An Early Prediction of At-Risk Students under Online Learning Environment: Using Informative Features from Students' Clickstream Data

Shi Pu, Purdue University, spu@purdue.edu

Yu Yan, Penn State University, yzy122@psu.edu

Qiong Zhu, Penn State University, qxz132@psu.edu

The proliferation of distance education brings new types of data that are traditionally not available in education. Among them, students' clickstream data on virtual learning environment is particularly relevant to identify at-risk students. First of all, the clickstream data is available at an early stage (even before the start of the course), making it suitable for early prediction. Second, the data has potential to reveal students' engagement and learning strategies (e.g., Sinha, Jermann, Li, & Dillenbourg, 2014), thereby are promising candidate features. Last, the data is readily available or easier to collect than other new sources of data, like eye tracking and postures.

Recent research has explored several possibilities of utilizing clickstream data to predict students' success in the context of Massive Online Open Courses (MOOC). It usually involves aggregating students' clicks at activity level (e.g., Kloft, Stiehler, Zheng, & Pinkwart, 2014) or appending their aggregated clicks in each temporal segments (e.g., Xing, Chen, Stein, & Marcinkowski, 2016). Though achieving inspiring success, their approaches are still far from refined. Aggregating clickstream data loses subtle information on student behavior (e.g., how the clicks fluctuate across time). In addition, this method lacks a strong connection to the established literature on student learning, making it only useful for the purpose of forecasting.

This study argues that a better utilization of clickstream data is to investigate how students' effort is distributed across time and activity types. We hypothesized that the summary statistics of the effort distribution can be informative features to predict students' success. In addition, the student effort distribution can reveal student self-regulation and learning strategies, which not only improves the prediction but also provides useful implications for education practitioners and researchers.

We have tested our idea on a recent publicly available dataset in the Open University (UK) (Kuzilek, Hlosta, & Zdrahal, 2017), a world-leading distance education institution. The dataset includes seven courses taken by 28,785 students in four time periods. We extracted students' first 30 days clickstream data since the course start and performed an early prediction for students' course success (failure) using both Random Forest and XGBoost. The preliminary results indicate that several summary statics of the effort distribution are promising features to identify at-risk students, including the entropy, variance, peak, and longest positives. The result suggests that the model powered by these features is able to reach a 0.48 f1 score, and 0.711 area under the curve (AUC) on the testing data. And the model performance is consistently better than a conventional model that relies only on students' sum of clicks over subjects or time. As students' click stream data is available in the early stage, these features will be useful to build early prediction models.

Reference

Sinha, T., Jermann, P., Li, N., & Dillenbourg, P. (2014). Your click decides your fate: Inferring information processing and attrition behavior from mooc video clickstream interactions. *arXiv preprint arXiv:1407.7131*.

Kloft, M., Stiehler, F., Zheng, Z., & Pinkwart, N. (2014). Predicting MOOC dropout over weeks using machine learning methods. In *Proceedings of the EMNLP 2014 Workshop on Analysis of Large Scale Social Interaction in MOOCs* (pp. 60-65).

Xing, W., Chen, X., Stein, J., & Marcinkowski, M. (2016). Temporal predication of dropouts in MOOCs: Reaching the low hanging fruit through stacking generalization. *Computers in human behavior*, *58*, 119-129.

Kuzilek, J., Hlosta, M., & Zdrahal, Z. (2017). Open University learning analytics dataset. *Scientific data*, *4*, 170171.