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

Christopher J. Urban1, Matthew L. Bernacki2, Robert D. Plumley1, Kathleen M. Gates1, Cynthia Demetriou3, A. T. Panter1, Kelly A. Hogan1, and Jeffrey A. Greene1

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

- 404 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

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%

Differential sequence mining (Kinnebrew, Loretz, & Biswas, 2013)

- A technique for comparing the behavior patterns of two different groups (Algorithm 1)

Results

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

<table>
<thead>
<tr>
<th>Pattern</th>
<th>High-Performing</th>
<th>Low-Performing</th>
<th>FDR Corrected p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>LECTURE_PPT_DOWNLOAD</td>
<td>67%</td>
<td>65%</td>
<td>0.41</td>
</tr>
<tr>
<td>CHECK_GRADES</td>
<td>69%</td>
<td>67%</td>
<td>0.31</td>
</tr>
<tr>
<td>LOGIN</td>
<td>66%</td>
<td>64%</td>
<td>0.42</td>
</tr>
<tr>
<td>EXAM_LOGIN</td>
<td>70%</td>
<td>68%</td>
<td>0.36</td>
</tr>
</tbody>
</table>

76% 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


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