A supervised data mining approach for identifying behavior sequences related to academic performance

Christopher J. Urban[1], Matthew L. Bernacki[2], Robert D. Plumley[1], Kathleen M. Gates[1], Cynthia Demetriou[3], A. T. Panter[1], Kelly A. Hogan[1], and Jeffrey A. Greene[1]

[1]University of North Carolina at Chapel Hill [2]University of Nevada, Las Vegas [3]University of Arizona

Project Goals

- Predict which students are likely to perform poorly and which students are not
- Understand behavioral differences between students who perform poorly and students who do not
- These are important first steps toward developing interventions to assist low-performing students

Participants

- 408 undergraduate students in an introductory Biology lecture course
- Used **Sakai**, a learning management system, to access course materials, check grades, and sign up for office hours (Figure 1)
- Completed homework assignments and quizzes using an online learning program called **Mastering Biology**

Data Preprocessing for Sequence Mining

- Treated all actions taken by a student during a single login session as a **sequence** (Figure 2; sequences of actions are delimited by right arrows)
- 3 or more repeated actions in a row during a single login session were condensed into one action
- Homework and quiz actions were timestamped based on how close to the due date students were working
- Dataset split into 4 parts, one per exam period
- Low-performing students were those who scored below a 70 on each period's exam

Methods

1. **Logistic regression**
   - Used to predict which students would perform poorly based on frequencies of various actions
   - Predictors selected using recursive feature elimination with 10-fold cross-validation
   - Model built with 90% of data and tested with 10%
2. **Differential sequence mining** (Kinnebrew, Loretz, & Biswas, 2013)
   - A technique for comparing the behavior patterns of two different groups (Algorithm 1)

Figure 1: Sakai user interface with possible student actions labeled

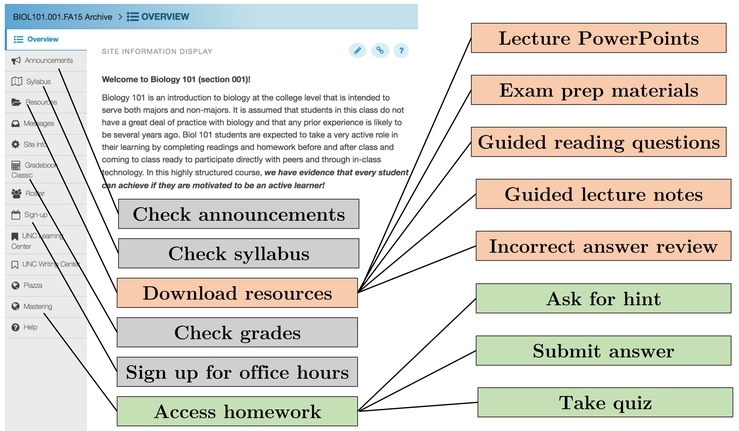

Figure 2: Transforming log data into sequences of student actions

Student ID	Timestamp	Action
student_01	2015-09-03 23:50	SUBMIT_CORRECT_HW_ANS
student_01	2015-09-03 23:52	SUBMIT_CORRECT_HW_ANS
student_01	2015-09-03 23:55	SUBMIT_CORRECT_HW_ANS
student_01	2015-09-03 23:57	SUBMIT_INCORRECT_HW_ANS
student_02	2015-09-04 12:15	LECTURE_PPT_DOWNLOAD
student_02	2015-09-04 12:16	CLASS_OUTLINE_DOWNLOAD
student_03	2015-09-07 15:45	CHECK_GRADES

Student ID	Action Sequence
student_01	SUBMIT_CORRECT_HW_ANS-NITE_DUE_MULTIPLE → SUBMIT_INCORRECT_HW_ANS-NITE_DUE
student_03	CHECK_GRADES
student_02	LECTURE_PPT_DOWNLOAD → CLASS_OUTLINE_DOWNLOAD

Algorithm 1. *Differential sequence mining*

1. Split the dataset into students scoring less than 70 and students scoring greater than or equal to 70 on a particular exam.
2. Find all behavior patterns occurring in at least 5% of either group's login sessions, i.e., find all **frequent patterns**. The percentage of login sessions a pattern occurred in is the pattern's **s-support.**
3. Compute the average number of times each frequent pattern occurs during a login session for both groups. This value is the pattern's **i-support**.
4. Conduct t-tests to compare the i-support for each pattern between groups.
5. Apply the Benjamini-Hochberg false discovery rate correction to all p-values.

Results

Table 1: Differentially frequent behavior patterns for low-performing students versus other students

Behavior Pattern	I-Support Difference (High-Performing - Low-Performing)	S-Frequent Group	FDR Corrected p-value
Exam 1			
None	-	-	-
Exam 2			
LECTURE_PPT_DOWNLOAD	0.02	High-performing	< 0.01
LOGIN → CHECK_GRADES	0.01	Both	0.04
SUBMIT_CORRECT_HW_ANS-AFTERNOON_DUE-MULTIPLE	−0.04	Low-performing	0.04
SUBMIT_INCORRECT_HW_ANS-AFTERNOON_DUE	−0.14	Low-performing	< 0.01
Exam 3			
LOGIN → CHECK_GRADES	0.02	Both	< 0.01
GUIDED_READING_QUESTIONS	0.01	High-performing	0.04
SUBMIT_CORRECT_HW_ANS-NITE_DUE-MULTIPLE	−0.20	Low-performing	< 0.01
SUBMIT_INCORRECT_HW_ANS-NITE_DUE-MULTIPLE	−0.22	Low-performing	< 0.01
Exam 4			
LECTURE_PPT_DOWNLOAD	0.03	Both	< 0.01
CURRENT_SEMESTER_EXAM_DOWNLOAD	0.03	Both	< 0.01
LOGIN → CHECK_GRADES	0.02	Both	0.01

Notes. FDR = false discovery rate.
Behavior patterns are sequences of actions delimited by right arrows.

- 67% to 79% test set accuracy achieved using logistic regression

Conclusions

- Students scoring greater than or equal to 70 on each exam downloaded more course materials and monitored their grades more frequently
- Low-performing students procrastinated on their homework more frequently
- Frequencies of certain online behaviors can be used to predict student performance

References

Kinnebrew, J. S., Loretz, K. M., and Biswas, G. (2013). A contextualized, differential sequence mining method to derive students' learning behavior patterns. *Journal of Educational Data Mining*, 5(1):190–219.

Acknowledgements

A portion of this research was funded by the Finish Line Project (P116F140018; Panter, PI), which is funded by the U.S. Department of Education's "First in the World" grant program.

Contact Information

- Email: cjurban@live.unc.edu