Use of Unsupervised Clustering to Characterize Learner Behaviors and Affective States while Using an Intelligent Tutoring System

Ma. Mercedes T. Rodrigo[a], Elizabeth A. Anglo[a], Jessica O. Sugay[a], Ryan S.J. d. Baker[b]

[a]*Department of Information Systems and Computer Science, Ateneo de Manila University, Loyola Heights, Quezon City, Philippines*
[b]*Human-Computer Interaction Institute, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA*

Keywords: Affect, Aplusix, learner modeling, interaction logs, clustering.

Abstract: This paper presents results from a preliminary analysis of interaction and human observation data gathered from students using an Aplusix, an intelligent tutoring system for algebra. Towards the development of automatic detectors of behavior and affect, this study tried to determine whether it was possible to identify distinct groups of students based on interaction logs alone. Using unsupervised clustering, we were able to identify that student behaviors within the software cluster into two categories, Clusters 0 and 1, associated with differing higher-level behaviors and affective states. Cluster 0 tended to reflect more collaborative work, whereas Cluster 1 reflected more solitary work. Cluster 1 students tended to exhibit more flow, suggesting that students in flow tend to work in a more individual fashion. An examination of the keystrokes used by each group showed that Cluster 0 used the arrow keys and cursor keys significantly more than Cluster 1. The Cluster 1, on the other hand, tended to type more mathematical operators or use the duplicate command more frequently than Cluster 0. This implies that frequent use of mathematical operators and frequent duplication of the problem may be evidence of flow within Aplusix.

Introduction

An intelligent tutoring system (ITS) is a computer program that makes use of artificial intelligence to help students learn a target instructional domain [5]. It maintains an internal student model to represent student capabilities or deficiencies and uses this model to guide its subsequent actions. In the past, student models tended to be based on a student's performance in cognitive assessments, i.e. correct or wrong answers in drills or tests within [6] or outside of the system. Recent literature extends the student model to include non-cognitive information.

Past studies have been successful at uncovering the effects of students' personality differences such as challenge-seeking and persistence [16], and learners' feedback preferences [9] on student interactions with an automated system and, consequently, on learning. Analysis of these interaction logs, for example, has led to the development of detectors for student behaviors such as "gaming the system," defined as systematic guessing or hint abuse to progress though a learning system [4], help-seeking[1], off-task behavior [3], confidence and interest [15], and student affective states such as confusion [8] and frustration [12]. Researchers have also drawn conclusions about system usability, usage patterns or user knowledge, expertise or proficiency in a subject area [2, 10].

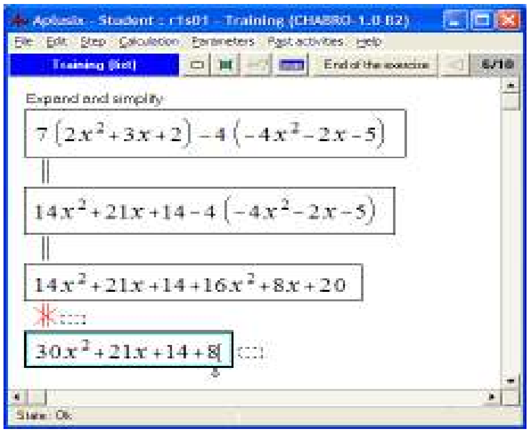
However, researchers seldom know a priori which of these characteristics will exhibit themselves in ways that can be automatically detected. When interaction logs are summarized into frequency counts, different learner strategies may look the same [11]. Hence, the analysis of interaction data requires some exploration of how student behaviors occur in shorter time-scales. Different approaches often have to be applied in order to obtain meaningful results. This paper presents a preliminary analysis of student keyboard interaction data and human observations. Using unsupervised clustering, we attempt to answer the following questions:

- What student groupings can be formed based on students' interaction logs?
- What are the behaviors and affective states exhibited by each of these groupings?
- Are the keystrokes used by one group distinct from those of the other group?

We hope that this analysis may lead to a model that can in the future be used to design an autonomous, intelligent agent capable of detecting the differences in students' behavior and affect.

Methodology

Aplusix II: Algebra Learning Assistant [13] is an intelligent tutoring system for mathematics. Topics are grouped into six categories (numerical calculation, expansion and simplification, factorization, solving equations, solving inequalities, and solving systems), with four to nine levels of difficulty each. Aplusix presents the student with an arithmetic or algebraic problem from the selected set. Students then solve the problem one step at a time. At each step, Aplusix displays equivalence feedback: two black parallel bars mean that the current step is equivalent to the previous step, whereas two red parallel bars with an X mean that the current step is not equivalent to the previous step (see Fig. 1). Aplusix does not indicate which part of the current step requires further editing; the student must discover this for himself/herself. When the student believes he is done, he can end the exercise. Aplusix then tells the student whether errors still exist along the solution path or whether the solution is not in its simplest form yet. The student has the option of looking at the solution, a "bottom out" hint with the final answer.


Aplusix - Student : r1s01 - Training (CHABRO-1.0.B2)
File Edit Step Calculation Parameters Past activities Help
Training (list)
End of the exercise
6/10
State: Ok

Expand and simplify:

$7(2x^2+3x+2)-4(-4x^2-2x-5)$

$14x^2+21x+14-4(-4x^2-2x-5)$

$14x^2+21x+14+16x^2+8x+20$

$30x^2+21x+14+8$

Fig. 1. A screenshot from Aplusix: Algebra Learning Assistant

The learners who participated in this study were first and second year high school students from four urban schools from within Manila and one provincial school located in Cavite, a province south of Manila. Students' age ranged from 12 to 15 with an average age of 13.5. One hundred and forty students participated in the study (83 female, 57 male). Students used Aplusix in groups of ten, one student per computer. Students first completed an information sheet that asked about their names, ages, and grade point averages. They took a 10-minute pre-test. After the pre-test, each student used Aplusix for 45 minutes, beginning at the first level of difficulty of expansion and simplification. Afterwards, the students took a 10-minute post-test.

A. Classroom Observations

The observations were conducted using the emotion (or affect in the literature) and behavior observation method developed in [17]. The observations were carried out by a team of six observers, working in pairs. The observers were Masters students in Education or Computer Science, and all but one had prior teaching experience. The observers trained for the task through a series of pre-observation discussions on the meaning of the affective categories. Observations were conducted according to a guide that gave examples of actions, utterances, facial expressions, or body language that would imply an affective state, and practiced the coding categories during an unrelated observation prior to this study.

As in [17], each observation lasted twenty seconds, and was conducted using peripheral vision, i.e. observers stood diagonally behind or in front of the student being observed and avoided looking at the student directly, in order to make it less clear exactly when an observation was occurring. If two distinct affective states were seen during an observation, only the first behavior observed was coded, and any behavior by a student other than the student currently being observed was not coded. Each pair of observers was assigned to three students and alternated between them. Since each observation lasted twenty seconds, each student was observed once per 180 seconds.

The usage categories coded were identical to those in [17], and are as follows: on-task solitary; on-task, giving and receiving answers; other on-task conversation; off-task conversation; off-task solitary behavior; inactivity; gaming the system.

The affective categories coded were also those used in [17] and were boredom, confusion, delight, surprise, frustration, flow [7], and the neutral state.

Some of these affective categories may not be mutually exclusive (such as frustration and confusion), though others clearly are (delight and frustration). For tractability, however, the observers only coded one affective state per observation. Thirteen pairs of observations were collected per student. Inter-rater reliability was acceptably high: Cohen's κ=0.83 for behavior and κ=0.63 for affect.

B. Dataset Creation

The Aplusix system automatically logs student actions. From these logs, the following information can be obtained: the problem attempted, its level of difficulty, the student's keyboard action (a character keystroke, a number keystroke, or a request for the score), the action number, the action's time. Of particular interest in this paper are the keyboard actions and their timing with respect to the students' behavioral and affective states as coded by the observers.

Since the observation times were fixed while the action times were variable, it was not immediately clear which actions took place during which observation. To address this issue, actions were synchronized with the observations using an unsupervised action filter based on a variable time window, as discussed in [18]. Using this technique, an observation is associated with all actions that occur within a dilation of time around that observation. Given, for example, a

2-minute time window, all keyboard actions logged between 1 minute before and 1 minute after each observation were matched with the observation. Because we were uncertain of which time window size would be most appropriate, we grouped the data into 2-, 2.5-, and 3- minute time windows. Then, for each time window, we counted the number of occurrences of a keyboard action. Finally, we labeled each time window according to its corresponding behavioral and affective observation. Each dataset contained approximately 3,000 records. Each record had 12 features (frequency counts), one behavior observation value and one affective state value. The 12 features are described below:

- Seven separate frequency counts for the use of character key, mathematical operator, logical operator, decimal key, parenthesis, fraction key, arrow key;
- Frequency of use of the cursor action (e.g., selection, de-selection, cut, copy, paste, drop);
- Frequency of use of the step features (undo, redo, duplicate);
- Frequency of use of the verification features (ask for the solution, ask for the score, ask for a verification of the solution);
- Frequency of use of the termination or ending features (say the exercise is solved, abandon, normal termination);
- Frequency of use of the commenting features (inserted a comment between steps, inserted a comment on the step).

The next task was to use data mining algorithms to answer the research question.

Clustering Using K-Means

Clustering is a class of machine learning techniques aimed at automatic detection of patterns in unlabeled data. Clustering operates on data points (feature vectors) in a feature space. Features can be any measurable property of the data, e.g., number of times a particular keystroke has been used in an observation.

The logged data consisted of 2,734 feature vectors corresponding to the 2,734 observations. The feature space consisted of the twelve separate attributes that represented the frequency of use of each keystroke as discussed previously. Typically the *k* value is determined by intuition from the researchers' knowledge of the data set. For this first analysis, the researchers selected k=2 and a time window of 2.5-minutes.

The k-means clustering algorithm included in the software Weka [19] was used to analyze the data. K-means chooses *k* random points to serve as cluster centers. All feature vectors are then assigned to their closest cluster center, using the Euclidean distance metric. Once all vectors have been assigned, the algorithm computes for centroid or mean of all each cluster. This centroid is then taken to be the new center value. The cluster assignment is then repeated. The whole process repeats until the cluster centers do not change.

Results and Discussion

A. Behavior and Affective States in the Cluster

K-means divided the data into two clusters, 0 and 1, with 1,662 and 1,072 records respectively. After the data was clustered, the behavior and affective state labels identified in the earlier synchronization step were re-attached to each record. Next the records in each cluster were counted and the distributions were tabulated first, according to behavior and then according to affective state. The distribution of the behaviors and affective states in the clusters are shown in Table 2 and Table 2, respectively. Logistic regression (in SPSS) was used to evaluate the significance of the differences of the distributions between clusters, with one regression equation for each behavioral/affective category, and a student set of terms included in order to account for non-independence between two observations of the same student. The Wald variant of the Chi-squared test (e.g. [14]) was used to evaluate the statistical significance of the cluster term in the logistic regression.

Table 1. Behavior Clusters (*k*=2). Statistically significant relationships are shaded in dark gray. Marginally significant relationships are shaded in light gray.

Cluster	**No of Observations**	**Working**	**Giving and Receiving Answers**	**Other On-Task Convers.**	**Off-Task Convers.**	**Off-Task Solitary**	**Inactive**	**Gaming**
0	1,662	79%	8%	12%	0.5%	0.2%	0.1%	1.2%
1	1,072	88%	5%	5%	0.1%	0%	0.6%	0.9%

Table 2. Affect Clusters (*k*=2). Statistically significant relationships are shaded in dark gray.

Cluster	**No of Observations**	**Boredom**	**Confusion**	**Delight**	**Flow**	**Frustration**	**Neutral**	**Surprise**
0	1,662	3%	16%	6%	70%	3%	1.3%	0.4%
1	1,072	1.4%	12%	5%	79%	2%	0.4%	0.3%

In terms of behavior, Cluster 0 tends to reflect more collaborative work, whereas Cluster 1 reflects more solitary work. Cluster 1 involves significantly more solitary work behavior, Wald χ^2(1, N=2,734)=15.11, p<0.001, whereas Cluster 0 involves more on-task conversation, Wald χ^2(1, N=2,734)=9.18, p<0.01. In addition to being more collaborative in constructive fashions, Cluster 0 involved marginally significantly more sharing of answers, Wald χ^2(1, N=2,734)=2.93, p=0.09.

In terms of affect, Cluster 1 demonstrated significantly more flow, Wald χ^2(1, N=2,734)=6.34, p=0.01. While there was the appearance of more confusion and boredom in Cluster 0, these relationships were not statistically significant, respectively, Wald χ^2(1, N=2,734)=0.01, p=0.95, Wald χ^2(1, N=2,734)=1.50, p=0.6.

These results appear to suggest that students who are more in a state of flow (in Cluster 1), on the other hand, tend to work in a more individual fashion. Since these students are not having difficulty (at least not more than they can handle), there is presumably not a need to seek help or attempt to regulate their negative affect through interpersonal communication. The flow state has been previously found to be associated with positive cognitive conflict and leads to intrinsic motivation to persist in a task and to learn. The absence of flow, on the other hand, appears to be associated with the decision to attempt to learn in more collaborative fashions, although it is not clear whether lower flow leads to greater collaboration, or vice-versa.

B. Keystrokes in the Clusters

As mentioned earlier, this study hopes to contribute to the larger goal of building autonomous, intelligent agents capable of detecting learning and non-learning behaviors. To achieve this goal, the measures of, say, productive and unproductive behaviors must be computationally tractable. In this case, the tractable data is composed of keystrokes. It is therefore of interest for us to examine whether the incidences of certain keystrokes were significantly different between clusters.

A total of 158,918 keystrokes were recorded for this data set. Cluster 0 accounted for 71,770 keystrokes, while Cluster 1 accounted for 87,148. The records in Cluster 0 averaged 26.25 keystrokes per observation. Cluster 1 records averaged 31.88. This implies that students who are in flow and working with the software, on average, type more than those who are interacting more with other students and experiencing confusion or boredom – suggesting that this may be a useful measure for detecting both student behavior and affect.

Figure 2 shows the distribution of the five most-often used keystrokes. Linear regression (in SPSS) was used to evaluate the significance of the differences of the distributions between clusters, with one regression equation for each behavioral/affective category, and a set of student terms included in order to account for non-independence between two observations of the same student. Arrow keys were used significantly more in Cluster 0, F(1, 2583)=17, p<0.001. Cursor commands, similarly, were used marginally significantly more in Cluster 0, F(1, 2583)= 38, p<0.001. Mathematical operations, by contrast, were used significantly more in Cluster 1, F(1, 2583)= 84, p<0.001. The duplication command was also used significantly more in Cluster 1, F(1,2583)= 137, p<0.001.

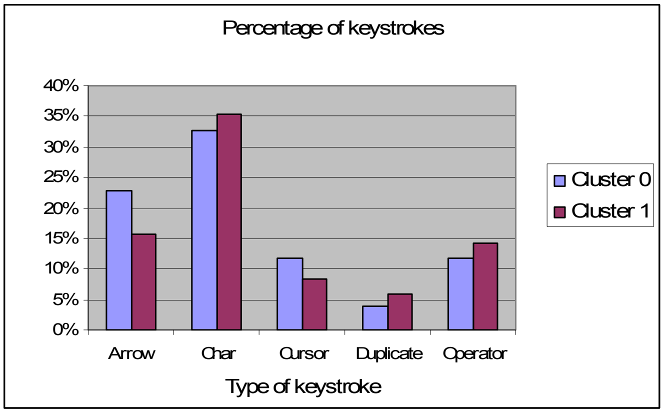


Fig. 2. Distribution of the Keystrokes for *k* =2

Based on these patterns, it is reasonable to conjecture that the frequent use of mathematical operators and the frequent duplication of the problem appear to indicate flow and engaged work. The use of arrow keys or cursor operations may, by contrast indicate that a student is collaborating with other students (or will soon collaborate with other students, or has just collaborated with other students).

Summary, Conclusions and Future Work

This paper was an exploratory study that tried to determine whether it was possibly to meaningfully cluster student interaction data at the keystroke level. Using the k-means clustering technique in Weka, the data was clustered into two, Clusters 0 and 1. On average, Cluster 0 records tended to use fewer keystrokes than records in Cluster 1. Cluster 0 records used significantly more arrow keys and cursor commands, whereas Cluster 1 tended to use more mathematical operators and more duplication.

An examination of the behavior and affect labels per record, gathered through human observation, showed that Cluster 0 was more associated with interaction between the student and other students (whether on-task and off-task), whereas Cluster 1 was more associated with solitary work and the affective state of flow.

The findings from this study are relevant to the design of intelligent agents [cf. 4]. Based on the results presented here, an affect-sensitive agent for Aplusix could monitor student keystrokes and intervene if the student continuously requests assistance from teachers or peers, to encourage the student to attempt to solve problems on his or her own before seeking help [cf. 1]. Several prior projects [cf. 4, 9, 12, 15, 16] provide examples of the possible responses to student behaviors or affective states. However, the exact interventions that would be most appropriate for Aplusix are beyond the scope of this paper.

In the next phases of the analysis, researchers will experiment with the value of k as well as the time window. The researchers will examine the clusters in terms of students' pre- and post-test scores to determine whether the clusters vary in terms of achievement. The researchers plan to include additional dimensions, i.e., number of correct or incorrect answers within a time span and the time lag between actions. Future work may also look at the duration of the switch from one keystroke to another. The pause duration may give insight on the behavior of the student: shorter pauses may indicate the learner is overusing a feature without thinking and hence may lead to poorer learning. A closer look at the sequence of the use of the arrow keys may also be helpful. Not all arrow key usages are necessarily impediments to learning. The ITS should be sensitive to the usage of a specific keystroke that is excessive and those which are not.

Acknowledgement

We thank Jean-Francois Nicaud of the Laboratoire d'Informatique de Grenoble for the use of Aplusix. We thank Maria Carminda Lagud, Sheryl Ann Lim, Alexis Macapanpan, Sheila Pascua, Jerry Santillano, Jessica Sugay, Sinath Tep, and Norma Jean Viehland, and Dr. Ma. Celeste T. Gonzalez for their assistance in organizing and conducting the studies reported here. We also thank the Ateneo de Manila High School, Kostka School of Quezon City, School of the Holy Spirit of Quezon City, St. Alphonsus Liguori Integrated School and St. Paul's College Pasig for their participation in the studies conducted. The work presented in this paper was made possible in part by a grant from the Ateneo de Manila University and by the Pittsburgh Science of Learning Center which is funded by the National Science Foundation, award number SBE-0354420.

References

[1] Aleven, V., McLaren, B., Roll, I., & Koedinger, K. (2004). Toward tutoring help seeing: Applying cognitive modeling to meta-cognitive skills. *7th International Conference on Intelligent Tutoring Systems*, 227-239, Amsterdam, The Netherlands: Springer-Verlag.

[2] Amershi, S. & Conati, C. (2006). Automatic recognition of learner groups in exploratory learning environments. In Mitsuru Ikeda, Kevin D. Ashley, Tak-Wai Chan (Eds.), *Intelligent Tutoring Systems, 8th International Conference, ITS 2006*. Jongli, Taiwan, June 2006, 463-472, Germany: Springer-Verlag.

[3] Baker. R. S. J. d. (2007). Modeling and understanding students' off-task behavior in intelligent tutoring systems. *ACM Computer Human Interaction 2007 Conference*, 1059-1068.

[4] Baker, R.S., Corbett, A.T., Koedinger, K.R. (2004). Detecting Student Misuse of Intelligent Tutoring Systems. *7th Conference on Intelligent Tutoring Systems*, 531-540, Amsterdam, The Netherlands: Springer-Verlag.

[5] Conati, C., Gertner A. & VanLehn K. (2002). Using Bayesian networks to manage uncertainty in student modeling. *Journal of User Modeling and User-Adapted Interaction, 12*(4), 371-417.

[6] Corbett, A.T., & Anderson, J.R. (1995) Knowledge Tracing: Modeling the Acquisition of Procedural Knowledge. *User Modeling and User-Adapted Interaction*, 4, 253-278.

[7] Csikszentmihalyi, M. (1990). *Flow: The Psychology of Optimal Experience.* New York: Harper and Row.

[8] D'Mello, S.D., Taylor, R. S., & Graesser, A.C. (2007). Monitoring affective trajectories during complex learning. *29th Meeting of the Cognitive Science Society*, 203-208.

[9] Forbes-Riley, K. & Litman, D. (2007). Investigating human tutor responses to student uncertainty. In A. Paiva, R. Prada, & R. W. Picard (Eds.), *ACII 2007, LNCS 4738*, 678-689.

[10] Guzdial, M., Santos, P., Badre, A., Hudson, S., & Gray, M. (1994). Analyzing and visualizing log files: A computational science of usability. Presented at HCI Consortium Workshop.

[11] Hulshof, C. D. (2004). Log file analysis. Encyclopedia of Social Measurement, Vol. 2, 577-583. Amsterdam: Elsevier Academic Press.

[12] McQuiggan, S. W., Lee, S., & Lester, J. C. (2007). Early prediction of student frustration. In A. Paiva, R. Prada, & R. W. Picard (Eds.), *ACII 2007, LNCS 4738*, 698-709.

[13] Nicaud, J-F., Bouhineau, D., & Chaachoua, H. (2004). Mixing microworld and CAS features in building computer systems that help students learn algebra. International Journal of Computers for Mathematical Learning *9*, 169-211.

[14] Pampel, F.C. (2000) *Logistic Regression.* Thousand Oaks, CA: Sage Publications.

[15] Porayska-Pomsta, K., Mavrikis, M., & Pain, H. (2007). Diagnosing and acting on student affect: The tutor's perspective. *User Model and User-Adapted Interaction, 18*(1-2), 125-173.

[16] Rebolledo-Mendez, G., du Boulay, B., & Luckin, R. (2006). Motivating the learners: Am Empirical Evaluation. In M. Ikeda, K. D. Ashley, & T.-W. Chan (Eds.), *ITS 2006,* 545-554, Heidelberg, Germany: Springer Berlin.

[17] Rodrigo, M. M. T., Baker, R. S. J. d., Lagud, M. C. V., Lim, S. A. L., Macapanpan, A. F., Pascua, S. A. M. S., Santillano, J. Q., Sevilla, L. R. S., Sugay, J. O., Tep, S., & Viehland, N. J. B. (2007). Affect and usage choices in simulation problem-solving environments. In R. Luckin, K. R. Koedinger, & J. Greer (Eds.) *Artificial Intelligence in Education: Building Technology Rich Learning Contexts that Work*, 145-154, Amsterdam, The Netherlands: IOS Press.

[18] Walonoski, J. & Heffernan, N., (2006) , Detection and Analysis of Off-Task Gaming Behavior in Intelligent Tutoring Systems, *ITS 2006*, 382-391, Heidelberg, Germany: Springer Berlin.

[19] Witten, I & Frank, E., (2005) *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd Edition, Morgan Kaufmann, San Francisco.