Identifying student behavior in MOOCs using Machine Learning: goals and challenges

Vanessa Faria de Souza
Graduate Program of Informatics in Education, Universidade Federal do Rio Grande do Sul (UFRGS),
Porto Alegre, Brazil.

Gabriela Trindade Perry (Corresponding author)
Graduate Program of Informatics in Education, Universidade Federal do Rio Grande do Sul (UFRGS),
Porto Alegre, Brazil.

Abstract

This paper presents the results literature review, carried out with the objective of identifying prevalent research goals and challenges in the prediction of student behavior in MOOCs, using Machine Learning. The results allowed recognizingthree goals: 1. Student Classification and 2. Dropout prediction. Regarding the challenges, five items were identified: 1. Incompatibility of AVAs, 2. Complexity of data manipulation, 3. Class Imbalance Problem, 4. Influence of External Factors and 5. Difficulty in manipulating data by untrained personnel.

Keywords: Machine Learning; MOOC; Student behavior

1. Introduction

Industry 4.0 – characterized by automation and data exchange in manufacturing technologies - is bringing great changes to society, education and to scientific development, which strongly affect the way people act and think. One of the consequences is the growth of Distance Educationinitiatives in several countries. In Brazil for example, it has been established in a comprehensive way, as the Brazilian Association of Distance Education's 2016 Census shows 7.2% of students are benefiting from it (ABED, 2016). Perhaps the most recent innovation, regarding distance education, was the expansion of MOOC (Massive Open Online Course) offers which have been attracting thousands of students, with significant enrollment numbers in recent years (Romero & Ventura, 2016). In 2012, edX, a non-profit startup created by Harvard University and the Massachusetts Institute of Technology, had 370,000 students in their first official courses. Coursera, founded in January 2011, has reached 1.7 million registered students and is growing "faster than Facebook", according to Wang, Hu & Zhou (2018). A MOOC aboutArtificial Intelligence, Stanford offered in 2011, attracted 160,000 students (Wang, Hu & Zhou, 2018). Romero & Ventura (2016), Wang, Hu & Zhou (2018). Greene, Oswald & Pomerantz (2015) and Hew, Qiao & Tang (2018) emphasize MOOC is a model of teaching and learning characterized by sustainable education.

Even so, this education model has failed to provide evidence to support its sustainability capability. More specifically, student attrition and dropout is the main limiting factor in consolidating MOOCs. An example is MOOC Introduction to Computer Science, offered by Harvard University in 2012, which had 150,349 students registered and 1388 students earning degrees (a 92% dropout). The MOOC Electronic Circuits course offered by MITx in March 2012 had 154,763 registered participants, and 7157 completed the course, (approximately 95% dropped out). This MOOC was offered again in September 2012 by edX, and had 46,000 registrations and 3008 conclusions, an average of 93% students leaving before completing [12]. The strong contrast between the registration fee and the abandonment rate increases skepticism about the sustainability of the MOOC. How to improve quantification is a key issue for the consolidation of this educational model (Wang, Hu & Zhou, 2018). We consider it would be noted that it is virtually impossible to provide the same quality of support in a class with thousands of students compared to ordinary classes with a few dozen students - even in classes with few students it might be difficult to know each student and identify their needs. Machine Learning algorithms might assist platform developers and teachers with this problem.

Nevertheless, there are benefits either for MOOC users – besides being mostly free and allowing access to world class specialists – and for providers. The large amount of data generated by the interactions opens up new possibilities for studying and understanding student behavior (Onah, Sinclair & Boyatt, 2014).There is a huge potential for studies with new and freely available methods and tools,which might assist in questions like these. For example, Educational Data Mining (EDM), an area of interdisciplinary research that deals with the development of methods to explore data originated in the educational contexts (Romero & Ventura, 2016). Learning Analytics (LA) is another emerging area of research. It stands for the measurement, collection, analysis and reporting of data about students and their contexts, with the objective of understanding and optimizing learning and the environments in which it occurs (Xing et al, 2015).

For this reason, our goal is toelaborate a literature to put togetherthe more prevalent goals and challenges regarding student behavior in MOOCs, using Machine Learning algorithms. The papers selected for this review were published from 2014 to 2018.

2. MOOCs

According to Hew, Qiao & Tang (2018), MOOCs are an opportunity for training and empowering the population, changing the way we teach and learn, thus requiring a new posture of educational institutions and professionals. However, one of the major challenges highlighted in this model is attrition and dropout. Although many enroll in the course, the number of completion is very small. Due to the peculiarities of these courses, Xing et al (2016) consider that calculating the dropout percentage over the total number of enrollees in snot the best way to measure MOOC efficiency, since the goals of students enrolled in face-to-face courses are generally the same, whereas students of MOOCs enroll for a multitude of reasons. This is another reason why identifying ways to reduce abandonment rates in MOOCs is a challenging task.

In the MOOC context, methods such as focus group, surveys, interviews and observations to collect data in this regard might not be very sufficient, because they are time-consuming and limited when the population is so large and diversified (Xing et al, 2015).Nevertheless, groups of researchers in the field of computer science have implemented models using machine-learning techniques to identify trends in the behavior of MOOC students. They make predictions by means of computational systems more suited to the large scale of data they manipulate. Many Brazilians have dedicated themselves to such studies, such as Manhaes et al (2011), Rodrigues, Medeiros & Gomes (2013) and Gotardo, Cereda & Hruschka (2013), as well as international researchers such as Hew, Xiao & Tang (2018), Wang, Hu & Zhou (2018), Greene, Oswald & Pomerantz (2015), Xing et al (2016) and Durksen et al (2016).However, due to characteristics and limitations of Machine Learning algorithms, this task, even with all current technological advances, is difficult to execute (Xing et al, 2015). Thus, the greatest challenge is to develop a method capable of predicting students' behavior, to enable the intervention of teachers, tutors, administrators and others, in order to rescue the student before he or she leaves the course (Wang, Hu & Zhou, 2018).

Thus we argue for the need for a study that lists the main studies carried out in the area in recent years and to describe the main advances and their respective challenges.

3. Machine Learning

Machine Learning is a subfield of Artificial Intelligence, dedicated to the development of algorithms and techniques that allow the computer to improve its performance in a given task. It is closely linked to Data Mining and Statistics. This area of research focuses on the properties of statistical methods, as well as their computational complexity (Sing & Purohit, 2015). Amongst its application are natural language processing, search engines, medical diagnostics, bioinformatics, speech recognition, handwriting recognition, computer vision and robot locomotion and prediction systems (Sing & Purohit, 2015).

A category of algorithms with wide use is classifiers, which are useful to classify unknown cases (Wu et al, 2008). According to Wu el at (2008), the most used classifier algorithms are: Naive Bayes, Support Vector Machines, Tree-Based Methods (such as C4.5 and Random Forest), IBK, Adaptive Boosting (Ada-Adaptive Boosting), Logistic Regression and Rule Based Method.

4. Research procedures

According to Vosgerau & Romanowski (2014), in a literature review the researcher gathers scientific documents obtained from a bibliographical survey, with the purpose of carrying out analyzes, contextualizing the problem under investigation. In this sense, in this research, a review of the literature was conducted with an emphasis on identifying studies that used Machine Learning to predict student behavior in MOOCs.

The databases used as source were IEEE Xplore Digital Library and Springer. We also look for paper in the journal Computers in Human Behavior, because it is well ranked and list MOOC, distance education and computational techniques in its scope. The databases, as well the journal, were chosen because they index papers related to Computer Science, Informatics in Education and Distance Education. The time frame was from 2014 to 2018. We used the strings "MOOC" and "MOOC" AND "Machine Learning" as search terms. The inclusion criteria were: (i) contain the terms "MOOC" or "Massive Open Online Course" in the title of the article; (ii) be available in English or Portuguese; (iii) contain the term "Machine Learning" in the title or in the keywords; (iv) present the use of machine learning in the prediction of student behavior in MOOCs. Exclusion criteria were: (i) using another idiom; (ii) do not present information about behavior prediction.

5. Results

Using the words "MOOC" and "MOOC" AND "Machine Learning", 55 papers were found. These papers were analyzed, and 11 were selected because they were in accordance with the fourth inclusion criterion, which established that they should present the use of Machine Learning in predicting student behavior in MOOCs. In the Journal Computers in Human Behavior 2 papers were found, and one paper that was not in English or Portuguese language was excluded. In the IEEE Xplore Digital Library 8 papers were found and Springer database 1 paper. The papers we analyzed and detail in the next sections are listed in Table 1. A summary of each paper can be found in Table 2.

Table 1 Papers selected for the literature review.

IEEE Xplore Digital Library	R. S. Baker, D. Lindrum, M. J. Lindrum, D. Perkowski, "Analyzing early at-risk factors in higher education e-learning courses". Students at Risk: Detection and Remediation, 2015.
	Y. Chen, Q. Chen, M. Zhao, S. Boyer, K. Veeramachaneni, H. Qu, "DropoutSeer: Visualizing learning patterns in Massive Open Online Courses for dropout reasoning and prediction," in IEEE Conference on Visual Analytics Science and Technology (VAST), Baltimore, MD, USA, 2016, pp. 111–120.
	N. Periwal, K. Rana, "An Empirical Comparison of Models for DropoutProphecy in MOOCs," in International Conference on Computing, Communication and Automation (ICCCA), Greater Noida, India, 2017, pp. 906–911.
	L. Wang, G. Hu, T. Zhou, "Semantic Analysis of Learners Emotional Tendencies on Online MOOC Education," Sustainability V. 10, N. 192, 2018.
	K. F. Hew, C. Qiao, Y. Tang, "Understanding Student Engagement in Large-Scale Open Online Courses: A Machine Learning Facilitated Analysis of Student's Reflections in 18 Highly Rated MOOCs,"International Review of Research in Open and Distributed Learning, V. 19, N. 3, 2018, pp. 69-93.
	J. Liang, C. LI, L. Zheng, "Machine Learning Application in MOOCs: Dropout Prediction," in 11th International Conference on Computer Science & Education (ICCSE 2016), Nagoya University, Japan, 2016, pp. 752–57.
	B. Hong, Z. Wei, Y. Yang, "Discovering Learning Behavior Patterns to Predict Dropout in MOOC," in 12th International Conference on Computer Science and Education (ICCSE), Houston, TX, USA, 2017, pp. 700–704.
	S. Jiang, A. Williams, K. Schenke, M. Warschauer, D. O'dowd, "Predicting MOOC performance with week 1 behavior," in 7th International Conference on Educational Data Mining, 2014.
Springer	S. Halawa, D. Greene, J. Mitchell, "Dropout prediction in moocs using learner activity features," in Proceedings of the European MOOC Summit (EMOOCs 2014)Lausanne, Switzerland, 2014.
Comp	J. A. Ruipérez-Valiente, p. J. Muñoz-merino, d. Leony, c. D. Kloos, "ALAS-KA: A learning analytics extension for better understanding the learning process in the Khan Academy platform", Computers in Human Behavior,

	V. 47, 2015, pp. 139-148.
	W. Xing, Chenx., J. Stein,M.Marcinkowski, "Temporal predication of dropouts in MOOCs: Reaching the low hanging fruit through stacking generalization,"Elservier, Computers in Human Behavior V. 58, 2016, pp. 119-129.

Table 2 Summary of each paper.

Authors	**Number of students**	**MOOC platform**	**Machine Learning algorithm**	**Goal**	**Challenges**
Jiang et al. (2014)	37.933	Coursera	Logistic regression	Dropout predictor	Influence of external factors
Halawa, Greene & Mitchell (2014)	Not informed	Not informed	Logistic regression	Student classification/grouping	Influence of external factors
Ruipérez-Valiente et al. (2015)	564	Khan Academy	Not informed	Student classification/grouping	Complexity of data manipulation
Baker et al. (2015)	70.000	Soomo Learning Environment	Logistic regression	Dropout predictor	Influence of external factors
Liang, Li & Zheng (2016)	200.904	XuetangX	Decision Tree	Feature Engineering	Complexity of data manipulation
Xing et al. (2016)	Not informed	Coursera, edX, Udacity,	Naïve Bayes, Decision Tree	Student classification/grouping	Complexity of data manipulation
Chen et al. (2016)	Not informed	Coursera, edX	Logistic Regression, Random Forest, Nearest-Neighbors, with cross validation	Student classification/grouping	MOOC platform incompatibility
Hong, Wei &Yang (2017)	96.529	XuetangX	Random Forest, Support Vector Machine, Multi Nomial Logistic Regression	Dropout predictor	Complexity of data manipulation
Periwal & Rana (2017)	235.772	MITx, HarvardX	K-Nearest Neighbor, Naïve Bayes, Decision Tree, Logistic Regression	Dropout predictor	Class Imbalance problem
Wang, Hu & Zhou (2018)	18.234	Chinese University (not informed)	Personalized algorithm	Dropout predictor	Data manipulation
Hew, Qiao & Tang (2018)	5.884	Coursetalk	K-Nearest Neighbor, Gradient Boosting Trees, Support Vector Machines, Logistic regression, Naive Bayes	Dropout predictor	Influence of external factors

5.1 Advances in student behavior prediction in MOOCs

The following the research goals were identified in the literature review.

5.1.1 Student classification/grouping: Chen et al (2016) created a system called DropoutSeer based on Machine Learning, which helps instructors and researchers to analyze the relationship between student performance and dropout. By grouping students with similar performance, the authors seek to identify students in need for assistance. Ruipérez-Valiente et al (2015) also developed a visual module. It extends the learning analysis support for the Khan Academy platform, so the information could be used to group students. Xing et al (2016) present a grouping generalization approach for constructing more robust and accurate forecasts about temporal prediction of dropouts in MOOCs, based on classification algorithms. In the research by Halawa, Greene & Mitchell (2014) an "integrated predictor" consisting of two components is presented. The “Predictor of the active mode” operates while the student is logged in and the “Predictor of absence mode” operates when the student has been absent for a certain period of time. The goal is to assist instructors and administrator to make appropriate interventions for each case.

5.1.2 Dropout predictors: Periwal & Rana (2017) described in their work four models using supervised learning algorithms with labeled data. Using these models, the authors were able to identify a range of characteristics that affect the prediction of dropout. In this sense, Jiang et al (2014) used logistic regression algorithms for predictions and developed a method to identify user interactions with online distance education platforms, using student's performance in the first week as a predictor of its permanence and success in the MOOC. The authors also report that one of the factors influencing the prediction of student abandonment is external factors, which are difficult to identify. In this perspective, Baker et al (2015) report the elements they characterized as early predictors of student success, thus defining the information that is worth investigating for the study of dropout in MOOCS, providing hints for interventions of teachers and instructors. The work by Hew, Qiao & Tang (2018) also brings an insight into the factors that affect student participation in MOOCs. Similarly, the authors rely on the application of machine learning classifiers to analyze sentences posted by the students. Wang, Hu & Zhou (2018) developed a SMA to predict students' emotional tendencies in order to analyze the acceptance of the courses based on data such as tasks, comments, forums and other information in MOOC platforms. The method can recognize students' emotional tendencies through semantic analysis, which provides an effective solution to personalized MOOC teaching, which can help those involved, achieve a reduction in dropout. Hong, Wei & Yang (2017) proposed a technique in which they apply two-layer cascade classifiers with a combination of three different classifiers to predict dropout. Experimental results indicate that this technique is promising, reaching 97% accuracy. Finally, Liang & Zheng (2016) used the rather unique procedure called “Feature Engineering”, an aid in the application of Machine Learning algorithms. This rationale is to create new data from collected data, which leads to improved forecasting and performance, via application of mathematical functions to the data vector. An analogy can be made with the Body Mass Index, which uses existing data to construct new data (not physiological, and therefore not part of the original database).

5.2 Challenges in student behavior prediction in MOOCs

Regarding the challenges of the prediction process, the results of the literature review pointed to the following items.

5.2.1 MOOC platform incompatibility: The main challenge highlighted by Chen el at (2016) concerns the structure of the MOOCs and student evaluation metrics, which are very diverse and evolve fast. For this reason, a unified solution, which could be used by several MOOC providers, is not yet possible.

5.2.2 Complexity of data manipulation: Hong, Wei & Yang (2017) and Liang & Zheng (2016) indicate that one of the main obstacles to prediction of abandonment in MOOCs is the available data, which need to be worked on before being used in algorithm training and in the process of prediction. Ruipérez-Valiente et al (2015) assert that dealing with the great diversity of student actions that can be captured by online platforms (for example, posting a question in a chat room and watching a video) is a complex task. In this perspective Xing et al (2016) add it is difficult to deal with the variability of data collected from MOOCs.

5.2.3 Class Imbalance problem: According to Periwal & Rana (2017), one of the major challenges of working with data from MOOCs is the issue of class imbalance. As the number of dropouts (major class) is much larger than the number of students who finish the course (minority) there is a need for a model that mitigates the effect of unbalanced data and predicts which students will actually dropout.

5.2.4 Influence of external factors: Jiang et al (2014) explain that external factors have a strong influence in the students' decision to leave – and it cannot be identified via interaction with the platform. Halawa, Greene &Mitchell (2014) say that such factors are quite heterogeneous and difficult to treat. In this same line of reasoning, baker et al (2015) report that one of the obstacles is that data on updated student profile is not available to the online MOOC platform. The work by Hew, Qiao & Tang (2018) corroborates these statements, noting that there is great complexity in determining why students drop out of MOOCs.

5.2.5. Difficulty in manipulating the methods by untrained personnel: According to Wang, Hu & Zhou (2018), Machine Learning algorithms are hard to understand by untrained personnel, such platform administrators, instructors and teachers. This is a challenge because to design a prediction routine using Machine Learning algorithms it is important to understand the problem, the users and the context – and this knowledge might not be available to programmers.

6. Conclusion

MOOCs have potential to be a sustainable model of education, providing students with a variety of possibilities and benefits. However, this scenario still seems distant, and the most worrying factor is the low completion rate. In, in this sense if those involved with MOOCs, such as platform administrators, teachers and tutors, could identify which students are most likely to give up, they could take actions to mitigate these dropouts or provide individualized assistance - Machine Learning algorithms could be useful in this sense. The process of identifying students at risk of leaving, even with all technological advances, usually is manual, subjective and not informed by data, depending mostly on the experience of teachers. This method is unable to meet the demands of MOOCs.

Through the literature review, 11 studies were identified, which presented relevant applications and results, showing that there are many researchers trying to improve MOOCs and make better use of the massive data generated by online platforms. The review reveals two important aspects: goals and challenges regarding application of Machine Learning algorithms in the MOOC context. Identifying goals in important to contextualize research questions and methods, while recognizing obstacles is important to foresee difficulties and avoid pitfalls.

In this scenario, Machine Learning is very opportune for treating data and extracting information in the large volume of data generated by student interactions with the MOOCs platforms. Even in the midst of large amounts of data, one can have a fairly accurate picture of student behavior, as shown by Jiang et al. (2014). These authors used Logistic Regression and developed a method that guaranteed that the performance of the student already in the first week could be a predictor of its permanence and success in the MOOC. In addition to predicting behavior, another frequent goal is grouping students, which may be the key to offering personalized assistance.

Difficulties were also encountered in the review. The first one is cross-platform incompatibility, which prevents data from being shared. Besides these platforms belong to different institutions, the course formats vary greatly, which makes proposing a general model very difficult. An unfolding of this problem is the complexity of data manipulation, requiring extensive treatment before used as input in prediction algorithms. Another obstacle refers to Class Imbalance Problem, characterized by one category being much larger than another. This disparity causes the algorithms to learn in a biased way, generating incorrect predictions. Another difficulty is the influence of external factors in student behavior, which is difficult to detect and that interferes with the students' attitudes. The last obstacle is the difficulty in manipulating the methods by untrained personnel, which indicates the need to develop systems that are more intuitive, so that the proposed solutions are not restricted to people with Computer knowledge.

References

[1] ABED - CENSO EAD 2016 [Online]. "Relatório analítico da aprendizagem a distância no Brasil", Available: http://www.abed.org.br [Acesso em 30 de setembro de 2018]

[2] A. Singh, A. Purohit, "A Survey on Methods for Solving Data Imbalance Problem for Classification,"International Journal of Computer Applications, V. 127, N.15, 2015, pp. 0975 – 8887.

[3] B. Hong, Z. Wei, Y. Yang, "Discovering Learning Behavior Patterns to Predict Dropout in MOOC," in 12th International Conference on Computer Science and Education (ICCSE), Houston, TX, USA, 2017, pp. 700–704.

[4] C. Romero, S. Ventura, S. "Educational data science in massive open online courses," Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, V. 7, N. 1, 2016.

[5] D. F. Onah, J. Sinclair, R. Boyatt, "Dropout rates of massive open online courses: behavioural patterns," in 14th, EDULEARN, EUA, 2014, pp. 5825-5834.

[6] D. S. R. Vosgerau, J. P. Romanowski, "Estudos de revisão: implicações conceituais e metodológicas," Revista Diálogo Educacional. Curitiba, vol. 14, n. 41, 2014, pp. 165-189.

[7] J. A. Greene, C. A. Oswald, J. Pomerantz, "Predictors of Retention and Achievement in a Massive Open Online Course," American Educational Research Journal, V. 52, N. 5, 2015, pp. 925–955.
[8] J. Liang, C. LI, L. Zheng, "Machine Learning Application in MOOCs: Dropout Prediction," in 11th International Conference on Computer Science & Education (ICCSE 2016), Nagoya University, Japan, 2016, pp. 752–57.
[9] J. A. Ruipérez-Valiente, p. J. Muñoz-merino, d. Leony, c. D. Kloos, "ALAS-KA: A learning analytics extension for better understanding the learning process in the Khan Academy platform", Computers in Human Behavior, Volume 47, 2015, pp. 139-148.
[10]K. F. Hew, C. Qiao, Y. Tang, "Understanding Student Engagement in Large-Scale Open Online Courses: A Machine Learning Facilitated Analysis of Student's Reflections in 18 Highly Rated MOOCs,"International Review of Research in Open and Distributed Learning, V. 19, N. 3, 2018, pp. 69-93.
[11]L. M. B. Manhaes, S. M. S. Costa, J. Zavaleta, G. Zimbrao, "Previsão de Estudantes com Risco de Evasão o Utilizando Técnicas de Mineração de Dados," in Proceedings of the 22th, SBIE, Campinas, Brasil, 2011, pp. 1500-1510.
[12]L. Wang, G. Hu, T. Zhou, "Semantic Analysis of Learners Emotional Tendencies on Online MOOC Education," Sustainability V. 10, N. 192, 2018.
[13]N. Periwal, K. Rana, "An Empirical Comparison of Models for DropoutProphecy in MOOCs," in International Conference on Computing, Communication and Automation (ICCCA), Greater Noida, India, 2017, pp. 906–911.
[14]R. Gotardo, P. Cereda, J. E. Hruschka, "Predição do Desempenho do Aluno usando Sistemas de Recomendação e Acoplamento de Classificadores," In Proceedings of the 24th, SBIE, Campinas, Brasil, 2013, pp. 2202-2212.
[15]R. L. Rodrigues, F. P. A. Medeiros, A. S. Gomes, "Modelo de Regressão Linear aplicado à previsão de desempenho de estudantes em ambiente de aprendizagem," in 24th, SBIE, Campinas, Brasil, 2013, pP. 607-616.
[16]R. S. Baker, D. Lindrum, M. J. Lindrum, D. Perkowski, "Analyzing early at-risk factors in higher education e-learning courses". Students at Risk: Detection and Remediation, 2015.
[17]S. Halawa, D. Greene, J. Mitchell, "Dropout prediction in moocs using learner activity features," in Proceedings of the European MOOC Summit (EMOOCs 2014)Lausanne, Switzerland, 2014.
[18]S. Jiang, A. Williams, K. Schenke, M. Warschauer, D. O'dowd, "Predicting MOOC performance with week 1 behavior," in 7th International Conference on Educational Data Mining, 2014.
[19]T. L. Durksen, M. W. Chu, Z. F. Ahmad, A. L. Radil, M. L. Daniels, "Motivation in a MOOC: a probabilistic analysis of online learners basic psychological needs," Springer, Soc. Psychol Educ., 2016.
[20]W. Xing, R. Wadholm, E. Petakovic, S. Goggins, "Group learning assessment: developing a theory-informed analytics". Journal of Educational Technology & Society, V. 18 N. 2, 2015, pp. 110-128,
[21]W. Xing, Chenx., J. Stein,M.Marcinkowski, "Temporal predication of dropouts in MOOCs: Reaching the low hanging fruit through stacking generalization,"Elservier, Computers in Human Behavior V. 58, 2016, pp. 119-129.

[22]X. Wu, V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, A. Mclachlan, B. Liu, P. S. Yu, Z. Zhou, M. Steinbach, D. J. Hand, D. Steinberg, "Top10 algorithms in data mining. Knowledge and Information Systems," Springer, Knowl Inf Syst, V. 14, 2008, pp. 1–37.

[23]Y. Chen, Q. Chen, M. Zhao, S. Boyer, K. Veeramachaneni, H. Qu, "DropoutSeer: Visualizing learning patterns in Massive Open Online Courses for dropout reasoning and prediction," in IEEE Conference on Visual Analytics Science and Technology (VAST), Baltimore, MD, USA, 2016, pp. 111–120.