International Journal of Engineering & Technology, 7 (4.32) (2018) 124-126

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Predicting MOOC Dropout Based on Learner's Activity Features

Soufiane Ardchir[1], Mohamed Amine Talhaoui[1], Houda Jihal[1], Mohamed Azzouazi[1]

[1] LTIM Laboratory, Faculty of science Ben Msik, Hassan II University Casablanca, Morocco
**Corresponding author E-mail:soufiane79@gmail.com*

Abstract

Over the last few years, open massive online courses (MOOC) have become very popular and greatly enhanced as they present a way of learning mostly free online used around the world by millions of participants. Despite all the characteristics and benefits of MOOC, however, one of the crucial problems associated with MOOC is their high dropout rate (completion rate below 13%), which questions the effectiveness of learning technology. The analysis of MOOC data provides a useful means of identifying characteristics that can help to understand the behavior of the learners and to accompany them in order to succeed in their learning. In this paper, we present a dropout predictor that uses student activity features based on machine learning methods for identification of students who are at risk of not completing courses.

Keywords: *MOOCs, Dropout prediction, Big data, Machine learning.*

1. Introduction

MOOCs (Massive Open Online Courses) can be defined as online courses, open to all, that involve several thousand learners simultaneously [1]. They combine digital resources (videos, texts, images ...) and educational activities (quizzes, forums, social networks, tutoring, evaluations ...). "

Specifically, a MOOC is an interactive training open to all that lasts several weeks. Throughout the MOOC, activities and quizzes allow the learners not only to practice but also to assess themselves.

They represent a new form of distance education whose popularity can be described as phenomenal [1]. This new form of learning has managed to draw the attention and interest of many economic and political actors to train a large number of people.

The first observations of MOOCs indicate that about 13% of learners go to the end of the course [2]. The dropout rate represent an obstacle to the success of the MOOC because a large number of learners cannot complete their study through the MOOC. Dropout is often linked to the difficulty to learn autonomously [3] and also to the lack of tools and methods that make it possible to personalize pedagogical activities in order to maintain and motivate learners in this type of environment [4].

This study proposes to predict student MOOCs dropout by using machine learning methods.

We focus on choosing better parameters to turn SVM machine learning. However, a careful selection of learning attributes and machine learning techniques is essential. In fact, inappropriate choice of machine learning techniques can lead to performance degradation.

Indeed, our objective is to develop a model, which is able to detect MOOCs student failure by using classification techniques. Besides, we propose the use of different SVM machine learning because MOOC is a complex problem, data both have high dimensionality and often highly unbalanced.

The final objective is to detect as early as possible the students who show these factors in order to anticipate and provide some type of assistance for trying to avoid and/or reduce MOOC failure.

The paper is organized as follows: Section II presents related works for predicting MOOC dropout. Section III describes data used and the information sources from we collected. Section IV describes the different experiments carried out and the results obtained. In section VI, we summarize the main conclusions and future research.

2. Related Work

Since the deployment of MOOCs, a large body of research reached significant findings related to student MOOC dropout. Voluntary dropout for some who seek by this modality to obtain resources or simply discover new pedagogical practices, or involuntary for others who could not persevere until the end of their project [5].

This success rate of learners in MOOC is often lower than that obtained by students following an identical course in classroom [2]. Many authors have studied this question of MOOC learner perseverance: importance of cooperation, quality of tutoring, as well as individual factors present in students [6].

Numerous research work has addressed the issue of MOOC dropout. We can quote the work of Clow [7], which explains these dropout rates by the demotivation of learners. Other studies have been set up around the engagement of MOOC learners [8].

Different research focuses on the problematic of MOOC dropout using machine learning algorithms to predict when a student will stop visiting the course based on his or her previous behaviors. Kloft et all [9] propose a machine learning framework based on svm for the prediction of dropout in Massive Open online Courses solely from clickstream data. The machine learning algorithm takes the weekly history of student data into account and thus is able to notice changes in student behavior over time.

Halawa et al. [10] have developed a prediction model that analyzes student activity to detect signs of lack of interest or attachment that may cause the student to drop out of MOOC or be away for extended periods. Their model has shown promising results since it is able to predict dropouts in almost half of the cases, even though students are still active. In [11] authors have been able to design effective hidden Markov models from discussion forums and video lectures to help predict learner retention, which can be used to infer general patterns of behavior between learners who complete the course and those who drop out of learning. In [12] authors proposed a prediction model focused on students who already demonstrate some engagement in a course through their participation in discussion forums. By designing a temporal modeling approach that prioritizes at-risk students based on when they are expected to drop out week by week instead of all learners.

Online learning suffers from some limitations; many previous applications using machine learning algorithms suffer from a lack of concern for the type of stability and reliability of the model needed for complex phenomena analysis such as MOOC learning [13]. The machine learning algorithms are also sensitive to the quality of the data, or the need for a technique for constructing from the existing data to improve the prediction result.

3. Methodology

3.1. Data Set

The dataset used in our work is the KDD Cup 2015 dataset extracted from XuetangX MOOC platform [14].The target of the competition is to predict whether a user will drop a course based on his or her prior activities within a time period of 30 days. If a user leaves no records for a specific course in the log during the next 10 days, the case is claimed as a dropout. This dataset contains interactions between students and MOOCs and containing the information of 79,186 users and 39 courses with 120,542 enrollment records in total. Each record has a binary label indicating its dropout status with “1” indicating that the user will drop the course. The dataset is unbalanced with 79.29 % enrollment records labeled as “1” and 20.71 % labeled as “0”. It also provides 8,157,277 user behavior logs within 30 days after the course starts, which contain the detailed user activities such as solving problems, watching videos or engaging in discussion. It does not include any demographic and historical data from past courses.

The dataset contain four csv files, a brief overview of the dataset attributes can be found in Table 1, 2, 3 and 4:

Table 1: object/module data

Features	Description
course_id	The course to which the module belongs
category	The category of the course module
module_id	The ID of a course module.
children	- The children modules of the course module.
start -	The time when the module was released to students

Table 2: enrollment_train

Features	Description
enrollment_id	Enrollment ID
username	Student ID.
course_id	Course ID.

Table 3: true_train

Features	Description
enrollment_id	Enrollment ID
dropout	Ground truth of dropout

Table 4. log_train

Features	Description
enrollment_id	Enrollment ID.
time	Time of the event.
source	Event source (server or browser).
event	In terms of event type, we defined 7 different event types problem Working on course assignments. • video - Watching course videos. • access - Accessing other course objects except videos and assignments. • wiki - Accessing the course wiki. • discussion - Accessing the course forum. • navigate - Navigating to other part of the course. • page_close – Closing the web page.
object	The object the student access or navigate to

3.2. SVM Machine Learning Technique

Support Vector Machine (SVM) is a widely accepted supervised learning model with associated learning algorithms that analyze data and recognize patterns, used for classification and regression analysis [15]. A SVM-based model is proposed to classify all learners into two classes: a learner dropout or not.

The support vector machines technique is pioneered by Vapnik [15] and has been successfully applied to many classification problems throughout the literature. This technique attempts to separate two classes of data using a hyperplane defined by support vectors, which are part of the data set.

Through its training, the support vector machines technique searches for the Optimal Separating Hyperplane (OSH), which is the optimal hyperplane that maximizes the margin between the two classes of the training dataset. To make this clearer, Figure 1 presents an example in 2 dimensions. In this example, several lines can bbve found to successfully separate the data into two classes. The optimal line, the OSH, is the one which lies half-way in the margin. The term ‘‘margin” refers to the sum of the distances *d* of those data that are closest to the line, which are defined as the support vectors.

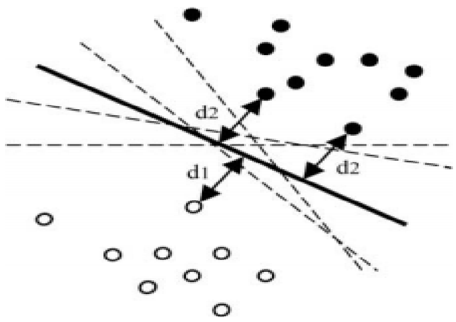

Fig.1. Support vector machines concept

SVMs can generate models with several kinds of decision borders; it depends on the parameters used on the kernel type, three different kernel functions are applied in our research. They include (1) the linear kernel, (2) the polynomial kernel, (3) the radial basis function. They are the most commonly used kernel in the world of research.

4. Experiment Results

4.1. Model Selection of SVM

In order to generate highly performing classifiers capable of dealing with real data, an efficient model selection is required. In this section, we present the experiments conducted to find efficient model for SVM.

We noticed that our dataset could have nonlinear relationships, which prompt us to try SVM. We first used linear SVM and then extended to others kernels, different SVM kernels are selected for prediction including linear kernel, poly kernel, and RBF kernel. Accuracy was calculated from predicted dropout and true dropout to measure performance of dropout prediction.

In our experiment, Grid-search technique has been used to find the best model for SVM with different kernels. This method selects the best solution by evaluating several parameter combinations of possible values. In our experiment, we used sckit-learn python package to turn our experimentation. We have considered the range of the parameter in the grid search. For the linear SVM, we only evaluated the inverse regularization parameter C; for the RBF and polynomial SVM kernel, we tuned both the C and gamma parameter. Note that the gamma parameter is specific to kernel SVMs. After we used the training data to perform the grid search, we used 10-fold cross-validation to evaluate the different SVM kernel, configured significantly with the same random seed to be sure that the same splits training data are done and that each algorithm is evaluated similarly. We obtained the score of the best-performing model as follow: for Linear SVM C= 0.001; but for both Gaussian SVM c = 0.10, gamma = 0.001. In this particular case, the RBF SVM model with 'C'= 0.1' and gamma=0.01 yielded the best k-fold cross-validation accuracy with 88.9 percent, Figure 2 shows estimations of SVM classifier accuracy of training set for all models.

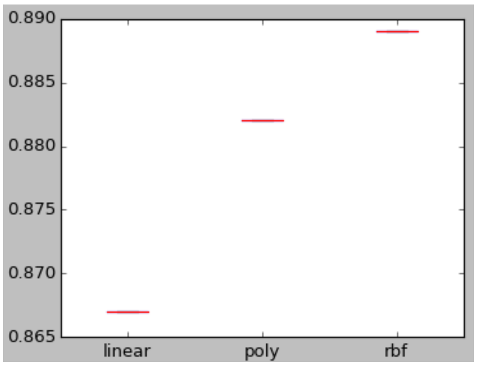

Fig.2. Estimation Accuracy SVM Classifier

5. Conclusion and Future Work

In this paper, we proposed a SVM machine learning method to model MOOC dropouts, based solely on log data based features and counts, Extracting the most relevant features is a crucial phase to produce a robust prediction model. Thus, we have taken an initial step towards early and accurately identifying dropout students, which can help instructors design intervention. We have tested multiple kernels of SVM method and we concluded that RBF kernel turn better.

Extracting the most relevant features is a primordial phase to produce a robust prediction model; we believe that the combination of relevant features and the addition of data identifying the student's profile will surely improve our prediction model. In the future work, we think to deepen mining the dataset and improve our model by applying techniques of optimization of the SVM because to have a good precision of the SVM is to optimize the separating hperplan.

References

[1] J. E. Luaran, “Massive Open Online Course (MOOC),” p. 165, 2013.

[2] K. Jordan, “Massive Open Online Course Completion Rates Revisited: Assessment, Length a...: EBSCOhost,” vol. 16, no. 3, pp. 341–358, 2015.

[3] S. L. Miller, “Teaching an Online Pedagogy MOOC,” MERLOT J. Online Learn. Teach., vol. 11, no. 1, pp. 104–119, 2015.

[4] S. Ardchir., M.A.Talhaoui, M.Azzouazi (2017) Towards an Adaptive Learning Framework for MOOCs. In: Aïmeur E., Ruhi U., Weiss M. (eds) E-Technologies: Embracing the Internet of Things. MCETECH 2017. Lecture Notes in Business Information Processing, vol 289. Springer, Cham

[5] C. Gütl, R. H. Rizzardini, V. Chang, and M. Morales, “Attrition in MOOC: Lessons Learned from Drop-Out Students,” Commun. Comput. Inf. Sci., vol. 446 CCIS, pp. 37–48, 2014.

[6] A. Margaryan, M. Bianco, and A. Littlejohn, “Instructional quality of Massive Open Online Courses (MOOCs),” Comput. Educ., vol. 80, pp. 77–83, 2015.

[7] D. Clow, “MOOCs and the funnel of participation,” Proc. Third Int. Conf. Learn. Anal. Knowl. - LAK ’13, p. 185, 2013.

[8] R. F. Kizilcec, C. Piech, and E. Schneider, “Deconstructing Disengagement : Analyzing Learner Subpopulations in Massive Open Online Courses,” Lak ’13, p. 10, 2013.

[9] M. Kloft, F. Stiehler, Z. Zheng, and N. Pinkwart, “Predicting MOOC Dropout over Weeks Using Machine Learning Methods,” Proc. 2014 Conf. Empir. Methods Nat. Lang. Process., pp. 60–65, 2014.

[10] S. Halawa, D. Greene, and J. Mitchell, “Dropout Prediction in MOOCs using Learner Activity Features,” eLearning Pap., vol. 37, no. March, pp. 1–10, 2014.

[11] G. Balakrishnan and D. Coetzee, “Predicting student retention in massive open online courses using hidden markov models,” Electr. Eng., 2013.

[12] W. Xing, X. Chen, J. Stein, and M. Marcinkowski, “Erratum: Corrigendum to ‘Temporal predication of dropouts in MOOCs: Reaching the low hanging fruit through stacking generalization’ (Computers in Human Behavior (2016) 58 (119–129) (S074756321530279X)(10.1016/j.chb.2015.12.007)) ,” Comput. Human Behav., vol. 66, p. 409, 2017.

[13] A. L’Heureux, K. Grolinger, H. F. Elyamany, and M. A. M. Capretz, “Machine Learning With Big Data: Challenges and Approaches,” IEEE Access, vol. 5, pp. 7776–7797, 2017.

[14] KDD cup 2015. The website may be down already. https://kddcup2015.com/

[15] C. Cortes and V. Vapnik, “Support-Vector Networks,” Mach. Learn., vol. 20, no. 3, pp. 273–297, 1995.