Using Learning Analytics to Investigate Patterns of Performance and Engagement in Large Classes

Hassan Khosravi
University of Queensland
Brisbane, Australia
h.khosravi@uq.edu.au

Kendra M. L. Cooper
Independent Scholar
Kelowna, Canada
kendra.m.cooper@gmail.com

ABSTRACT

Educators continue to face significant challenges in providing high quality, post-secondary instruction in large classes including: motivating and engaging diverse populations (e.g., academic ability and backgrounds, generational expectations); and providing helpful feedback and guidance. Researchers investigate solutions to these kinds of challenges from alternative perspectives, including learning analytics (LA). Here, LA techniques are applied to explore the data collected for a large, flipped introductory programming class to (1) identify groups of students with similar patterns of performance and engagement; and (2) provide them with more meaningful appraisals that are tailored to help them effectively master the learning objectives. Two studies are reported, which apply clustering to analyze the class population, followed by an analysis of a subpopulation with extreme behaviours.

CCS Concepts

•Social and professional topics → Computing education; Student assessment;

Keywords

Learning Analytics; Personalizing Learning; Clustering; CS1

1. INTRODUCTION

The popularity of computer science in post-secondary education has lead to higher students' enrollment, and larger class sizes in introductory programming courses to help meet the demand. As class sizes grow instructors face heightened challenges in motivating and engaging diverse populations (e.g., academic ability and backgrounds, generational expectations), monitoring students' achievements, and providing helpful feedback and guidance. Often, efforts to identify different subpopulations of the class rely on students' performance on summative assessments. For example, the authors have used midterm grades to identify struggling students by using a cut-off value (e.g., grade on the midterm <40%) to identify and reach out to students in need of more assistance. However, students with similar performance on summative assessments may have dissimilar patterns of engagement with the course content, and therefore may benefit from different recommendations. For example, one group of students may perform poorly on summative assessments while being highly engaged with the course and its content, whereas another group with poor performance may be totally disengaged. The first group may benefit from more preparatory content and training on soft-skills (e.g., time management) or academic skills (e.g., study skills), while the second group may be having personal difficulties and might benefit from meeting with student advisers as a first step in regaining their confidence and/or motivation.

Researchers explore solutions to challenges arising from large classes from alternative perspectives, including learning analytics (LA), which is emerging as a new, interdisciplinary area that explores the "measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environments in which it occurs" [18]. This area is interdisciplinary, drawing upon research, methods, and techniques from education, educational psychology, visualization, and machine learning. LA approaches rely on the availability of substantial data sets (student, course data). This is now feasible, as learning management systems provide capabilities for the automated, unobtrusive collection of access patterns for course materials (when, how much time, downloading, and so on).

Here, LA techniques are applied to investigate data that span summative, formative, and behavioural features. A historical data set for a large, flipped, introductory programming course is used. The goal of the investigation is to reveal smaller groups of students with similar patterns of performance and engagement. The course staff can then provide these groups with tailored appraisals to help effectively master the learning objectives.

Two studies are presented in this paper. The first explores the formative, summative, and behavioural data for the entire class. The second explores a subpopulation in the class: students with extremely high levels of activity using on-line resources. The research methodology uses established best practices in the LA community [8, 11] including the k-means clustering algorithm. The results indicate analyses based on multiple dimensions reveal smaller groups of students with patterns that are not evident using only summative data; this supports providing more meaningful student appraisals.

The remainder of this paper is organized as follows: re-

SIGCSE '17, March 08 - 11, 2017, Seattle, WA, USA

ACM 978-1-4503-4698-6/17/03…$15.00
DOI: http://dx.doi.org/10.1145/3017680.3017711

lated work is presented in Section 2, and an overview of the research methodology is presented in Section 3. A study on clustering the entire class population is presented in Section 4 and a study on clustering a subpopulation is presented in Section 5. Conclusions and future work are in Section 6.

2. RELATED WORK

Research results in two closely related areas, LA in general and LA in CS education, are briefly discussed in this section.

LA has enormous potential to improve post-secondary education. In earlier work, LA studies strive to uncover interesting relationships and patterns among students' summative evaluations and their interactions with the learning environment, with a focus on predicting academic performance. The goal of identifying potentially weak students, based on their log data, can be traced back to [13]. More recently, Barber et al. [2] used a predictive analytic model to identify students in danger of failing a course in which they are currently enrolled; Jayaprakash et al. used binary classification to detect undergraduate students in academic difficulty [10]; and Brooks et al. [6] used time series interaction analysis to predict student achievement in summative evaluations. The analysis of students on many other tasks such as clustering and analyzing learner subpopulations (e.g., [9, 11]), recommending relevant context dimensions for Technology Enhanced Learning (e.g., [21]), visualizing student's progress (e.g., [7]), measuring students' emotions (e.g., [16]), and characterizing aspects of on-line social learning (e.g., [17]) have also received substantial attention.

LA specifically focused on CS education has received attention over the last decade. For example, Blikstein [4] analyzed snapshots of students' code and assessed their behaviour in open-ended programming tasks; Piech et al. [14] used a collection of machine learning algorithms to model how students in a CS course learn to program; Ahadi et al. [1] used classification for mining students' source code in order to identify students in need of assistance in an introductory programming course; and Porter et al. [15] predicted student success using clicker grades from early on in the semester in a CS course. Clustering techniques have been used for categorizing novice programmers [12], investigating the mechanisms of how students begin to learn to program [3], and using automated log analysis to reveal patterns in students' programming [5].

3. RESEARCH METHODOLOGY

An overview of the research methodology is presented in this section: 3.1 introduces the programming course used in the study; 3.2 describes how data have been organized; and 3.3 presents the LA approach used for investigating the patterns of performance and engagement, which includes discussions on the algorithms, techniques, and tools used.

3.1 Example Programming Class

Historical data from a required, introductory course in C programming for engineering students, APSC 160, at The University of British Columbia are used in this research. This course focuses on program design and problem solving, and has over 1000 students and 70 Teaching Assistants (TAs) each year.

This course is fully flipped; students are provided with screencasts (voice over PowerPoint) that introduce the material to be covered in the subsequent class. Lectures start with an in-class clicker quiz that is used to assess student comprehension of the learning goals presented by the screencasts. The remainder of the lecture is allocated to group in-class exercises that provide hands-on experience with this material. Students hand in a copy of their answers, which are later marked based on active in-class participation. A team of TAs attend the lectures to proactively engage with the students and help them with the exercises. A sample solution is released one day after each lecture, allowing students to complete the exercises before checking the solutions.

The lab component is run in the form of an examination. For each lab, students are provided with a pre-lab, which includes two detailed programming problems for self-study with full sample solutions. Students are encouraged to use the first example to refresh their minds of the content and do the second problem under exam conditions to better prepare themselves for the actual lab test. TAs are present in the lab, but their primary task is invigilation.

The course has two midterm examinations. To help students prepare, they are provided with a set of practice questions and solutions for each of the examinations. Students are also provided with a full sample solution for each of their examinations to help them understand their mistakes and better prepare for the final.

3.2 Data Organization

In preparation, the 78 available scores of the class raw data from the first 8 weeks (15 lectures excluding the first lecture) of the course are organized into a vector of nine features spanning summative (S), formative (F), and behavioural (B) dimensions as scaled, normalized values with a mean of 0 and standard deviation of 1. **S features**, labeled S_1, S_2 and S_3, use a total of 7 scores to represent the first 3 values of the vector. These data are organized into: S_1 (labs) the average lab grade of students for the first 5 labs; S_2 (midterm 1) the first midterm grade; and S_3 (midterm 2) the second midterm grade. **F features**, labeled F_1 and F_2, use a total of 30 scores to represent the next two values of the vector. These data are organized into: F_1 (clickers) the average clicker grade over 15 lectures and F_2 (worksheets) the average grade of students for the in-class exercises over 15 lectures. **B features**, labeled $B_1 \ldots B_4$, use a total of 41 scores to represent the last 4 values of the vector. These data are organized into: B_1 (screencast views) the total number of views of screencasts for the 15 lectures; B_2 (worksheet solution views) the total number of solutions (out of 15) students access; B_3 (pre-lab exercise views) the total number of views of the 5 pre-lab exercises; and B_4 (examination/solution views) the total number of files out of the four practice questions with solutions for midterms; and two examination solutions for the midterms students access.

3.3 LA Approach to Investigate Patterns of Performance and Engagement

This section discusses the LA approach used for investigating patterns of performance and engagement on the course data described above; the results are reported in Sections 4 and 5. The approach has three main steps: select the population for the study; identify; and analyze and create appraisals for the clusters.

Selecting the population data to study: The first step is to select the population and consequently the data for

a study. The selection step is included to provide a re-usable approach that can be applied on a collection of studies. The study presented in Section 4 uses the data for the entire class, whereas the study presented in Section 5 uses data for a distinct subpopulation in the class.

Identifying the clusters: Current best practices in LA based on established algorithms, techniques, and tools for clustering are used. An important step in using k-means is determining the number of clusters (K). Most existing methods for determining the number of clusters present it as a model selection problem, in which the clustering algorithm is run with different values of K, and the best value of K maximizing or minimizing a criterion is selected.

The gap statistic method of Tibshirani et al. [20] determines the number of the clusters; it attempts to find clusters that have the properties of internal cohesion and external separation that is challenging to find on student populations that are scattered across the feature space, resulting in over-fitting or under-fitting the data. For the data set used in this work, the method recommended the use of a minimal (2) or a maximal (14) number of clusters, neither of which are satisfying. However, the values provide a useful range of K values to explore (MIN=2, MAX=14), which is aligned with the range considered in previous studies [9, 11]. Values in this range are explored using an alternative approach - the "elbow" method, which can be traced back to [19]. This method aims to obtain the number of clusters by computing and plotting the sum of square errors (SSE) for a range [MIN..MAX] of values of K. The goal is to manually choose a K at which the marginal gain drops significantly, producing an angle (elbow) in the graph. To account for the random initialization of centroids in k-means, recommendations of [9, 11] are followed; for each value in the range, 100 executions of the k-means algorithm are run and the solution with the highest likelihood is selected. The R Studio and Tableau tools are adopted to conduct the studies. The k-means results for the selected value of K are summarized in tables and figures. Final examination grades are used to provide context for the analysis and appraisal results.

Analyzing the clusters and creating appraisals: Once the clusters are selected, they are explored in more depth. The relatively small number of clusters makes this analysis feasible. The normalized, average values of the summative, formative, and behavioural features are abstracted onto a scale - very low (VL), Low (L), medium (M), high (H), and very high (VH) - to help reveal patterns within and among the clusters, considering multiple dimensions. Very low and very high average values exceed one standard deviation from the mean; medium describes values that are close to the class average.

Based on the results of exploring each cluster, appraisals can be created. In general, the appraisals consist of: feedback on student accomplishments with respect to engagement activities (in-class, on-line) and performance outcomes (examination scores) and/or identifying the need for additional information to better understand the students' situations underlying, for example, low in-class engagement. In all appraisals, reminders and invitations to drop by during the scheduled instructor/TA office hours and organized review or problem sessions can be included.

When the students' situations are consistently strong in a cluster (e.g., high engagement and on track to strongly achieving the learning objectives), they can be provided with feedback to recognize their accomplishments and provide additional challenges they may wish to consider (e.g., participating in research projects or volunteering as a peer mentor). When the engagement and performance values in a cluster are mixed, there are many possible reasons behind situations including issues around: (1) technical content; (2) soft-skills (time/stress management); (3) academic skills (study/examination skills); (4) interest/value in the course content; or (5) non-academic issues (family, illness). Based on the authors' experience (more than 20 years teaching experience combined), possible reasons underlying the mixed results are provided as conjectures; recommendations for the students are outlined to help them accomplish the learning objectives for the course.

4. ANALYZING THE ENTIRE CLASS POPULATION

This study analyzes the class population by applying the methodology presented in Section 3. After identifying the clusters, the elbow method is used (refer to Figure 1) for determining the number of clusters; K=5 produces strong results. The results obtained from running k-means with five clusters, identified as C_1, C_2, .., and C_5, are reported in Table 1 and Figure 2. The table contains the normalized, average values for the nine features spanning the S, F, and B dimensions (S_1, S_2,... defined in Section 3.2) as well as cluster statistics on the size of the cluster and the associated median (Q2) on final examination grades. The clusters are ordered with respect to their final examination median grades. Figure 2 visualizes the results presented in Table 1. On the left, a 3-dimensional point plot visualizes the average value of attributes in each dimension; each cluster is labeled with the median examination grade. On the right, the Box and Whiskers plot summarizes the clusters' associated final examination data (median, 25th, 75th percentile, maximum and minimum grade). The overall class median on the final exam is 77. This data set exhibits summative values that are indicative of performance on the final examination.

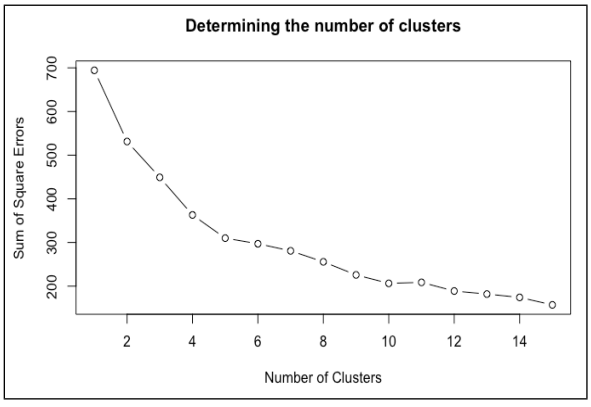

Figure 1: Using the elbow method for determining the number of the clusters for the entire class population.

C_1 consists of 39% of the class population. These students mostly perform extremely well on summative assessments. They frequently attend the lectures and are highly engaged with the in-class activities. In addition, they have high levels of interaction with the on-line course content.

Features		Clusters				
		C_1	C_2	C_3	C_4	C_5
Summative	S_1	0.52	0.29	0.13	-0.15	-1.95
	S_2	0.72	0.51	-0.51	-0.81	-0.67
	S_3	0.68	0.29	-0.06	-0.69	-1.01
Formative	F_1	0.44	0.06	0.32	0.05	-2.21
	F_2	0.43	0.08	0.21	0.13	-2.20
Behavioural	B_1	-0.26	-0.63	1.73	-0.04	-0.03
	B_2	0.33	-2.10	0.49	0.40	-0.28
	B_3	-0.17	-0.58	1.75	-0.08	-0.34
	B_4	0.27	-0.85	0.31	0.18	-0.76
Cluster Stats	Q2	87.8	80.49	70.73	60.98	54.88
	%	39%	13%	12%	26%	10%

Table 1: Using k-means to cluster the entire class population: nine features across S, F, and B dimensions.

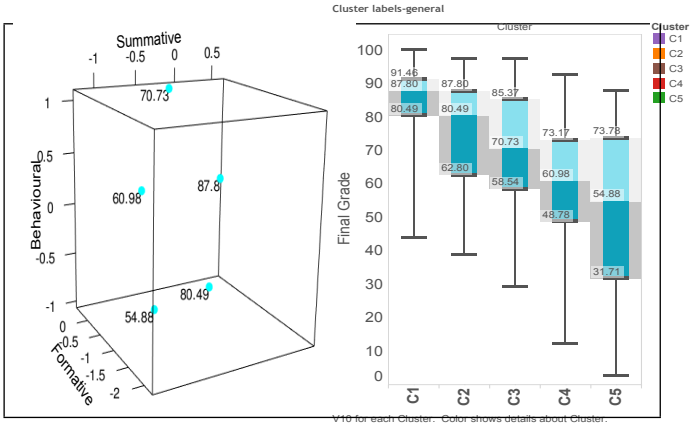

Figure 2: Visualize clusters and associated final exam data for entire class population. Left: 3D point plot of centroids. Right: Box and Whisker plot of final exam grades.

C_2 consists of 13% of the students. These students mostly perform well on summative assessments. Their class participation and engagement with the in-class activities are both slightly higher than the class average. Students in this cluster spend very little or no time interacting with the course content available on-line.

C_3 consists of 12% of the students. These students mostly perform poorly on summative assessments. Their participation and engagement with the in-class activities and engagement in on-line activities are significantly higher than the class average, making them by far the most engaged with the course content available on-line.

C_4 consists of 26% of the students. The students mostly perform very poorly on summative assessments. Their class participation and engagement with the in-class activities are both slightly higher than the class average. Students in this cluster spent less time than average interacting with the on-line course content.

C_5 consists of 10% of the students. These students perform extremely poorly on summative assessments. They rarely attend the lectures and are disengaged during the in-class activities. Students in this cluster spent very little or no time interacting with the on-line course content.

The results of exploring the clusters for patterns and an outline of the appraisals for the students in each cluster are presented in Table 2. The results reveal that some clusters can be quite similar in one dimension and very different across the other two dimensions. For example, students in C_2 and C_3 have somewhat similar performances in the S dimension, with students in C_2 doing slightly better. However, their engagement level both in-class and outside class are significantly different; these are close to the extremes of the spectrum. Similarly, students in C_4 and C_5 both do quite poorly on the S dimension, with the C_4 students doing slightly better; however, C_5 students have very low values across the F and B dimensions, whereas C_4 students have moderate values for those. Finally, students in C_2 and C_5 both have low engagement with the online material; however, students in C_2 have much higher grades both in their summative and formative assessments. The analysis of the clusters using all three dimensions is valuable to provide more meaningful appraisals.

A possible limitation of clustering the entire class population is that it may obscure students with extreme patterns of performance or behaviour, as the results are reduced by averaging. The second study explores the analysis of subpopulations with extreme patterns of engagement.

5. ANALYZING STUDENTS WITH EXTREME PATTERNS OF ENGAGEMENT

The goal of this study is to analyze an extreme subpopulation, providing a better understanding and appraisals for these students. A simple criterion for selecting students with extreme patterns is to consider the top X% of students with the highest or lowest measurements in each of the summative, formative, and behavioural dimensions. Experiments on two of these subpopulations are carried out: (1) *overly engaged participants*, which are those with the highest number of interactions with on-line course material, and (2) *infrequent participants*, which are the students with extremely low performance on formative (in-class) assessments. Due to space limitations, only the study on overly engaged participants is presented.

The methodology in Section 3 is applied. Students with the highest 20% of the average, behavioural values in the class are selected as the subpopulation (96 students). After identifying the clusters, the elbow method is used; $K = 3$ produces strong results.

The results obtained from running k-means with three clusters, identified using E_1, E_2, and E_3, are reported in Table 3 and Figure 3. The table contains the normalized, average values for the nine features as well as cluster statistics on the size of the cluster and the associated median (Q2) on final examination grades. The clusters are ordered with respect to their final examination median grades. Figure 3 visualizes the results presented in the table. This data set also exhibits summative values that are indicative of performance on the final examination.

E_1 consists of 12% of the students in the class. These students mostly perform relatively well on summative assessments. Their class participation and engagement with the in-class activities are higher than the class average. Their engagement with the on-line activities are very strong.

E_2 consists of 4% of the students in the class. These students mostly perform poorly on summative assessments. Their class participation and engagement is lower than the class average. Their engagement with the on-line activities are very strong.

Id	Features	Conjecture	Appraisals
C_1	H, H, M	Strongly engaged and achieving students infers strong interest in the course content with strong technical-, soft-, and academic skills.	Recognize their accomplishments. Provide additional, optional research or peer-mentoring opportunities to sustain engagement.
C_2	H, M, L	Moderate levels of engagement and relatively high achievements infers possibility of previous experience in coding, allowing them to perform well without high engagement. Significant drop in performance on S_3 compared to S_2 potentially because of lack of previous knowledge of content covered later in the term.	Recognize their accomplishments with an alert on drop in performance. Provide additional, advanced challenges early in the semester to improve engagement.
C_3	M, H, H	Strongly engaged and moderate achievements infers lack of soft-, academic skills, and/or issues with the technical content. Significant improvement on S_3 compared to S_2 shows students' desire to do well.	Recognize their hard work and efforts to ("catch-up"). Provide discussions/training on how they can study more effectively. Recommend peer-mentors to them.
C_4	L, M, M	Moderate levels of engagement and low achievements infers lack of soft-skills and issues in the technical contents as they are not benefiting from the in-class, on-line material. Possibly also a lack of interest in deeply learning the topic.	Recognize their effort to engage in-class and on-line, noting their low performance. Provide discussion/training on soft- skills; additional fundamental content and challenges to work on technical content.
C_5	VL, VL, L	Disengaged and low achievements infers lack of soft-, academic skills, non-academic issues, or problems with the technical content. Students may be at risk.	Reach out to early-alert or student consultation services or invite students to meet the course staff in a 1 on 1 session.

Table 2: Feature abstraction, conjectures, and appraisals for each of the clusters in the entire class population.

Features		Clusters		
		E_1	E_2	E_3
Summative	S_1	0.41	0.06	-1.78
	S_2	-0.14	-0.39	-1.27
	S_3	0.31	-0.28	-1.61
Formative	F_1	0.45	-0.01	-0.60
	F_2	0.38	-0.10	-0.68
Behavioural	B_1	0.95	2.37	1.25
	B_2	0.51	0.51	0.51
	B_3	0.93	2.91	0.87
	B_4	0.45	0.40	0.45
Cluster Info	Q2	80.49	63.41	43.9
	%	12%	4%	4%

Table 3: Using k-means to cluster the overly engaged subpopulation: nine features across three dimensions.

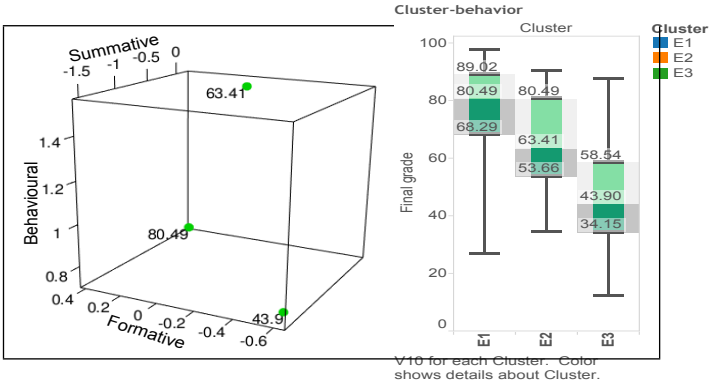

Figure 3: Visualize clusters and associated final exam data for the overly engaged subpopulation. Left: 3D point plot of centroids. Right: Box and Whisker plot of final exam grades.

E_3 consists of 4% of the students in the class. These students perform extremely poorly on summative assessments. Their class participation and engagement is significantly lower than all of the other clusters. Despite heavy interactions with the on-line course content, they exhibit extremely poor performance.

The results of exploring the clusters for patterns and an outline of the appraisals for the students in each cluster are presented in Table 4. The smaller number and size of the clusters makes the exploration simpler in comparison to analyzing the entire class. Here, the results indicate these students, a subpopulation who are extremely engaged with on-line activities, have diverse patterns of performance across their summative and formative assessments. For example, students in E_1 and E_3 have similar, strong engagement with on-line resources. Students in E_3, despite their strong engagement outside the class, are not engaging with in-class activities and are unable to learn the material and fail the course, whereas students in E_1 are very engaged in the classroom and excel in the examinations. The distinctions across the dimensions provide a better understanding of the students' situations, supporting more meaningful appraisals.

6. CONCLUSIONS AND FUTURE WORK

Advancements in technology and the birth of LA have initiated a revolution in education. In this work, LA is utilized to identify groups of students with similar patterns of performance and engagement. The goal is to provide tailored appraisals to each group that helps them effectively master the learning objectives of the course. Results align with previous studies in that summative assessments are the most valuable indicator for predicting final examination performance. Results also reveal that some clusters can be similar on one dimension and very different across other dimensions. As such, to provide more personalized feedback and recommendation, it is important to consider multiple dimensions in the analysis. These findings are true for the entire class population as well as subpopulations with extreme patterns of engagement (overly engaged, disengaged). Due to space limitations, the results exploring the disengaged subpopulation are going to be presented in subsequent work.

The preliminary results presented in this paper are promising, indicating the application of LA for personalizing education has potential; however there are limitations in the work

Id	Features	Conjecture	Appraisals
E_1	H, H, H	Strongly engaged and achieving students infers strong interest in the course content with strong technical-, soft-, and academic skills. Significant improvement on S_3 compared to S_2 shows students' desire to boost their grade and perform better.	Recognize their accomplishments and efforts in improving their performance. Provide mentorship opportunity to sustain engagement.
E_2	L, L, VH	Heavily engaged outside the classroom, but low achieving students infers lack of soft-skills and issues in the technical contents as they are not benefiting from the on-line material. Low in-class engagement may be a sign that students are trying to replace lectures with on-line activities.	Recognize their hard work outside classroom, noting low performance. Provide discussions/training on the the benefits of flipped classrooms, and how they can study more effectively. Recommend peer-mentors to them.
E_3	VL, VL, H	Highly engaged outside the classroom and very low achievements infers lack of soft-, academic skills, and/or issues with the technical content. Extremely low in-class engagement may be a sign that students are shy or uncomfortable working with peers.	Recognize their effort to engage outside classroom, noting extremely low performance. Provide more preparatory material and invite them to attend office hours.

Table 4: Feature abstraction, conjectures, and appraisals for each of the clusters in the overly engaged subpopulation.

to consider. As an example, the on-line engagement of students is approximated with the number of times the material is accessed without any indication of the duration or involvement for the engagement. In the next steps of the research additional kinds of data from the learning management system may be collected to better approximate the on-line engagement. Surveys and interviews can also be conducted to test the validity of the provided conjectures. The additional validation is anticipated to provide further insights that can be embodied into appraisal templates.

An interesting future direction is to use the approach of [8, 11], and consider clustering students based on temporal patterns of performance and engagement. Their work on this topic has been focused on MOOCs. Comparing results of on-campus classes and MOOCs can provide further insight between the differences and similarities of the two.

ACKNOWLEDGMENTS. This research was supported by The University of British Columbia's Carl Wieman Science Education Initiative.

7. REFERENCES

[1] A. Ahadi, R. Lister, H. Haapala, and A. Vihavainen. Exploring machine learning methods to automatically identify students in need of assistance. ICER '15, pages 121–130. ACM, 2015.

[2] R. Barber and M. Sharkey. Course correction: Using analytics to predict course success. LAK '12, pages 259–262. ACM, 2012.

[3] M. Berland, T. Martin, T. Benton, C. Petrick Smith, and D. Davis. Using learning analytics to understand the learning pathways of novice programmers. *Journal of the Learning Sciences*, 22(4):564–599, 2013.

[4] P. Blikstein. Using learning analytics to assess students' behavior in open-ended programming tasks. LAK '11, pages 110–116. ACM, 2011.

[5] P. Blikstein. Multimodal learning analytics. LAK '13, pages 102–106. ACM, 2013.

[6] C. Brooks, C. Thompson, and S. Teasley. A time series interaction analysis method for building predictive models of learners using log data. LAK '15, pages 126–135. ACM, 2015.

[7] E. Duval. Attention please!: learning analytics for visualization and recommendation. LAK '11, pages 9–17. ACM, 2011.

[8] R. Ferguson and D. Clow. Consistent commitment: Patterns of engagement across time in massive open online courses (moocs). *Journal of Learning Analytics*, 2(3):55–80, 2015.

[9] R. Ferguson and D. Clow. Examining engagement: Analysing learner subpopulations in massive open online courses (moocs). LAK '15, pages 51–58. ACM, 2015.

[10] S. M. Jayaprakash, E. W. Moody, E. J. Lauría, J. R. Regan, and J. D. Baron. Early alert of academically at-risk students: An open source analytics initiative. *Journal of Learning Analytics*, 1(1):6–47, 2014.

[11] R. F. Kizilcec, C. Piech, and E. Schneider. Deconstructing disengagement: Analyzing learner subpopulations in massive open online courses. LAK '13, pages 170–179. ACM, 2013.

[12] E. Lahtinen. A categorization of novice programmers: a cluster analysis study. PPIG '16, pages 32–41, 2007.

[13] Y. Ma, B. Liu, C. K. Wong, P. S. Yu, and S. M. Lee. Targeting the right students using data mining. KDD '00, pages 457–464, 2000.

[14] C. Piech, M. Sahami, D. Koller, S. Cooper, and P. Blikstein. Modeling how students learn to program. SIGCSE '12, pages 153–160. ACM, 2012.

[15] L. Porter, D. Zingaro, and R. Lister. Predicting student success using fine grain clicker data. ICER '14, pages 51–58. ACM, 2014.

[16] B. Rienties and B. A. Rivers. Measuring and understanding learner emotions: Evidence and prospects. *LACE*, 2014.

[17] S. B. Shum and R. Ferguson. Social learning analytics. *Educational technology & society*, 15(3):3–26, 2012.

[18] G. Siemens and R. S. J. d. Baker. Learning analytics and educational data mining: Towards communication and collaboration. LAK '12, pages 252–254. ACM, 2012.

[19] R. L. Thorndike. Who belongs in the family? *Psychometrika*, 18(4):267–276, 1953.

[20] R. Tibshirani, G. Walther, and T. Hastie. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B*, 63(2):411–423, 2001.

[21] K. Verbert, N. Manouselis, X. Ochoa, M. Wolpers, H. Drachsler, I. Bosnic, and E. Duval. Context-aware recommender systems for learning: a survey and future challenges. *Learning Technologies, IEEE Transactions on*, 5(4):318–335, 2012.