Using Association Rule Mining to Uncover Rarely Occurring Relationships in Two University Online STEM Courses: A Comparative Analysis

Hannah Valdiviejas
Department of Educational Psychology University of Illinois at Urbana–Champaign, IL, USA
hsv2@illinois.edu

Nigel Bosch
School of Information Sciences and Department of Educational Psychology University of Illinois at Urbana–Champaign, IL, USA
pnb@illinois.edu

ABSTRACT

Metacognition is a valuable tool for learning, particularly in online settings, due to its role in self-regulation. Being metacognitive is especially crucial for students who face exceptional difficulties in academic settings because it grants them the ability to identify gaps in their knowledge and seek help during difficult courses. Here we investigate metacognition for one such group of students: college students traditionally underrepresented in STEM (UR-STEM) in the context of two online university-level STEM courses. Using an automatic detection tool for metacognitive language, we first analyzed text from discussion forums of the two courses; one as a prototype and another as a replication study. We then used association rule mining to uncover fine-grained relationships in the online educational context between underrepresented STEM student status, online behavior, and self-regulated learning. In some cases, we inverted association rules to find associations for underrepresented minoritized students. Implications of the results for teaching and learning STEM content in the online space are discussed. Finally, we discuss the issue of using association rule mining to analyze commonly occurring patterns amongst an uncommon smaller subset of the data (specifically, underrepresented groups of students).

Keywords

Metacognition, Association rule mining, Rare itemsets, STEM

1 INTRODUCTION

The troubling underrepresentation of certain groups of people in STEM majors and careers is a multifaceted and complex issue that does not have one single cause and therefore one single solution. Thus, in this paper we utilize a multi-step research design that involves innovative ways to capture what may or may not be contributing to the underrepresentation of certain students in STEM, specifically through online STEM courses at the university level. In the current study we use student demographic data to understand fine-grained relationships in online learning behaviors, analyzed in ways that are not common in this field of research. Specifically, we inverted what association rule mining was originally constructed to do, which we will discuss in this paper.

1.1 Metacognition and the Online Space

Especially in higher education contexts, where learning responsibilities often fall more on the student than the instructor, it is important to understand the behaviors related to students' academic successes and failures. One behavior that oftentimes separates a successful student from a struggling student is metacognition [11]. Amongst metacognitive research, three main branches of metacognition have been distinguished: metacognitive knowledge, metacognitive monitoring, and metacognitive regulation [7]. For the sake of this research, we focus on metacognitive monitoring, as it is the critical point in order for metacognitive regulation to take place [3].

Metacognitive monitoring, or being conscious of what you do and do not know, is especially critical in online courses, because the burden of guiding and monitoring learning rests more on the student than in traditional learning environments [18]. To be a successful student in an online setting, where being self-regulated is crucial to academic success, the ability to be aware and strategize one's thinking is of the utmost importance. Students who accurately assess their mastery of a concept know how to take effective measures for studying that reflect this judgment of learning. This is called *calibration* and it can be detected through metacognitive monitoring [6]. Traditionally, metacognition in educational contexts has been analyzed according to interventions and surveys; however, this has been shown to be unreliable [15]. More often than not, metacognitive monitoring, a form of self-regulated learning, occurs subconsciously, making it difficult for students to accurately report this [9]. It is for this reason that we use an automatic metacognitive language detection tool [5], in order to avoid invalidities in traditional metacognition measurement.

1.2 Underrepresented Students in STEM

In the United States, an important issue remains unsolved year after year: that is, the vast underrepresentation of African American, Hispanic, Native American, first-generation, and non-male students in STEM majors and careers [2]. As if this were not troublesome enough, with each vertical stage in the academic process, the underrepresentation of these students gets worse [2]. A large underrepresentation of these students, and in turn, a large overrepresentation of people who do not identify with these demographic markers poses a serious bias in the trajectory of the nation, with only a small and homogenous group of people controlling sectors of business and research that are the engines of the nation's economy and innovation [14].

With the concern for students underrepresented (UR) in STEM existing throughout these students' educational trajectory, we argue that much can be learned by examining behaviors related not just to what might impede, but also what might support, these students' success in their STEM college courses to later improve representation in STEM fields. As online education continues to grow [1], its flexibility has made it a very attractive option for underrepresented students in STEM [4]. While online education does offer many options and benefits that traditional face-to-face education does not, it must not only improve access to college courses among traditionally underserved students, but it must also support the academic success of these students. The purpose of this investigation is to document and understand some of the affordances of the online context for UR-STEM students in online STEM college courses.

1.3 Association Rule Mining

Unlike correlation analysis, which is bivariate, association rule mining can discover relationships among multiple variables at the same time [17]. Specifically, association rule mining aims to find "if-then" rules of the variables, in the form of "antecedent → consequence," where antecedent and consequence are conditions that some variable(s) has certain value(s). While association rule mining is extremely useful for exploratory analyses of large data, researchers have only recently attempted to grapple with a main issue of this tool: its inability to catch important, yet uncommon association rules [15]. This shortcoming of association rule mining poses an obstacle.

A handful of prior association rule mining research endeavors have expressed concern and proposed methods to remedy this issue. For example, [13] proposed confabulation-inspired association rule mining for finding rare itemsets. [12] stressed the importance of high-utility infrequent itemsets in fields like biology, banking, retail, and market basket analysis because of how infrequent itemsets find the hidden rules of association among the data items. In their research, they propose a Utility Pattern Rare Itemset (UPRI) algorithm to handle these scenarios. In terms of educational data mining, [16] explains that researchers will likely only find normal behavior in association rule mining because that is the most frequent behavior. To remedy this issue, [16] developed a new algorithm based on the Apriori approach to mine fuzzy specific rare itemsets from quantitative data, consisting of sets of items that rarely occur in the database together.

The current study aims to bring awareness to using association rule mining to catch rules amongst an already known subset of the participants, within the large dataset, rather than first mining in order to *discover* a subset group of the data that has characteristics in common. In this particular case it is minoritized underrepresented STEM students within a normal STEM online course. We applied association rule mining to explore the associations among variables pertaining to these students. For example, a possible rule in this study might be "non-male → no prior online experience." That is, if the student is a non-male, they are likely to have no prior online experience. Given that association rule mining tends to find frequent itemsets, we propose a modified approach in order to answer our research questions.

We ask the following research questions (RQs):

RQ1. What fine-grained relationships amongst underrepresented STEM students, their demographic information, and their metacognitive language can be uncovered through association rule mining?

RQ2. Although created to find commonly occurring sets of rules, can association rule mining to be used to find sets of rules in an uncommon population (underrepresented students in STEM), within a larger set of data?

2 METHOD

In order to answer our research questions, we used demographic information from students in two online STEM courses and discussion forum posts from the same two courses to uncover fine-grained relationships between online learning behavior and student demographic variables.

2.1 Participants and Data Source

2.1.1 Discussion Forum

We analyzed all forum posts (7,040) from 205 students from one (8-week) term of Course A as well as all forum posts (6,086) from 77 students from one (16-week) term of Course B at a large Midwestern public university in the United States. All prompts that corresponded to the forum posts were open-ended with much flexibility for students to answer. Data included all of the students' discussion forum posts as well as their final course grades, which were provided to us by university data curators. Specifically, there were four levels of grades: A, B, C, and D or lower (we combined D and F grades to avoid identifying students from this small group). In both courses, forum participation was required as part of students' participation grades. Students were required to regularly post questions they had, or to answer other students' questions. Online forum activity was 25% of their grade for students in Course A and 5% of their grade for students in Course B.

We used the [5] metacognition tool in order to count metacognitive phrases spontaneously produced by the students in their forum posts. We used this count to relate evidence of self-regulated learning behaviors to students' background information. This tool also categorizes metacognitive language as being positive or negative; however, for the sake of this study, we only used total count.

2.1.2 Participants

Table 1 describes students' demographic characteristics. Note that the total number of students across the subsamples is greater than the total of all students because some students belonged to more than one group. We do not report intersectional group level findings of students who fit multiple UR categories, to protect students' identities and comply with FERPA regulations.

2.2 Data Analysis

Association rule mining has been used in educational contexts to find out relationships between variables, particularly in datasets with many variables [10], like in the current datasets (e.g., ethnicity, prior online experience, ACT score (a standardized test used for college admissions in the United States), grades, metacognitive language count).

Initially we used association rule mining tool as it was intended to be used but only found obvious associations, like those who are STEM majors are likely to have prior subject experience, with none of them dealing with underrepresented students in STEM. This is because their actions were not frequent compared to those in the majority (i.e., STEM majors) and therefore did not get detected as association rules. The current study's process of association rule mining was inverted, meaning that the minimum support and lift values were set low because the target population was vastly underrepresented in the dataset. This included taking the inverse of many dummy variables where the majority was reflected rather than the minority; for example, we changed the variable "STEM major" to "Non-STEM major" so that we were mining for rules associated with the minority rather than the majority and the unlikely versus the obvious likely. In other words, all of the variables were changed to reflect the minority rather than majority in order to avoid excluding uncommon associations in these courses, especially dealing with minority groups. Therefore, we were actually looking for sets of *less* likely associations, relative to the total amount of associations, rather than likely associations. To identify interesting rules, the FP-Growth algorithm was used with a minimum Support value of 0.10, because the minimum population size of some underrepresented student category groups that we looked at (non-males, racial/ethnic minoritized students, and first-generation) were just above 15% of the total population. In other words, if the minimum Support value was set higher than 0.15, it would not capture any of the association rules of the target population and if the minimum Support value were set right at 0.15, it would only capture those association rules in which all of the students pertaining to a specific category exhibited a particular rule. We selected a maximum Lift value of .89 since we wanted to find rules that were not associated with each other. High association, or associations that occur more than expected, are indicated by a Lift value > 1 in traditional uses of association rule mining. Therefore, a Lift value < 1 translates to events that happened less than expected. Through trial and error, we discovered that a Lift value set any lower than 0.89 would be too general and would generate too many rules. A Lift Value set higher than 0.89 gets too close to a high association value, excluding too many rules related to the underrepresented population we were interested in. Rules satisfying the criteria are defined as "interesting" in the sense that they were less likely to happen.

3 RESULTS

3.1 Descriptive Statistics

205 students in Course A produced a total 11,417 metacognitive phrases in 7,007 forum posts. The average number of metacognitive words per student was 55.69 (SD = 24.18). The final exam score was out of 170 points, and scores were approximately normally distributed. The minimum score a student received on the final exam was 69.26 and the maximum was 180 (with extra credit). The 77 students in Course B produced a total of 475 metacognitive phrases and 1,939 forum posts. The mean number of metacognitive phrases per student was 6.17 (SD = 5.07). Table 1 shows a percentage breakdown of the variables used in association rule mining in order to conceptualize Support values. *URM* signifies underrepresented racial/ethnic minoritized students in STEM (African American, Hispanic, and/or Native American), *First-gen.* signifies first generation college student (neither parent completed a higher education degree), *No Prior OL* refers to a student having no prior experience with an online course, a higher poster is a student who posts more than the class average (34 for Course A and 13 for Course B), *Low Exam* refers to the student getting a score lower than the class mean, *Course Rep.* refers to students taking the course for a second time (repeating), and *Non-tr. Age* refers to students older than 22.

Table 1. Student breakdown of variables used

Course A	205 Students	Course B	77 Students
Non-males	25%	Non-males	47%
URM	15%	URM	19%
First-Gen	16%	First-Gen	22%
No Prior OL	25%	No Prior OL	29%
High Poster	45%	Course Rep.	19%
Low Exam	47%	Non-Tr. Age	31%

3.2 RQ Answers

Table 2 shows the association rules that were likely to take place, or associations with a Lift value > 1 and Table 3 shows the association rules with a Lift Value < 1 that were less likely than average to occur. The meaning of each variable follows that of Table 1. The new variables include *Low Total MC* which signifies the student produced less metacognitive language than the average of that class, *High Total MC* phrases refers to students producing more than the average for that class, and prior subject experience refers to students who have had experience with their current course's subject. The strongest associations have Lift values > 1.00 and the weakest association all have Lift values < 1.00.

3.3 Likely Association Rules

The two rules from Course A in Table 2 involve the *likely* associations among variables. In particular, the rule "High poster → Non-male and isolates a strong association regarding who, of the underrepresented students in STEM, is engaging most in beneficial educational behaviors like posting often. "First generation → Low total metacognition" suggests that first-generation students are not engaging metacognition as much as their peers.

The last two rules from Course B in Table 2 involve *likely* associations. The rule "Non-male, Non-traditional age group → Low grade" suggests that non-males who are older than 21 are likely to receive lower grades than their peers. The rule "URM → More than 4 metacognitive comments, Low grade" indicates if a student identifies as a URM, they are likely to engage in high amounts of metacognitive language but receive a low grade.

Table 2. Likely associations (Lift > 1)

	Antecedent	Consequence	Support	Lift
Course A	High poster	Non-male	0.13	1.16
	First-Gen	Low total MC	0.10	1.15
Course B	Non-male, Non-tr. age	Low grade	0.12	1.50
	URM	High total MC, Low grade	0.07	1.66

3.4 Less Than Average Association Rules

The first four rules from Course A in Table 3 involve the *unlikely* associations among variables. These are not simply the inverse of the most likely rules, because the minimum Support value was not changed, only the Lift. The rules "High poster → Low metacognition" and "High poster → Low exam" suggest that students who post often rarely exhibit low amounts of metacognition and rarely get low exam grades. The next two rules, "Low total metacognition, Low Exam → Prior subject experience" and "Low metacognition → Non-male", indicate that the relationships between low metacognitive language and low exam score are rarely found amongst students with prior subject experience and non-male.

Table 3. Unlikely Associations (Lift < .89)

	Antecedent	Consequence	Support	Lift
Course A	High Poster	Low MC	0.09	0.55
	High Poster	Low Exam	0.09	0.70
	Low MC, Low Exam	Prior Subject Experience	0.06	0.73
	Low MC	Non-male	0.09	0.86
Course B	First-Gen, URM	No prior OL	0.06	0.79
	High total MC	Course repeat, Low grade	0.15	0.84
	First-Gen, Non-male	High total MC	0.19	0.87

The next three rules in Table 3 are unlikely associations from Course B. The rule "First generation, URM → no prior online experience" describes that if a student identifies as first-generation and as an URM, they are likely to have prior online experience. The rule "More than 4 metacognitive phrases → Course repeat, Low grade" indicate that it is unlikely for students to have negative educational outcomes if they are engaging in high amounts of metacognitive language. Lastly, the rule "First generation, Non-male → More than 4 metacognitive comments", suggesting if a student is a first-generation and a non-male, it is highly unlikely that they are engaging in a high amount of metacognitive language production.

4 DISCUSSION

Based on the association rule mining analysis that was performed on data from an online Course A, there is evidence that suggests increased posting in this online course is associated with beneficial educational outcomes, like engaging in metacognitive learning strategies and obtaining a high exam grade. A more obvious rule uncovered through this analysis is that prior subject experience is also associated with beneficial educational outcomes. Some insight that rule mining provided about this course is that non-male students, although underrepresented in STEM, generally did well in this course while first-generation students did not fare as well.

Association rule mining also uncovered important information about students in Course B. A stark difference from Course A is that non-male students did not do as well in this course as in Course A. In Course B, being a non-male older than 22 years old was associated with getting a lower grade in the course. Being a non-male in general as well as being a first-generation college student was associated with uttering the least number of metacognitive phrases of all groups compared (gender, first-generation, and URM).

Underrepresented racial/ethnic minoritized students were the most likely group of students, amongst those compared, to produce metacognitive language; however, being a minoritized student was still associated with getting a lower grade in the course. This is an interesting finding because in Course A, the production of metacognitive language was positively related to course outcome however, in Course B it was not. Through association rule mining it is seen that the more metacognitive phrases a student produced, they less likely they were to display non-beneficial educational behaviors (i.e., repeat the course or receive a low grade).

Perhaps the most interesting finding of this analysis is that underrepresented racial/ethnic minoritized and first-generation college students were very likely to have prior online experience, but only for one course. Initially, before mining for association rules, we thought that a possible factor exacerbating the STEM achievement disparity was the digital divide, or the lack of experience that certain populations have with technology [8]. However, there is evidence that this is not the case. Along with research explaining that online education is an attractive option for underrepresented students [19], we see it is likely that underrepresented students have had prior experience with online education. Knowing this, educational researchers could hone in on this advantageous likelihood of experience with online courses to help lessen the underrepresentation of these students in STEM. The fact that this finding was only present in one course and not the other entertains explanations related to how there might be underlying similarities amongst students related to the types of courses they take, even within the STEM discipline.

4.1 Implications

Right now is a crucial time in higher education because of the apparent transition into more of an online state that ever before. We also know that online education is an attractive option for underrepresented students in STEM for various reasons (e.g., flexible class time). That being said, much work needs to be done in understanding academic outcomes in online education, especially for student underrepresented in STEM, because although it has great positive potential it also has the potential to worsen the lack of certain students in STEM majors and field.

The current study also indicates that association rule mining can, in fact, be used in other ways that it was not intended for, and in this case, to find commonly occurring sets of rules in an uncommon population (URMs), within a larger set of data. This opens the possibility for association rule mining to become a prevalent tool to be used among education researchers, especially to generate hypotheses about intersectional relationships that traditional statistical analyses might not uncover.

4.2 Future Directions

Association rule mining is intended to find variables that have strong associations to each other, in order to single out patterns not obvious by simply looking at the data. Using association rule mining was an issue when analyzing uncommon or non-majority populations, and therefore uncommon categories in the dataset, because the data miner has to take the inverse of what association rule mining was constructed to do. It is for this reason that we promote new algorithms or new ways of dealing with specific-rare itemsets, keeping in mind nuanced approaches that might be easier to use for educational researchers who are not entirely familiar with data mining techniques. Also, algorithms for rule mining that are specifically tailored to analyze unlikeliness or even likeliness but in minority subsets of data within the larger dataset would be very useful for reliable results and interpretation as well as facility in usage of educational data mining techniques. Future studies could also include extending these methods to more courses with varied demographics to determine the generalizability of using association rule mining in this way.

5 CONCLUSION

We took a novel approach to uncover relationships between student variables and course success by mining these variables for association rules in order to get a better understanding of the how UR-STEM students interact with online STEM courses.

We mined for unlikely as well as likely associations. We found interesting relationships that could prompt further analysis. These findings could be beneficial to an instructor, to provide clear direction about which students need direct help or additional resources, and thereby enhance positive outcomes in a course. These findings could also prove to be beneficial to online curriculum creators as well as university policy-makers because of specific information regarding an at-risk population (first-generation and racial/ethnic minoritized students) in the leaky STEM pipeline.

6 ACKNOWLEDGMENTS

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education through Grant R305A180211 to the Board of Trustees of the University of Illinois. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

7 REFERENCES

[1] Allen, I.E. and Seaman, J. 2013. Changing course: Ten years of tracking online education in the United States. Sloan Consortium.

[2] Estrada, M., Hernandez, P. R., & Schultz, P. W. (2018). A Longitudinal Study of How Quality Mentorship and Research Experience Integrate Underrepresented Minorities into STEM Careers. CBE - Life Sciences Education, 17(1).

[3] Gourgey, A., (1998). Metacognition in basic skills instruction. Instructional Science, (1/2), 81

[4] Gregory, C. B., & Lampley, J. H. (2016). Community college student success in online versus equivalent face-to-face courses. Journal of Learning in Higher Education, 12(2), 63-72.

[5] Huang, E., Valdiviejas, H., & Bosch, N. (2019). I'm sure! Automatic detection of metacognition in online course discussion forums. In Proceedings of the 8th International Conference on Affective Computing and Intelligent Interaction (ACII 2019) (pp. 241–247). Piscataway, NJ:

[6] Lingel, K., Lenhart, J., & Schneider, W. (2019). Metacognition in mathematics: do different metacognitive monitoring measures make a difference? ZDM - Mathematics Education, 51(4), 587–600.

[7] Miller, T. M., & Geraci, L. (2011). Training Metacognition in the Classroom: The Influence of Incentives and Feedback on Exam Predictions. *Metacognition and Learning*, 6(3), 303–314.

[8] Moore, R., Vitale, D., Stawinoga, N., & ACT Center for Equity in Learning. (2018). The Digital Divide and Educational Equity: A Look at Students with Very Limited Access to Electronic Devices at Home. Insights in Education and Work. In ACT, Inc. ACT, Inc.

[9] Perry, N. E., & Winne, P. H. (2006). Learning from learning kits: Study traces of students' self-regulated engagements with computerized content. Educational Psychology Review, 18, 211-228.

[10] Rojanavasu, P. (2019). Educational Data Analytics using Association Rule Mining and Classification. 2019 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering (ECTI DAMT-NCON), Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering (ECTI DAMT-NCON), 2019 Joint International Conference On, 142–145.

[11] Rovers, S., Clarebout, G., Savelberg, H., de Bruin, A., van Merriënboer, J. (2019). Granularity matters: comparing different ways of measuring self-regulated learning. *Metacognition & Learning,* 14(1), 1–19.

[12] Shrivastava, S., & Johari, P. K. (2016). Analysis on high utility infrequent ItemSets mining over transactional database. 2016 IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT), Recent Trends in Electronics, Information & Communication Technology (RTEICT), IEEE International Conference On, 897–902.

[13] Soltani, A., & Akbarzadeh-T., M. (2015). A new tree-based approach for evaluating rule antecedent constraint in confabulation based association rule mining. International Journal of Knowledge-Based and Intelligent Engineering Systems, 19(1), 1-14.

[14] U.S. Bureau of Labor Statistics. (2019). Bureau of Labor Statistics. Consumer Price Index, 1–2.

[15] Veenman, M. V. J., Van Hout-Wolters, B. H. A. M., & Afflerbach, P. (2006). Metacognition and learning: Conceptual and methodological considerations. Metacognition and Learning, 1, 3-14.

[16] Weng, C.-H. (2011). Mining fuzzy specific rare itemsets for education data. Knowledge-Based Systems, 24(5), 697–708

[17] Xiao Hu, Weng-Lam Cheong, C., & Kai-Wah Chu, S. (2018). Developing a Multidimensional Framework for Analyzing Student Comments in Wikis. Journal of Educational Technology & Society, 21(4), 26–38.

[18] Xu, D., & Jaggars, S. (2011). The effectiveness of distance education across Virginia's community colleges: Evidence from introductory college-level math and English courses. Educational Evaluation and Policy Analysis, 33, 360-377.

[19] Xu, D., Jaggars, S. (2014). Performance gaps between online and face-to-face courses: Differences across types of students and academic subject areas. Journal of Higher Education, 85, 633-659.