Identifying Successful Learners from Interaction Behaviour

Judi McCuaig
School of Computer Science
University of Guelph
Guelph, ON, Canada
judi@uoguelph.ca

Julia Baldwin
School of Computer Science
University of Guelph
Guelph, ON, Canada
baldwinj@uoguelph.ca

ABSTRACT

The interaction behaviours of successful, high-achieving learners when using a Learning Management System (LMS) are different than the behaviours of learners who are having more difficulty mastering the course material. This paper explores the idea that conventional Learning Management Systems can exploit data mining techniques to predict the success or failure of students without requiring the results of formal assessments. This paper describes a study with a second semester computer science class that shows that the success or failure of a learner can be predicted using information about learner interactions with course materials and learner self-reports of subject matter confidence.

1. INTRODUCTION

A Learning Management System gathers and records a rich set of data about the educational materials for a course and about the learners using the course. The data is used by the LMS in producing reports about the learners and the educational materials. Data gathered by a LMS is also frequently used in data mining applications to make predictions about learners. Three common uses for the data are: to predict the learners' likely academic performance and adjust either the LMS or the face to face instruction accordingly, to automatically adapt the LMS or the content to the needs or learner preferences of individual learners, or to predict learner affect in an effort to monitor and react appropriately to learner engagement with the course.

This paper presents a study that shows that successful learners exhibit different interaction behaviours with the LMS than less successful learners. Our results indicate that the results of formal assessments are unnecessary to accurately predict which learners will be successful with the course content and which learners will struggle with the course.

2. MINING LMS METADATA

A common goal for mining the metadata produced by learning management systems is to predict learner achievement. Most achievement-predicting systems produce a prediction of final grade (see for example [**?**]). Others predict or model the knowledge that the learner has mastered, with the goal of adapting the materials in the course to more closely match the learner's immediate learning needs (see for example [**?**]). The data mined by such systems almost always include the correctness or score for learning objects that generate grades, elapsed times for completing activities, and the learner interactions with the LMS.

LMS interaction data usually consists of a list of the elements viewed by learners as well as dates and timestamps to identify the viewing. Also, in many LMS the specific types of LMS objects (ie. forums, quizzes, help pages) are identified in the logs, making the type of learning object another piece of metadata that is frequently collected. All of this data can be used in learner classification activities.

3. MINING FOR LEARNER SUCCESS

A study was conducted during the winter of 2011 in a single semester computer science class on programming in the C language. 122 learners participated in the study, which captured a wide variety of data for the entire duration of the course. The work presented in this paper looks at the relationship between the learner's activities on the LMS and their ultimate success in the course. For this portion of the study, the learner's final grade was used as the ground truth representation of overall learner success. A ranking of final grades in a single course, regardless of the grade distribution, will most likely result in the more successful learners appearing in the top ranks, and the less successful learners appearing in the bottom ranks.

3.1 Data Collection

Throughout the semester experimenters collected information about participant interactions with course material. Data was obtained from the course LMS, from the Subversion (SVN) server, from in-class clickers, and from formal observations.

Data included information about the learners habits with respect to lecture attendance and participation, starting and finishing labs and assignments, and studying from online materials. The resulting data set included the dates and time(s) that participants read a lab or assignment description, dates and times that assignments and labs were submitted, the date and time that a participant had their lab graded (labs were graded in-person), lecture attendance, the dates that problem sets were completed, and the amount of time taken to complete quizzes and exams. Additional data was generated through the collection of self assessment surveys, which asked learners to rate their own confidence and mastery of the course material.

Each week, along with the self-assessment questions, learners were given an ungraded set of multiple choice problems to help them evaluate their understanding of the course con-

tent. The weekly problem set consisted of seven randomly selected questions. Every problem set consisted of 2 hard, 2 medium and 3 easy questions that spanned the entire course content. Learners had the option of indicating that a question was about content they had yet to learn rather than guess at an answer. A measure of the participant's self-reported confidence as well as the participant's success with problem sets was calculated for each week of the course.

Because participation in this study was voluntary, and not all students in the course participated, the data collection process could not interfere with the normal activities of the learners in the course. As a result, the study activities were ungraded and did not affect the academic grade of the learner. One of the side-effects of this restriction was that participants sometimes placed little importance on the study-specific activities (the problem sets and self-evaluation).

3.2 Analysis

A goal of this work is to determine if learner success (or failure) can be predicted using data that is passively, or semi-passively captured as the learner works through the course. For the purposes of this work, passively and semi-passively captured data is data that can be collected exclusively from the learner's interactions with the LMS, and without the need for separate action on the part of the course instructor or teaching assistants. In some cases the data may come from log files and in others it may be directly captured from questions asked of the learners.

3.2.1 Self Assessments

Independent self-assessment activities such as ungraded problem sets for self-checking and opportunities for self-reflection about personal confidence and progress are one type of data that can be semi-passively captured. The data capture requires input from the learner, but not from course administration or instructors.

During this study, participants completed weekly problem sets that were kept at the same difficulty for the entire semester. The problem sets were difficult for learners at the beginning of the semester and should have seemed easier as the learners mastered the course material. No problem sets were assigned in weeks 11 and 12. The data collected during the study shows that the median scores on the problem sets slowly increased over the semester. The increase is not dramatic, nor is it consistent every week, but higher median scores did occur in the last third of the semester as can be seen in Table 1.

The participants' weekly scores on the problem sets show a weak correlation with their final grade, however the total problem set score shows a higher correlation (.73) with final grade (see Figure 1).

Participants also assessed their own confidence in programming skills each week. When examined week-by-week, the correlation between learner confidence and final grade is consistently positive except for a one week anomaly shortly after a difficult quiz. The scores from the weekly confidence assessments were averaged across two separate 5 week periods in the course, weeks 0 through 5 and weeks 5 through 10 (no confidence assessments were done in weeks 11 and 12). The correlation between those mean confidence scores and participant final grade was calculated. The results can be seen in Table 2.

Table 1: Median Scores for Problem Sets by Week

Week	Mean	Median
0	3	3
1	4	4
2	4	4
3	4	4
4	4	5
5	4	4
6	4	4
7	5	5
8	4	4
9	5	5
10	4	5

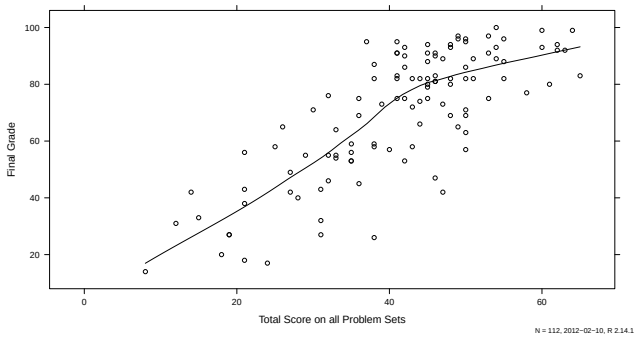

Figure 1: Problem Set Scores vs Final Grade

The increased correlation coefficient for the latter weeks of the course suggests that participants became more realistic about their own programming abilities as the semester progressed, and that in the latter half of a semester, participants may be a good source of estimations of their own success with the course.

3.2.2 Independent Course Work Habits

Learners with strong independent work habits are often more successful. Since assignments were the largest 'problems' given as independent homework in this course, the learner's interactions with assignment related information was examined to better understand the relationship between independent coursework interactions and overall success. As can be seen in Figure 2, there is a relationship between the total number of times assignments were viewed and the participant's final grade (a positive correlation coefficient of .42). Even though the correlation is moderate, the relationship between final grade and assignment reading habits bears more investigation.

In order to further investigate work habits, two additional pieces of data about student work habits were calculated from the LMS logs: the number of times participants used the course LMS (views), and the number of individual days that participants used the course LMS (days active). Days active is a count of the number 24 hour periods that the participant logged in to the site at least once. Views is a count every logged interaction that a participant had with the LMS.

The number of views of LMS material has a clear relationship with the final grade of the participant (see Figure 3)

Table 2: Mean Confidence Correlations with Final Grade

Weeks	Final Grade Correlation
0 through 4	.34
5 through 10	.60
Whole Semester	.52

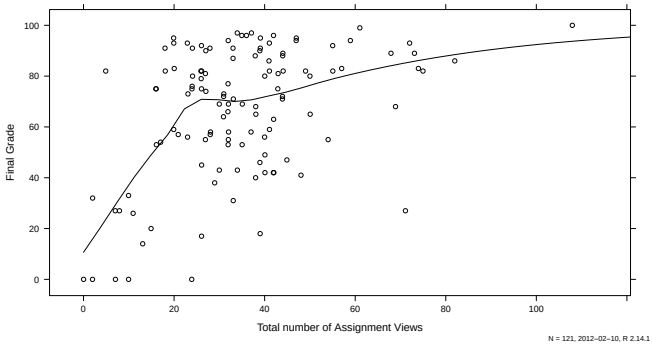

Figure 2: Total Assignment Views vs Final Grade

While the relationship is positive, a wide variance in grades is evident, especially around the 500 views point in the graph, indicating that a simple count of number of interactions is unlikely to discriminate between successful and unsuccessful learners. However, the total number of LMS views has a positive correlation coefficient of .56 with final grade.

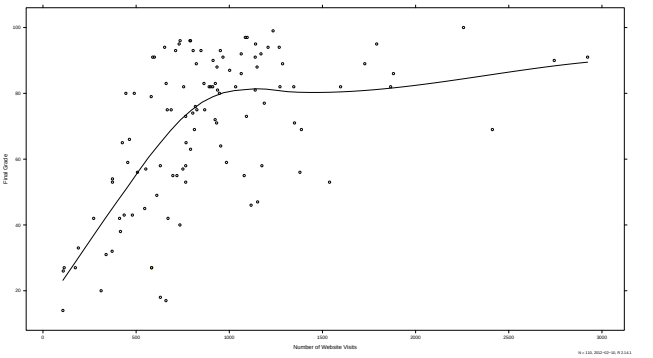

Figure 3: Total Course LMS Views vs Final Grade

The relationship between the number of days a participant was active on the LMS and their final grade is even stronger. As can be seen in Figure 4, the relationship appears to be nearly linear and the variation is similar for the entire graph. The correlation coefficient between the total number of days active and the participants' final grade is .73. It is as strong a relationship as the problem set scores, but the data capture requirements are completely passive, while problem set scores require direct action from the learner and the ability to automatically grade the practise problems.

A deeper analysis of days active provides even more fodder for consideration. When a cumulative total for days active is calculated at quarterly intervals (i.e. weeks 3, 6, 9, and 12) and a correlation coefficient is calculated for each subtotal, we find that after only three weeks of the course, the number of days active has a correlation coefficient of .53 with

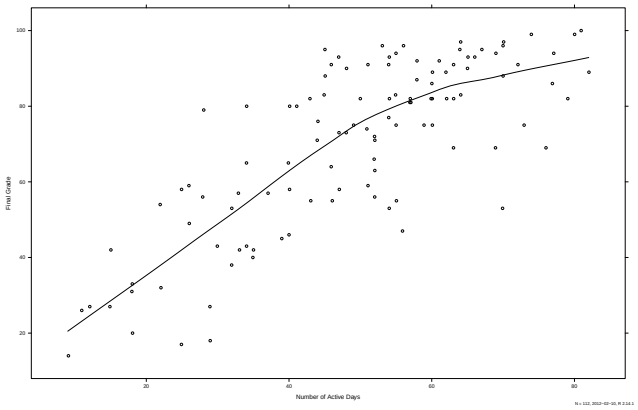

Figure 4: Total Days Active vs Final Grade

participant final grade. After six weeks of class the correlation coefficient is .62 and after nine weeks of class it is .70. Students who are active in the LMS across a number of days appear to be more successful with the course material.

3.3 Predicting Success

To further investigate the role of participant activity in predicting student success, decision trees were constructed using some of the study data. The trees were constructed with Rattle, a data-mining addon for the stats package R. This tree construction used the rpart(recursive partioning) model builder with the default parameters.

Table 3 shows the list of data available to the model builder.

Table 3: Data Available to Decision Tree Model Builder

final grade categories
participant mean confidence for the overall semester
participant mean confidence for weeks 0-4 and 5-10
participant mean expertise for the semester
total LMS visits
total days active
cumulative days active for weeks 0-3, 0-6 and 0-9
mean time taken to complete problem sets

For this particular set of students, there were far more final grades of A and F than of B, C or D. In particular, very few participants received a grade of C. As a result, the grade categories of C and D were clumped together prior to constructing the decision tree and the first decision tree (Figure 5) was constructed for final grade categories of A, B, C&D, F. The discriminating attributes for the decision tree turned out to be the total number of days the participant was active on the LMS for the course, the average time a participant took to complete the weekly problem sets, and the participant's own confidence in his/her programming ability (in one case averaged over the entire semester, in others averaged over weeks 5 through 10). Two of these discriminators (days active and confidence) had high correlations with final grade, but the time taken to solve the weekly problem sets showed no correlation with final grade when examined in our preliminary work.

The decision tree shows that with passively and semi-passively

collected data, and with no direct measurement of participant domain knowledge, the extremes of success (the A learners and the F learners) can be identified with high confidence. Of particular note is the prediction of a final grade of F simply from the number of days active and the participant's own confidence in their abilities at the end of the semester. Also interesting is the predictability of a final grade of A for a confident participant who is active on the LMS for the course for more than 55 days. A high number of active days alone is enough to predict student success.

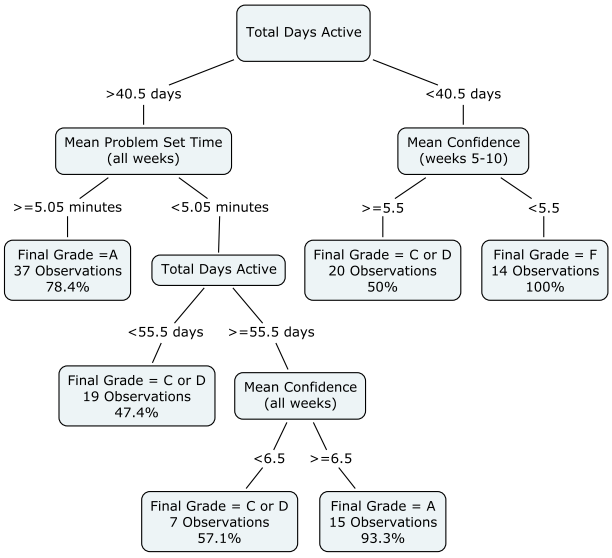

Figure 5: Decision Tree for Final Grade with D and C Clumped

A slightly different decision tree results if the aggregate bins for final grade are changed so that B and C grades are considered together, but the discriminating attributes remain the same. The second decision tree, built with the same algorithm and parameters is shown in Figure 6.

This tree predicts that if a participant used the course LMS for fewer than 40.5 days, the participant is most likely to get a D or an F as a final grade, and that the difference between a prediction of D and F is the participants own confidence in their programming. B or C grades are discriminated from A grades by the average time taken to complete the weekly problem sets (participants who spent more than 5 minutes on average were more likely to get an A).

While this analysis is still preliminary and further work will hopefully increase the confidence in the predictions, the capability to predict the extremes of success using data that can be automatically collected while students are working within a course is quite exciting.

4. MOVING FORWARD

This work has shown that it is possible to predict learner success without requiring data about student mastery of content (i.e. grades).

Two pieces of data stand out as a result of this investigation, the total number of days a learner is active in the course, and the scores on self-check problem sets. While both of these

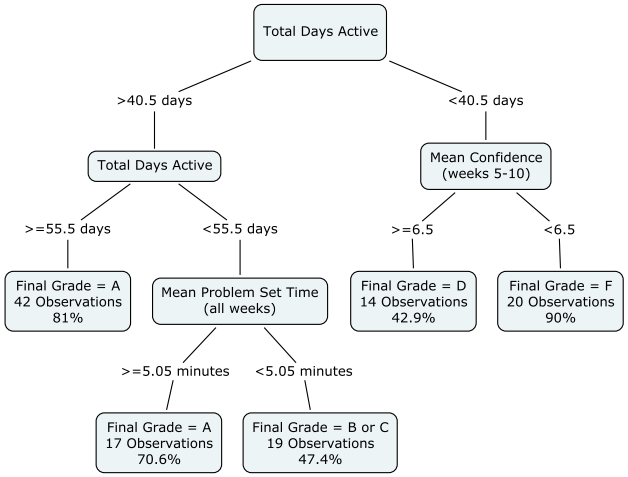

Figure 6: Decision Tree for Final Grade with B and C Clumped

pieces of data can be captured automatically, given suitable self-check questions, the number of days active meets our secondary objective of finding relevant measures that do not require examination of the learner's domain knowledge and, as such, will be the subject of our immediate detailed investigations.

This work is part of a larger effort to enhance existing Learning Management Systems with the ability to react intelligently to learner behaviours. For example, a LMS that could use its own log files to identify learners who appeared to be struggling with a courses would be extremely valuable to instructors. It could notify course instructors about students who might need extra help, and it could offer help to students. Because the underlying models in this work are based on learner behaviour rather than on evaluations of domain knowledge, such an enhanced LMS would require no subject matter knowledge and would be reusable across subject domains. The potential for such a system to improve student retention and success in both distance education and in LMS-supported face to face courses might be quite high. We are optimistic about our future endeavors.

5. REFERENCES

[1] S. Kotsiantis. Use of machine learning techniques for educational proposes: a decision support system for forecasting students' grades. *Artificial Intelligence Review*, pages 1–14, 2011.

[2] G. Weber and P. Brusilovsky. ELM-ART: an adaptive versatile system for web-based instruction. *International Journal of Artificial Intelligence in Education*, 12:351—384, 2001.