Data-driven learner profiling based on clustering student behaviors: learning consistency, pace and effort

Shirin Mojarad[1], Alfred Essa[1], Shahin Mojarad[1], Ryan S. Baker[2]
McGraw-Hill Education[1], University of Pennsylvania[2]
{shirin.mojarad, Alfred.essa, s.a.mojarad}@mheducation.com, Rybaker@upenn.com

Abstract. While it is important to individualize instruction, identifying and implementing the right intervention for individual students is too time-consuming for instructors to do manually in large classes. One approach to addressing this challenge is to identify groups of students who would benefit from the same intervention. As such, this work attempts to identify groups of students with similar academic and behavior characteristics who can benefit from the same intervention. In this paper, we study a group of 700 students who have been using ALEKS, a Web-based, adaptive assessment and learning system. We group these students into a set of clusters using six key characteristics, using their data from the first half of the semester, including their prior knowledge, number of assessments, average days and score increase between assessments, and how long after the start of the class the student begins to use ALEKS. We used mean-shift clustering to select a number of clusters, and k-mean clustering to identify distinct student profiles. Using this approach, we identified five distinct profiles within these students. We then analyze whether these profiles differ in terms of students' eventual degree of content mastery. These profiles have the potential to enable institutions and instructors using ALEKS to identify students in need and devise and implement appropriate interventions for groups of students with similar characteristics and needs.

Keywords: Group intervention, ALEKS, clustering, student profiling, grit.

1 Introduction

Interventions delivered by instructors can change students' course of learning, guiding them to improve their outcomes [1]. However, despite the success of specific broad-based interventions [2], it is unlikely that any single intervention will be ideal for all students. A major limitation to the development of broad-based, classroom-wide interventions is that students' characteristics are highly variable from student to student. This variability has made it difficult to identify and apply the right intervention in classroom and online platforms supporting instructors.

This realization has led educators to consider personalized interventions, where each individual receives an intervention tailored to their needs. However, it is time-consuming for instructors to identify and implement the right intervention for individual students, especially if they have to do so manually in large classes. Fortunately, students are not fully unique either; what enhances learning and progress for students with specific characteristics could apply to other students with similar characteristics. Researchers have demonstrated that it is possible to identify groups or clusters within students [3], [4], suggesting that it may be possible to use these methods to identify groups of students who could potentially benefit from the same intervention.

There are several published studies that have clustered students into meaningful groups with a goal of driving intervention. Conati et al. provided initial evidence on a user modeling framework for exploratory learning that can automatically identify meaningful student interaction behaviors and can be used to build user models for the online classification of new student behaviors. They built supervised classifiers to recognize categories of student behavior initially identified using an unsupervised clustering approach [5]. Additionally, Rodrigo et al. used unsupervised clustering on data gathered from an intelligent tutoring system to determine whether it was possible to identify distinct groups of students based on interaction logs alone. With the aim of automatic development of detectors of behavior and affect, they identified two student behaviors clusters associated with differing higher-level behaviors and affective states [6]. Another study categorized learners in 13 MOOC courses based upon their interaction with the course, using k-means clustering, to suggest possible improvements in course design and delivery. They identified learners' classes as Uninterested, Casuals, Performers, Explorers and Achievers, where each class of learners had distinct interaction with the course and followed a certain learning approach. Another study has analyzed engagement patterns on four MOOCs and identified two clusters with seven distinct patterns of engagement, suggesting that patterns of engagement in the MOOCs under study were influenced by decisions about pedagogy [7].

In this paper, we use students' characteristics to identify groups of students who could benefit from the similar intervention. Section 2 describes the data and clustering techniques, section 3 shows some exploratory analysis, and section 4 and 5 include the results and discussion.

2 Data

2.1 ALEKS

ALEKS is a Web-based artificially intelligent learning and assessment system. Its artificial intelligence is based on a theoretical framework called Knowledge Space Theory (KST) [8]. KST allows the representation of a large number of knowledge states that constitute a domain in form of a knowledge map. Therefore, KST allows for a precise description of a student's current knowledge state (KS), and what they are ready to learn next. ALEKS's assessment engine enables estimation of students' KS by a diagnostic test taken when the student starts using the system. ALEKS then conducts assessments during students' progress through the course to continuously update students' KS and decide on what the student is ready to learn next. These progress assessments are taken if the student have achieved certain amount of learning progress or have not used the system within 60 days. Research has shown that using ALEKS after school is as effective as interacting with expert instructors [9].

2.2 Data

The data used in this study is from 18 classes in higher education using ALEKS for Beginning Algebra. The data is comprised of information about each assessment the students have taken and has a 5725 total number of assessments for all students. After removing students who did not take the initial assessment, or took it multiple times (making it difficult to estimate their initial knowledge in the class), the dataset has 628 students. These are 16 week classes and we calculate a set of attributes for students up to week 8 in the class. Below attributes are computed for each student:

- Initial assessment score percentage
- Total number of assessments
- Average days between assessments
- Days since class start initial assessment was taken
- Average percentage score increase between assessments.

In addition to above attributes, we have each students' latest assessment score percentage in ALEKS at the end of class.

The above attributes are named based on how ALEKS works and the student behavioral characteristics in ALEKS. Prior knowledge is students' initial knowledge assessment score. Average days between assessments is treated as a measure of students' consistency of working in ALEKS. As mentioned in section 2.1, since each assessment is triggered based on both how much students have learned and/or how much time they've spent, more frequent assessments indicates consistency in learning and time spent. The total number of assessments taken could indicate students' effort in ALEKS, both in learning and time spent. The average increase in percentage of mastery between each assessment indicates how fast/slow the student is mastering the topics and is referred to as pace. These attributes are calculated at/before week 8 (the middle of the course) by filtering the data only for the assessments taken in the first 8 weeks of the class.

3 Exploratory Data Analysis

3.1 Initial knowledge and outcome

Students' initial score distribution versus final score distribution is a good indication of how much students have progressed over the class period. Figure 1a shows these distributions on the same plot. It is noticeable that students mostly start with low and medium initial knowledge but the outcome could vary from low to very high scores. Students' progress over the class period can be represented as their outcome minus their initial knowledge. Figure 1b shows how varied students' progress is, over the class period in the same domain on ALEKS.

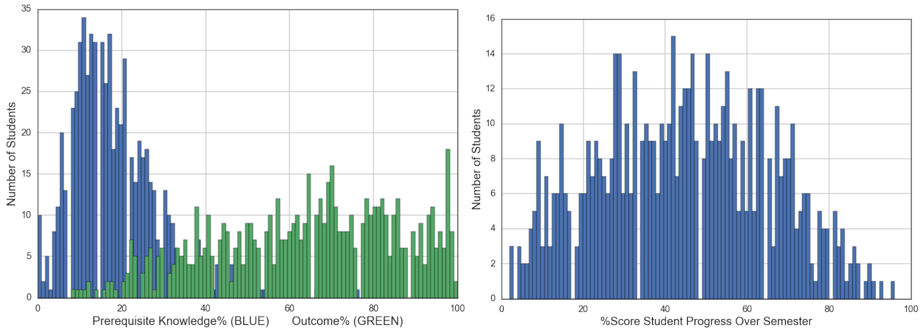

Fig. 1. a) Students' initial knowledge distribution and outcome distribution. b) Students' progress over the class period, computed as the difference between students' initial knowledge and outcome in the class.

3.2 Student progress versus pace

Investigating the relationship between students' average items learnt between assessments (progress) and the average number of days between their assessments (pace) reveals that a typical student learns about 18 items over 7 days. Assessments in ALEKS are triggered based on student's learning and time spent. Progress assessments are triggered when a student has learnt 20 topics or spent 10 hours in ALEKS. This is shown in figure 2, where the highest density is the point at 18 items on the y-axis, representing the average number of items students learnt between assessments, and at 7 days on the x-axis, representing then average days between students' assessments. However, there are students who are considerably faster, learning a given number of items over 5 days or less, and students who are considerably slower, learning the same number of items within 10 days or more.

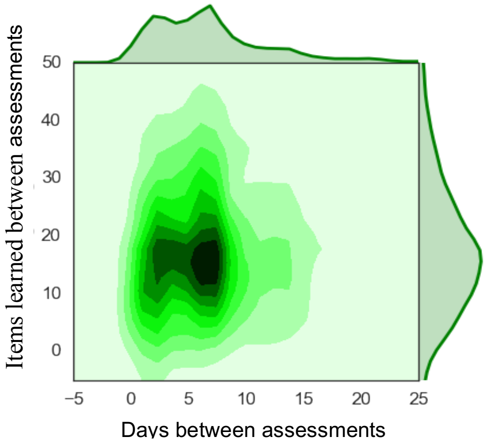

Fig. 2. Relationship between students' average items learned between assessments (progress) and days between assessments (pace).

4 Methodology

We create student profiles within ALEKS using clustering. Specifically, we adopt a three-step process. First, we need to address the high level of correlation between the student characteristics that are input to our clustering algorithm; for example, the Pearson correlation coefficient between last assessment score percentage and total number of assessments for each student is 0.67. To address this concern, we use principal component analysis (PCA), which reduces the data dimensionality by replacing the higher-dimensional original data with a smaller number of non-correlated derived vectors ("principal components"), linear combinations of the original attributes.

Next, we determine the number of clusters using Mean-Shift Clustering [12], which delineates arbitrarily shaped clusters in the data (in this case using a Gaussian kernel) and detects the modes of the density using kernel density estimation. Hence, the number of modes can be used to decide the number of clusters in data by finding the centers of mass within the data.

Finally, we use k-means clustering to find a set of student groups with similar characteristics, choosing this algorithm due to its high interpretability and its property of having non-overlapping clusters. K-means is an iterative clustering technique, where data is partitioned into k clusters by assigning each data point to the cluster with the nearest centroid, and the cluster centroids are iteratively refined until the within-cluster sum of squares (of the distance between each data point and its cluster centroid) cannot be reduced further. In this study, we used Euclidean distance as the distance measure.

5 Results and Discussion

5.1 Preliminary Step: Principal Component Analysis

PCA revealed five principal components (PCs), listed in Table 1. Table 1 shows these factor weightings and the proportion of variance explained by each component. The first three PCs from table 1 cumulatively explain nearly 90% of variance in the data. Therefore, we use the first three PCs as the input to mean shift clustering and k-means. Note that the first three principal components do not load strongly on the variable "Delay in Start". As such, this variable is effectively not included in the cluster analysis and is not used later in section 5.2 to explain cluster characteristics.

Table 1. Attributes' weights and explained variance proportion for principle components.

PC	Prior Knowledge	Consistency	Pace	Effort	Delay in Start	Explained Variance proportion
1	0.06	0.27	0.29	-0.91	0.00	0.46
2	-0.99	0.00	0.13	-0.02	-0.00	0.31
3	0.04	0.81	0.43	0.38	0.02	0.11
4	-0.10	0.51	-0.84	-0.12	0.02	0.10
5	0.005	0.02	-0.00	0.00	-0.99	0.004

5.2 Learner Profiles

Mean-shift clustering identified 5 possible clusters in the data. Hence, we set k-means clustering to look for five distinct groups of students, using the data from the first three PCs. To interpret these clusters, we then look at the average values for each variable in each cluster, and determine whether each cluster has low, medium or high average values for each variable, in relation to other clusters, shown in Table 2.

Table 2. Average attributes for each cluster.

Label	Size	Prior Knowledge (% score)	Consistency (days)	Pace (% score increase)	Effort (# of assessments)
1	190	13.6 (Very Low)	9.3 (Average)	8 (Average)	5.1 (Low)
2	243	17.8 (Average)	8 (Average)	6.9 (Average)	9.5 (Average)
3	50	15.6 (Average)	23.3 (Very Low)	12.9 (High)	4 (Very Low)
4	62	18 (Average)	5.2 (High)	5 (Low)	16.5 (Very High)
5	83	40 (Very High)	10.2 (Average)	6.8 (Average)	6.5 (Low)

We use t-tests to measure the statistical significance of the difference in each characteristic between clusters, using a Benjamini & Hochberg post-hoc correction [10] to control for running multiple comparisons. Table 3 shows the statistical significance (p-value) of the difference between each attribute between each pair of clusters. All differences between groups that have p<0.05, indicated in bold, are also statistically significant after a Benjamini & Hochberg post-hoc correction, except for the prior knowledge difference between clusters 2 and 3 (p=0.04). As Table 3 shows, each pair of clusters is different in at least three of the four variables.

Table 3. Statistical significance (p-value) of the difference between each attribute between each pair of clusters.

Clst 1	Clst 2	Prior Knowledge (% score)	Consistency (days)	Pace (% score increase)	Effort (# of assessments)
1	2	**<.001**	**<.001**	**<.001**	**<.001**
1	3	0.089	**<.001**	**<.001**	**<.001**
1	4	**<.001**	**<.001**	**<.001**	**<.001**
1	5	**<.001**	0.12	**<.001**	**<.001**
2	3	0.04	**<.001**	**<.001**	**<.001**
2	4	0.88	**<.001**	**<.001**	**<.001**
2	5	**<.001**	**<.001**	0.73	**<.001**
3	4	0.20	**<.001**	**<.001**	**<.001**
3	5	**<.001**	**<.001**	**<.001**	**<.001**

5.3 Learner Profile Names

Using the cluster characteristics in table 2, we have identified each learner profile with a name. Below are the cluster names and description of each profiles.

1. Strugglers: this group starts with a very low prior knowledge, puts in low effort and has an average pace of learning.
2. Average Students: this group of learners are average in all characteristics.
3. Sprinters: this group starts with average prior knowledge. They have low consistency in learning and low effort, but have a high pace.
4. Gritty: this group has an average prior knowledge. They have high consistency and high effort, but work at a slow and steady pace.
5. Coasters: this group starts with very high prior knowledge. However, they have average pace and consistency, and put in low effort.

The name of the gritty group is inspired by Angela Duckworth's definition of grit as perseverance and passion to achieve long-term goals [11]. This group shows similar characteristics to what Duckworth defines as grit, maintaining consistency and high effort throughout the class.

A good way of understanding these clusters comes from an external qualitative study conducted by an independent research agency to identify three key student personas in higher education. These personas were characterized based on organizational effort, study effort, persistence, self-confidence, and social extroversion:

- The struggler: characterized by low effort, persistence, self-confidence, social extroversion and very high organizational efforts.
- The planner: characterized by very high organizational and study effort in addition to high persistence, self-confidence and social extroversion.
- The average student: characterized by very low organizational effort, low study effort and persistence, and average self-confidence and social extroversion.

These personas characterize a converging set of categories that represent one or more groups of students we identified in our current cohort of students. The planner characteristics are very similar to gritty students. The struggler persona represents the strugglers and sprinters in our study while the average student persona covers the characteristics of average Students and coasters.

5.4 Learner Profiles and Mastery

We can use final percentage mastery in ALEKS to verify whether the profiles are associated with differences in students' outcomes at the end of the class. Based on their learning characteristics in the first 8 weeks of the class, we expect the strugglers to achieve a low outcome, Average Students to achieve an average outcome, Sprinters to achieve an average to low outcome, the Gritty group to achieve high outcome and Coasters to achieve an average to high outcome.

Table 4 shows the final percentage mastery for each group and the corresponding standard deviation. As we can see, the groups differ in their mastery in the hypothesized fashions. For example, gritty learners finish the class with a very high outcome in ALEKS while strugglers are at the other end. We then conduct a set of t-tests to measure the statistical significance differences in final mastery between groups, using a Benjamini & Hochberg post-hoc correction [10] to control for running multiple comparisons. Table 5 shows the mean differences between different groups and their statistical significances. All differences between groups that have p<0.05 are also statistically significant after a Benjamini & Hochberg post-hoc correction. We find that Groups 1 and 3 (Strugglers and Sprinters) and groups 2 and 5 (Average Students and Coasters) have similar average final mastery. What differentiates these groups is their characteristics in terms of consistency, pace and effort.

Table 4. Final percentage mastery and corresponding standard deviation for each profile.

Label	Learner Profile	% Final Mastery	Standard Deviation	% Final Mastery
1	Strugglers	44.8	17.9	Very Low
2	Average Students	72.4	15.4	Average
3	Sprinters	48.9	19.5	Low
4	Gritty	88.6	11.8	Very High
5	Coasters	72.7	15.1	Average

Table 5. Mean difference between different groups and their statistical significance. All differences between groups that have p<0.05 are also statistically significant after a Benjamini & Hochberg post-hoc correction.

Group	Comparison Group	Mean Difference in Outcome	P-value
1	2	-27.6	<.001
1	3	-4.1	0.15
1	4	-43.8	<.001
1	5	-27.9	<.001
2	3	23.5	<.001
2	4	-16.2	<.001
2	5	-0.3	0.86
3	4	-39.7	<.001
3	5	-23.8	<.001
4	5	15.9	<.001

6 Discussion and Conclusion

In this study, we aimed to address the challenge of deciding and applying individual interventions for students by identifying groups of students whose learning patterns were similar, and who could potentially benefit from the same intervention. We used cluster analysis techniques to identify groups of students with similar academic and behavior characteristics who can benefit from the same intervention. To accomplish this, we used PCA and two clustering techniques to identify distinct groups of students within 628 students using ALEKS. Our analyses used five characteristics of these students up until the midterm of the class: students' starting score in the course (prior knowledge), average score increase between assessments (progress), total number of assessments taken (effort), average days between assessments (pace), and days since class start that initial assessment was taken (delay in start). We find five clusters, described in Table 2 in terms of their average characteristics. Translating these averages to word descriptions gives us the learning pattern of students in each cluster (Table 3), producing five student profiles -- Strugglers, Average Students, Sprinters, Gritty, and Coasters -- which can be communicated to instructors. The clusters can then be used by instructors to devise appropriate interventions in time to still take meaningful action – at the course's midterm. Used appropriately, with the proper interventions, these profiles can ensure that institutions are effectively using the information from adaptive learning systems such as ALEKS to deliver relevant learner intervention. For example, the Strugglers group could benefit from extra instruction while the Sprinters group could benefit from nudges from the learning platform or instructor to maintain learning consistency and put in higher efforts [12].

The next step for this project is to devise and implement interventions using the guidelines given in this study and study the effectiveness of these interventions on individuals within each group, to understand the effectiveness of this grouping for intervention. Another area of future work would be to identify how early each student's learning group could be identified to enable the instructors to intervene early in the course – if a cluster could be identified earlier, it would enable faster and possibly more effective intervention. In addition, learner profiles could be validated further using relevant surveys at the beginning of the semester, before students start using ALEKS. For example, we may be able to use the short grit survey [16], to see if students who self-report as gritty tend to also behave in a gritty fashion within ALEKS.

Through these approaches, we can better understand the learner profiles we are developing and use them to improve student outcomes, helping adaptive learning systems to achieve their goals for helping every student succeed.

7 Acknowledgements

This paper is based on work supported by the McGraw-Hill Education Digital Platform Group. We would like to extend our appreciation for all the informational support provided by ALEKS team at McGraw-Hill Education, specially Jeff Matayoshi, applied research scientist, and Eric Cosyn, director of applied research. Any opinions, findings, conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect positions or policies of the company.

References

[1] X. Lin-siegler, C. S. Dweck, and G. L. Cohen, “Instructional interventions that motivate classroom learning,” *J. Educ. Psychol.*, vol. 108, no. 3, pp. 295–299, 2016.

[2] D. Paunesku, G. M. Walton, C. Romero, E. N. Smith, D. S. Yeager, and C. S. Dweck, “Mind-Set Interventions Are a Scalable Treatment for Academic Underachievement,” *Psychol. Sci.*, vol. 26, no. 6, pp. 784–793, 2015.

[3] F. Bouchet, J. M. Harley, G. J. Trevors, and R. Azevedo, “Clustering and Profiling Students According to their Interactions with an Intelligent Tutoring System Fostering Self-Regulated Learning,” *JEDM - J. Educ. Data Min.*, vol. 5, no. 1, pp. 104–146, 2013.

[4] C. R. Beal, L. Qu, and H. Lee, “Classifying learner engagement through integration of multiple data sources,” *Proc. Natl. Conf. Artif. Intell.*, vol. 21, no. 1, p. 151, 2006.

[5] S. Amershi and C. C. Conati, “Combining Unsupervised and Supervised Classification to Build User Models for Exploratory,” *JEDM-Journal Educ. Data Min.*, vol. 1, no. 1, pp. 1–54, 2009.

[6] M. M. T. Rodrigo, E. A. Anglo, J. O. Sugay, and R. S. J. D. Baker, “Use of Unsupervised Clustering to Characterize Learner Behaviors and Affective States while Using an Intelligent Tutoring System,” in *International Conference on Computers in Education*, 2008.

[7] R. Ferguson and D. Clow, “Examining engagement: analysing learner subpopulations in massive open online courses (MOOCs),” in *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge - LAK ’15*, 2015, pp. 51–58.

[8] J. C. J.-C. Falmagne, N. Thiéry, E. Cosyn, J.-P. Doignon, and N. Thiery, “The Assessment of Knowledge, in Theory and in Practice,” *Form. Concept Anal.*, vol. 3874, no. 949, pp. 61–79, 2006.

[9] S. D. Craig *et al.*, “The impact of a technology-based mathematics after-school program using ALEKS on student’s knowledge and behaviors,” *Comput. Educ.*, vol. 68, no. October, pp. 495–504, 2013.

[10] Y. Benjamini and Y. Hochberg, “Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing,” *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 57. WileyRoyal Statistical Society, pp. 289–300, 1995.

[11] A. L. Duckworth, C. Peterson, M. D. Matthews, and D. R. Kelly, “Grit: Perseverance and passion for long-term goals.,” *J. Pers. Soc. Psychol.*, vol. 92, no. 6, pp. 1087–1101, 2007.

[12] K. E. Arnold, M. D. Pistilli, and K. E. Arnold, “Course Signals at Purdue: Using Learning Analytics to Increase Student Success,” *2nd Int. Conf. Learn. Anal. Knowl.*, no. May, pp. 2–5, 2012.