Exploring Stopout Prediction and Transfer Learning in MOOCs

Alex Huang, Erik Hemberg, Una-May O'Reilly, MIT CSAIL

There is significantly larger student stopout in online courses compared to traditional classroom settings. This issue needs to be addressed in order to use MOOCs to improve career advancement and development. Being able to identify at-risk learners early would support more personalized interventions and potentially decrease stopout rates [2]. Machine learning models for stopout can only be trained using historical data from past courses. The challenge is that they then often do not transfer accurately to a different course or new set of learners when they are used for stopout prediction because assumptions integrated within the model no longer hold. *Transfer learning* refers to how effectively we can use a machine learning model trained on data from one class to predict stopout in other classes [1].

We explore the problem of stopout prediction and transfer learning in MOOCs. At the first week a learner does not submit a required assignment, we labeled her as having stopped out. In contrast to prior works, we model all learners without segregating them by how engaged they are and we use a combination of 14 weekly engagement related features extracted from clickstream (e.g. number of distinct problems attempted, average time between submissions and assignment due date) and forum data (e.g. number of replies to a learner's post) to train a model that can identify students that are at risk of dropout. To train the machine learning models for each week, we use 75% of the data while using the remaining 25% as a test set for evaluation. We insist on using a standard Machine Learning library[1] with default parameters as a non-expert would. We use data from 3 edX MOOCs, statistics per class are: 1. 24.00x: 19,376 unique users and 772,460 distinct feature values, 2. 3.086x: 10,370 unique users and 308,207 distinct feature values, 3. 6.002x: 22,670 unique users and 827,221 distinct feature values,

As expected, direct model transfer is inaccurate, while combining models from different courses reduces transfer prediction error [3, 1]. Inaccuracy could be explained by extremely sparse feature vectors, i.e. some learners had many zero features. To address this we first removed every learner with only 1 non-zero feature, then trained and recorded stopout prediction model accuracy. Then we repeated this for learners with 2 non-zero features, etc. We found that a dataset with learners having a minimum of 3 non-zero features yields more accurate stopout prediction models than 2 or 4 non-zero features. We compare different stopout prediction modeling methods observing that a multilayer perceptron method performs very well. The logistic regression method is also adequate and yield a clearer understanding of why a learner is at risk. Using the same machine learning library to automatically tune the stopout prediction models is likely to further improve transfer learning accuracy.

References

[1] Sebastien Boyer and Kalyan Veeramachaneni. Transfer learning for predictive models in massive open online courses. In International Conference on Artificial Intelligence in Education, pages 54–63. Springer, 2015.

[2] Scott Crossley, Luc Paquette, Mihai Dascalu, Danielle S McNamara, and Ryan S Baker. Combining click-stream data with nlp tools to better understand mooc completion. In Proceedings of the Sixth International Conference on Learning Analytics & Knowledge, pages 6–14. ACM, 2016.

[3] Jacob Whitehill, Kiran Mohan, Daniel Seaton, Yigal Rosen, and Dustin Tingley. Mooc dropout prediction: How to measure accuracy? In Proceedings of the Fourth (2017) ACM Conference on Learning@ Scale, pages 161–164. ACM, 2017.

[1] `http://scikit-learn.org/stable/`