Review and Research Challenges

Fisnik Dalipi
Linnaeus University, Sweden & University College of Southeast Norway

Ali Shariq Imran, Zenun Kastrati
Norwegian University of Science and Technology
This is a presentation of the paper: “MOOC Dropout Prediction Using Machine Learning Techniques: Review and Research Challenges”, which was accepted and presented at the EDUCON’2018 – IEEE Global Engineering Education Conference.

The full paper can be downloaded from the IEEEXplore website: https://ieeexplore.ieee.org/
Introduction

- Massive Open Online Courses (MOOCs) are now the most recent topic within the field of e-learning.
- Some researchers even predict MOOC will represent the fourth stage of online education evolution, after the third stage of LMS.
- MOOCs are progressively becoming an integral part of a learning process in higher education settings.
- Also proven that the MOOC approach when complemented with a flipped classroom pedagogy would also bring advantages toward enhancing the learning experience of students substantially.
- But ...
The biggest threat to MOOCs
Reasons behind the dropout

Student related
- Lack of motivation
- Lack of time
- Insufficient background

MOOC related
- Course design
- Lack of interactivity
- Hidden costs
State-of-the-art: Frequency of occurrences of ML for MOOC dropout prediction
Challenges of dropout prediction using ML

- Lack of enough sample data
- Managing big masses of unstructured data
- Data variance
- High data imbalance
- Availability of publicly accessible dataset
- Lack of standard for creating and representing clickstream data
- Student schedule related challenges
Challenges of dropout prediction using ML

- Lack of enough sample data
- Managing big masses of unstructured data
- Data variance
- High data imbalance
- Availability of publicly accessible dataset
- Lack of standard for creating and representing clickstream data
- Student schedule related challenges

- For unbiased classification results
- Lack of negative samples, i.e. HarvardX
  - E.g. Out of 641138 registered students only 17687 obtained certification
- However, the effectiveness of the ML relies on the availability of enormous amount of both positive and negative samples
- A significant difference in observable sample data affects
  - the generalization of the deep neural network model during the training step, but also
  - comprises the classification accuracy.
Challenges of dropout prediction using ML

- MOOC comprises of big masses of unstructured data, missing data occurs and is hard to validate.
- A multitude of ML techniques can’t be applicable.
  - E.g. Hidden Markov Models, since:
    - it relies on a finite data set, and
    - cannot be applied if observations are missing.
- Researchers’ reaction: replacing missing observations with mean values.
  - Not a realistic choice for many of the observed features.
Challenges of dropout prediction using ML

- In MOOC's self-paced learning pedagogy, students have the freedom to decide what, when, and how to study.
- This might lead to considerable data variance, which may produce less accurate and reliable ML models.
- Particularly for SVM and Naïve Bayes, since their performance can quickly turn poor when dealing with imbalanced classes.
- Overall, the high data imbalance may result in biasing of the classifier towards the majority of the class.

Other challenges include:

- Lack of enough sample data
- Managing big masses of unstructured data
- Data variance
- High data imbalance
- Availability of publicly accessible dataset
- Lack of standard for creating and representing clickstream data
- Student schedule related challenges
Challenges of dropout prediction using ML

- MOOC platforms are often reluctant to publish the data due to confidentiality and privacy concerns.
- The de-identified information which is made public is also restricted to system generated events (clickstream data) only, while most of the user-provided data is omitted.
- The lack of user-provided data can be a hindrance in successfully estimating the reasons behind the dropouts.

- Lack of enough sample data
- Managing big masses of unstructured data
- Data variance
- High data imbalance
- Availability of publicly accessible dataset
- Lack of standard for creating and representing clickstream data
- Student schedule related challenges
The user data MOOC platforms gather and the clickstream data they generate doesn’t necessarily conform to a standard format, due to lack of available standards - such as IEEE LOM for learning objects.

The lack of a standard for creating and representing clickstream data for MOOC implies that a network trained and validated for one MOOC on a given platform might not be interoperable across different MOOC on another platform, or even on the same platform if a MOOC gathers and collects data differently.

Lack of standards gives rise to another challenge, i.e., how to create compelling ML models and predictive solutions that can be generalized across different MOOCs?
Students want to construct timetable based on their preferences

Satisfying such preferences of students to be accommodated on different timetables is a complicated problem which can be solved using Artificial intelligence (AI) optimization techniques.

These AI techniques are specialized to be used for scheduling and thus students’ time constraints can be solved using these techniques.
Towards useful and effective predictive solutions

- Encouraging social interactions and simulating peer-evaluation activities could also potentially represent an efficient way to mitigate the dropout.
- To transfer and evaluate models from one MOOC to another, and to perform real-time prediction with ongoing courses.
- To allow instructors track and monitor students activity across different modules of the course.
- Predicting models can benefit from the data provided by the students that don’t reveal the identity of the user.
- To analyze posts, feedback, and comments shared by the candidate on the discussion blogs and social forums of the MOOC.
Conclusions

- This paper provides a comprehensive review of the most recent and relevant research endeavors on machine learning application toward predicting, explaining and solving the problem of student dropout in MOOCs.

- It also identifies some of the critical challenges associated with student dropout prediction and provides recommendations and proposal to assist researchers employing various machine learning techniques in solving it timely and efficiently.

- Additionally, this paper floats the idea of unification of a clickstream data and student-provided data as a standard, similar to various learning objects metadata standards.

- Future work will follow on data collection for the development and analysis of deep learning algorithms towards solving the MOOC student dropout.
Thank You!

QUESTIONS?