Who You Are or What You Do: Comparing the Predictive Power of Demographics vs. Activity Patterns in Massive Open Online Courses (MOOCs)

Christopher Brooks
School of Information
University of Michigan
Ann Arbor, MI 48103
brooksch@umich.edu

Stephanie Teasley
School of Information
University of Michigan
Ann Arbor, MI 48103
steasley@umich.edu

Craig Thompson
Department of Computer Science
University of Saskatchewan
Saskatoon, SK S7N 5C9
craig.thompson@usask.ca

L@S 2015, Mar 14-18, 2015, Vancouver, BC, Canada
ACM 978-1-4503-3411-2/15/03.
http://dx.doi.org/10.1145/2724660.2728668

Abstract
Demographics factors have been used successfully as predictors of student success in traditional higher education systems, but their relationship to achievement in MOOC environments has been largely untested. In this work we explore the predictive power of user demographics compared to learner interaction trace data generated by students in two MOOCs. We show that demographic information offers minimal predictive power compared to activity models, even when compared to models created very early on in the course before substantial interaction data has accrued.

Author Keywords
MOOC; learning analytics; predictive modeling; student success; demographics, interaction data mining; activity

ACM Classification Keywords
K.3.1 Computer Uses in Education

Introduction

Demographics, which we broadly construe here to be information known about learners that does not change while they interact with learning systems, are a valuable tool in building models to predict learner success. For example, [1] used performance demographics, such as previous grade point averages, to predict student success in traditional higher education courses, while [2] used sociocultural demographics, such as gender and military service, with the same goal. A major theme of the research in this area is to inform the development of *early warning systems* [3] which can identify when learners are struggling and intervene, typically through dashboards or providing awareness cues.

MOOCs provide a very different learning environment than traditional place-based educational institutions; the stakes are lower, attrition rates are higher, learners have different motivations and goals, and the information about the prior educational experience of learners is limited. In attempts to better understand MOOC students, vendors and institutions offering MOOC content have begun to actively survey learners in an attempt to generate demographic profiles as well as to identify students' goals and preferences. However survey response rates tend to be quite low, and the usefulness this information is unknown. Therefore, the question we address here is:

To what extent is demographic information valuable for building predictive models of student success in MOOCs?

To explore this question, we compare the value of demographic data versus learner activity data when building predictive models of student success. Because definitions of success are numerous (perhaps more so in MOOCs than in traditional higher education), we contextualize this work around the basic concept of academic success: the student passes a course. In a MOOC, this means that the student has scored high enough on tests or quizzes to earn a certificate of completion.

Activity data in online learning systems is collected through clickstream logging, and we use the temporal interaction modeling techniques described in [4] to process MOOC clickstreams and generate comparative models. These models are absent of learner demographics, and are based only on the regularity by which a learner interacts with resources in the online courseware. Finally, we compare both the demographic and activity approaches to a hybrid model that includes both demographic and activity data.

Methodology

We considered data from a MOOC course which was a broadly accessible introduction on Internet technologies. This MOOC ran three times each over 11 weeks, with tens of thousands of learners enrolled in each offering.

To capture demographic data, we used three sources including (i) the Coursera demographic survey, which asks questions about student gender, age, race, and language capabilities, (ii) data about whether a learner was paying for the course or not (e.g. in the *signature track*) as well as their signup date, and (iii) geolocation information derived from the learners clickstream data, including country, state/province, and city. The total number of demographic attributes (features) used for our models was 97.

Week	1	2	3	4	5	6	7	8	9	10	11	End
Hybrid	0.344	0.502	0.595	0.645	0.687	0.698	0.711	0.711	0.741	0.751	0.743	0.755
Activity	0.328	0.495	0.591	0.639	0.686	0.689	0.705	0.725	0.754	0.775	0.781	0.810
Difference	0.016	0.007	0.004	0.006	0.000	0.010	0.007	-.014	-.014	-.024	-.037	-.054

Table 1: Weekly average κ for hybrid and activity models.

For the activity-based models, three different kinds of content accesses were considered: lecture video views, quiz attempts, and discussion forum accesses. As the data is based on end-user activity, predictive models were generated for each day of the course. The total number of attributes (features) used for data mining ranged between 371 (for the first day of the course) and 1,091 (for the last day of the course). The third comparison, the hybrid models, were created through the merger of the data from the demographic and activity-based models described previously.

The first two sessions of the course were used to train models to predict success, while the third session of each course was used as a validation set. The Weka toolkit version 3.7.12 was used with the J48 classifier. The training of the model was done by restricting the data to only those who had filled out both the demographics (7,791 of 61,820 learners, 12.6%) and those who had interacted in the course (23,818 of 61,280 learners, 38.9%) which was a total of 4,130 learners (6.7%) after balancing. The validation was done on the full unbalanced third session of the course (23,902 learners), to simulate the performance of an actual early warning system.

Results

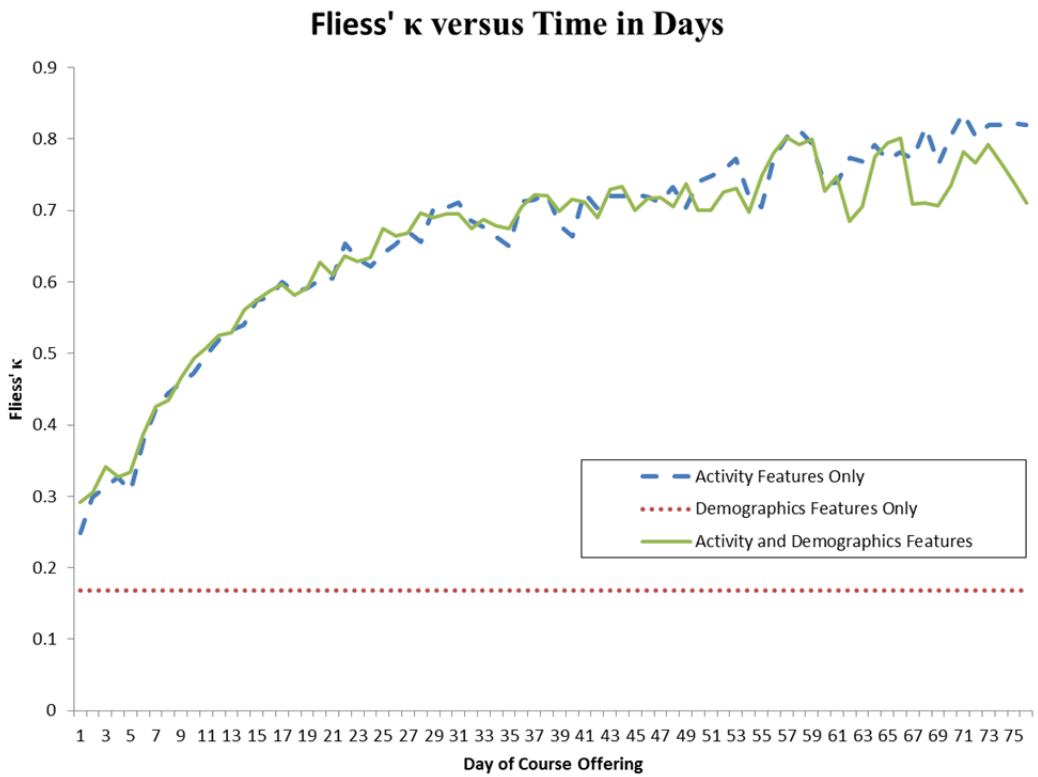
Shown in Figure 1 are the accuracies (in Fliess' κ) of all three models over time. The dotted straight line in the figure is the accuracy of demographics features alone, which does not change over time. The dashed line is the activity-based modeling approach, while the solid line is the hybrid model.

The demographics model performed poorly, with κ=0.17. Even at the beginning of the course when the clickstream data is minimal, the activity model has a κ=0.24 which outperformed demographics models considerably. Interestingly, the addition of demographics to the activity model made minimal difference. Table 1 shows the difference in accuracy between the activity and hybrid models broken down on a weekly averages. It is notable that for the first seven weeks of the course the hybrid model has a slight ($\kappa \leq 0.01$) benefit over the activity model, but for the last five weeks of the course it has a slight ($\kappa \leq 0.05$) disadvantage over the activity model.

Conclusions

Demographics, including your location and whether you are paying for a certificate, have minimal predictive power when determining the academic achievement of learners enrolled in MOOCs. While there are many exciting questions this work surfaces (e.g. Is this true in other kinds of MOOCs, or just this one? Are some demographics of particular value versus others?), the one most interesting to us for follow up is understanding whether this is true for traditional residential education as well. In traditional early warning systems, student demographics (which include

Figure 1: Fliess' κ versus time in days.



prior achievement, a value often not available in MOOCs) models) are used for predictive modeling. Is this simply because activity models are starved of data in traditional blended formats, or are there key demographic features, such as prior achievement, which may help increase the quality of success models in MOOC environments as well?

References

[1] S. Jayaprakash, E. Moody, E. Lauría, J. Regan and J. Baron, "Early Alert of Academically At-Risk Students: An Open Source Analytics Initiative," *Journal of Learning Analytics,* vol. 1, no. 1, 2014.

[2] M. Sharkey, "Academic analytics landscape at the University of Phoenix," in *1st International Conference on Learning Analytics and Knowledge (LAK11)*, Banff, AB, 2011.

[3] L. Macfadyen and S. Dawson, "Mining LMS data to develop an "early warning system" for educators: A proof of concept," *Computers & Education,* vol. 54, no. 2, pp. 588-599, 2010.

[4] C. Brooks, C. Thompson and S. Teasley, "A Time Series Interaction Analysis Method for Building Predictive Models of Learners using Log Data," in *5th International Conference on Learning Analytics and Knowledge 2015 (LAK15)*, Poughkeepsie, NY, 2015.