Investigating Students' Interaction Profile in an Online Learning Environment with Clustering

Gökhan Akçapınar
Computer Education and Instructional Technology
Hacettepe University
Ankara, Turkey
e-mail: gokhana@hacettepe.edu.tr

Arif Altun
Computer Education and Instructional Technology
Hacettepe University
Ankara, Turkey
e-mail: altunar@hacettepe.edu.tr

Erdal Coşgun
Biostatistics
Hacettepe University
Ankara, Turkey
e-mail: erdal.cosgun@hacettepe.edu.tr

***Abstract*—The aim of this study is to identify clusters of students who interact with an online learning environment in similar ways. The study included analyzing three-month interaction data from 74 undergraduates in the online learning environment using the Self Organizing Map (SOM) clustering method. The results of analysis revealed the existence of three distinct groups of students, labeled by their interaction (non-active, active, very active) and course success (low learning, medium learning, high learning). These are the preliminary results of the study and the cluster data which was obtained here is intended to be used in further studies for classifying new students or adaptation and personalization purposes.**

Keywords-component; student behavior modeling, student profiling, interaction data, clustering

I. INTRODUCTION

Profiling students and modeling student-related attributes (skills, knowledge, performance, behavior etc.) are topics that researchers explore for effective adaptive learning environments [1, 2]. Modeling student behaviors can be used for identifying similar students in online learning environments [3], teachers can use this information in order to classify new students depending on their characteristics [4] or to group students for working together in collaborative activities [5].

Modeling learning-related attributes of students is essential in adaptive learning environments for the adaptation and feedback mechanisms as well [6]. There are different educational objectives for using these models in adaptive systems, such as: for the purpose of identifying and eliminating gaming and off-task behaviors of students in Intelligent Tutoring Systems (ITS) [7, 8], to build student models in exploratory learning environment for the online classification of new student behaviors [9], etc.

Data mining techniques such as, classification and clustering are frequently used for student profiling based on students' activities in online learning environments [10]. Automatic classification is an indispensable part of adaptive learning environments. Prior to making any adaptation or intervention such as, task selection, navigation, content etc. the system has to classify the current state of a student [11]. Although clustering is normally an unsupervised process for grouping similar elements into clusters, the groups which were identified by the clustering analysis can be used for classifying new student and enable real-time adaptations [12]. Therefore, the aim of this study is to identify clusters of students who exhibit similar interaction behaviors in online learning environments.

II. METHODS

The data that was used in the study was obtained from an online learning environment which was designed by the researchers. The participants of the study were 74 undergraduates who were taking the Computer Hardware course. In addition to the in-class course, the students performed activities in the online learning environment through the term which could help their learning. These activities can be summed as, writing reflections about the concepts which they learned in the course, reading, commenting and assessing (like/dislike) reflections written by other students, asking questions in discussions, writing answers to questions, assessing questions and answers which were written by others, following announcements and course resources. In addition to data about these activities, students' navigation paths and login information is stored in the database within the system.

A. Dataset

The raw data which was used in the research includes three-month interaction data from students (n = 74) in the online learning environment. This data was analyzed by the researchers and the attributes which reflect aforementioned students' behaviors were identified and analyzed by the means of Self Organizing Map (SOM) clustering analysis. These attributes are shown in Table 1.

978-1-4799-4038-7/14 $31.00 © 2014 IEEE
DOI 10.1109/ICALT.2014.40

TABLE I. ATTRIBUTES USED FOR EACH STUDENT.

No.	Attributes	Description
1	n_Login	login count
2	d_Usage	total time spent on the environment
3	n_Post	post count
4	n_UniquePost	number of unique days on which a post was written
5	n_Tag	number of tags used in posts
6	n_PostNav	number of navigation on posts written by other students
7	n_PostAss	number of assessing posts written by other students
8	n_DissNav	number of navigation on the discussion section
9	n_Answer	number of writing answers to questions in the discussion section
10	n_QuestionAss	number of assessing questions in the discussion section

B. *Data Analysis*

The aim of the clustering analysis is to reduce distances within the cluster and increase distances between clusters. Therefore, elements in one cluster will be more similar to each other than with those from other clusters [13].

In clustering analysis, in how many clusters the data is to be divided can be predefined [5] or the optimal number of clusters in the data can be found by using cluster validity indexes [7, 10]. In this study, number of clusters set as 3 by the authors and clustering analysis was performed by the Self Organizing Map (SOM) method as per the process shown in Figure 1.

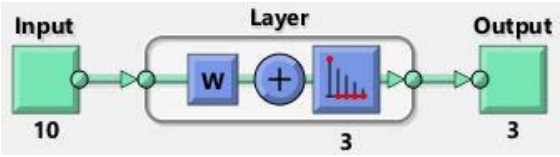

Figure 1. SOM analysis process.

SOM initiates the analysis with a pattern which includes the Neural Network structure and is similar to a grid structure which is in a certain size in the beginning (For example: 1x3, 2x3). This size, also determines the number of clusters (in this study, 3 was selected as the number of size). In the first stage of the clustering analysis process, individuals are randomly appointed to the matrix which was auto-generated. Then, the distance between the next individual and the firstly appointed ones is calculated as the Euclidean distance and the next individual is appointed to the closest cell in the matrix. This process continues until all individuals are appointed to a cluster.

One-Way ANOVA test is used for analyzing whether there is any difference in terms of course success between the student groups which were determined by the clustering analysis. MATLAB was used for the clustering analysis and SPSS software was used for the One-Way ANOVA test and descriptive statistics. Since the distances between the data points are important for the clustering analyses, the attributes were standardized by turning them into Z scores in the preprocessing stage.

III. RESULTS

The graphics which were obtained as a result of the SOM analysis made by the MATLAB software are shown in Figure 2. The graphic on the left (Figure 2.a) shows the inter-cluster distances. The dark colors in the graphic indicate clusters which are away from each other and light colors indicate closer ones. This graphic can be interpreted as, Cluster 1 is the most distinct one, Cluster 2, and Cluster 3 however, are closer to each other. The graphic on the right (Figure 2.b) shows the number of students in each cluster.

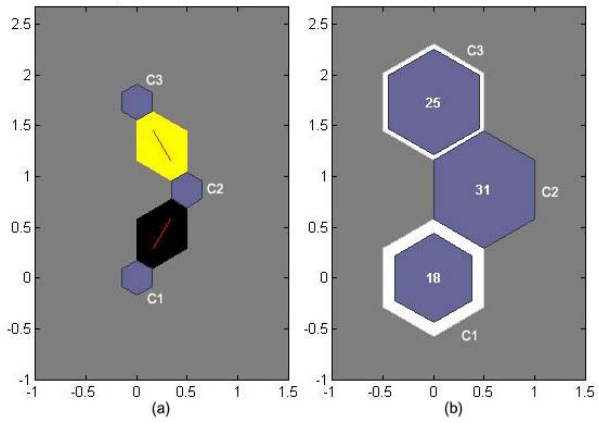

Figure 2. (a) Inter-cluster distances (b) Number of students in each cluster

When cluster means in Table 2 are examined, it is seen that students in Cluster 1 (n = 18) made more logins into the environment, spent more time, wrote more posts, wrote posts on more different days, used more tags in their posts, visit more posts written by other students, visited the discussion section more often, wrote more answers to the questions, made more comments about the posts and assessing more questions and posts compared to students in other clusters. Students in Cluster 3 (n = 25) however, are seen to have made much fewer logins into the environment, wrote much fewer posts, used fewer tags in their posts, visited the discussion section much less, wrote almost no answers to the questions, made fewer comments about the posts written by others and assessing less questions and posts. Finally, Cluster 2 (n = 31) is characterized by students with values somewhat smaller than Cluster 1 but greater than Cluster 3.

One-Way ANOVA test, which was performed to analyze whether there is any difference between the students in different clusters in terms of course success, introduced that there was a statistically significant difference between clusters in terms of course success ($F = 34.048$, $p = .001$). Cluster means are as follows: Cluster 1 ($n = 18$, $\bar{x} = 71.67$, $sd = 10.05$), Cluster 2 ($n = 31$, $\bar{x} = 59.55$, $sd = 12.41$) and Cluster 3 ($n = 25$, $\bar{x} = 40.12$, $sd = 14.79$). In terms of course success, the resulting clusters can be "labeled" as High

Learner (Cluster 1), Medium Learner (Cluster 2) and Low Learner (Cluster 3) according to Kardan and Conati [14]'s classification. Or, based on cluster means they can be labeled as Very Active Students (Cluster 1), Active Students (Cluster 2) and Non-Active Students (Cluster 3) as Romero, Ventura, and García [5] did by using only the interaction data.

TABLE II. CLUSTER MEANS

Attributes	Clusters		
	1 (n = 18)	*2 (n = 31)*	*3 (n = 25)*
	$\bar{x}$	$\bar{x}$	$\bar{x}$
n_Login	93.4	54.1	17.2
d_Usage (min)	3538	1655	703
n_Post	99.3	71.2	4.8
n_UniquePost	9.1	6.3	.9
n_Tag	102.4	63.1	5.5
n_PostNav	271.3	112.4	21.5
n_PostAss	144.6	56.4	14.3
n_DissNav	358.2	121.9	34.4
n_Answer	8.8	3.5	1.2
n_QuestionAss	52.8	35.5	14.7

IV. CONCLUSION

In the study, three-month interaction data from the students in the online learning environment was analyzed by the clustering method and three different groups were identified and analyzed which consisted students who exhibited similar interaction patterns. Determining the ideal number of clusters or comparing the performances of different clustering algorithms are outside the scope of this study. The number of clusters was predefined as 3 by the researchers and SOM was used as the clustering method. When the cluster means are examined, it is seen that the formed clusters reflect 3 different student groups. The presence of a statistically significant difference between the clusters in terms of course success supports this finding. It was also found that, students with lower activities in the environment had a lower course success; students with moderate activities in the environment had a moderate course success; and students with higher activities in the environment had a higher course success.

The results which were obtained here can be used in further studies for classifying new students and providing feedback about their possible achievements during their course attendance. Besides, like Amershi and Conati [9] did, these cluster variables can be used as inputs for the classification models which work integrated into adaptive environments.

REFERENCES

[1] Peña-Ayala, A., *Educational data mining: A survey and a data mining-based analysis of recent works.* Expert Systems with Applications, 2014. **41**(4, Part 1): p. 1432-1462.

[2] Romero, C. and S. Ventura, *Data mining in education.* Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2013. **3**(1): p. 12-27.

[3] Valdiviezo, P., R. Reátegui, and M. Sarango. *Student Behavior Patterns in a Virtual Learning Environment*. in *Eleventh LACCEI Latin American and Caribbean Conference for Engineering and Technology (LACCEI'2013)*. 2013. August 14 - 16, 2013 Cancun, Mexico.

[4] Lopez, M.I., et al. *Classification via clustering for predicting final marks based on student participation in forums*. in *5th International Conference on Educational Data Mining, EDM 2012*. 2012. Chania, Greece.

[5] Romero, C., S. Ventura, and E. García, *Data mining in course management systems: Moodle case study and tutorial.* Computers & Education, 2008. **51**(1): p. 368-384.

[6] Bouchet, F., et al. *Identifying Students' Characteristic Learning Behaviors in an Intelligent Tutoring System Fostering Self-Regulated Learning*. in *5th international conference on educational data mining (EDM)*. 2012.

[7] Baker, R.J.d., et al., *Generalizing Detection of Gaming the System Across a Tutoring Curriculum*, in *Intelligent Tutoring Systems*, M. Ikeda, K. Ashley, and T.-W. Chan, Editors. 2006, Springer Berlin Heidelberg. p. 402-411.

[8] Baker, R.S., et al., *Off-task behavior in the cognitive tutor classroom: when students "game the system"*, in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2004, ACM: Vienna, Austria. p. 383-390.

[9] Amershi, S. and C. Conati, *Combining Unsupervised and Supervised Machine Learning to Build User Models for Exploratory Learning Environments.* Journal of Educational Data Mining, 2009. **1**(1): p. 71-81.

[10] Bienkowski, M., M. Feng, and B. Means, *Enhancing Teaching and Learning Through Educational Data Mining and Learning Analytics: An Issue Brief*. 2012: Washington, D.C.

[11] Hämäläinen, W. and M. Vinni, *Classifiers for Educational Data Mining*, in *Handbook of Educational Data Mining*. 2010, CRC Press. p. 57-74.

[12] Bouchet, F., et al., *Clustering and profiling students according to their interactions with an intelligent tutoring system fostering self-regulated learning.* Journal of Educational Data Mining, 2013. **5**(2).

[13] Chien-Sing, L. and Y.P. Singh. *Student modeling using principal component analysis of SOM clusters*. in *Advanced Learning Technologies, 2004. Proceedings. IEEE International Conference on*. 2004.

[14] Kardan, S. and C. Conati. *A Framework for Capturing Distinguishing User Interaction Behaviours in Novel Interfaces*. in *The 4th International Conference on Educational Data Mining (EDM 2011)*. 2011.