

Studentlyzer for Analyzing and Visualizing E-learning Data

Zongsheng Zhao, Yating Lei, Yi Dou, Yik Him Ho, Henry C.B. Chan and Chetwyn C.H. Chan

Abstract — With the increasing popularity of e-learning in higher education institutions, there is a need to develop data analytics tools to analyze e-learning data, student learning behavior and student performance. In recent years, there has been growing interest in educational data mining, which can provide useful insights into student learning behavior, providing holistic analysis. This paper presents an online data analytics tool called Studentlyzer, which applies data mining to analyze student data. It can cluster student datasets using K-means clustering, and visualize the graphical results through a web browser. Two real-world student e-learning datasets, the Open University Learning Analytics Dataset (OULAD) and Educational Processing Mining (EPM) dataset, were used to demonstrate Studentlyzer's usefulness. The results provide valuable insights about students. In general, Studentlyzer can help identify students who are similar (e.g., with similar study behavior) and provide useful information about student performance and student behavior (e.g., their correlation).

Keywords — educational data mining, e-learning, clustering, online learning behavior

I. INTRODUCTION

Today, with the advent of information and communications technologies (ICT), e-learning has become increasingly popular for use in higher education. Unlike conventional learning, e-learning allows students to learn more flexibly and at their own pace [1]. Previous studies have shown that e-learning can greatly facilitate active learning [2-6]. An e-learning system can provide a large amount of student data, such as enrollment and attendance records, login times, number of posts in discussion forums, as well as assessment results (e.g., Grade Point Average (GPA)). There has been considerable interest in developing effective methods to analyze and mine e-learning data in higher education.

A. What is EDM?

Educational data mining (EDM) seeks to use data mining methods for educational purposes [7]. The primary purposes are to find patterns, discover learner behavior and predict performance based on educational data (e.g., student data) using data analytics and computing methods [8, 9]. EDM is interdisciplinary, involving a collaboration between researchers from different fields, such as computer science, education, psychology and mathematics. In general, there are four major types of EDM methods: (a) Prediction Models - including classification, regression, latent knowledge estimation; (b) Structure Discovery - including clustering, factor analysis; (c) Relationship Mining - including association rule mining, causal data mining, correlation mining and sequential pattern mining; and (d) Model

Discovery - including prediction, clustering, or in some cases knowledge engineering [10].

B. Related Work

EDM is a promising research area in e-learning. According to a survey by Peña-Ayala [8], there were 240 related works published between 2010 and 2013 (first quarter). These works can be classified based on the following: algorithms [11], disciplines [12], educational systems [13], methods [14] and tasks [15]. With the advent of e-learning, coupled with a large amount of educational data, there is also a growing trend toward specialized EDM research (i.e., to focus on a specialized educational domain).

Using a decision tree model, Quadri and Kalyankar [16] predicted student dropout rates based on certain factors, such as gender, grade, attendance rate, family income, parental education level, and whether or not a student was working. They found that among all factors, family income is the most important factor influencing student dropout rates. The study also indicated that student performance can be predicted based on certain factors, such as gender, family income and whether he/she is a firstborn child. With the aim of investigating the accuracy of an ID3 decision tree model for performance prediction purposes, Baradwaj and Pal [17] studied the performance or results of 50 students at VBS Purvanchal University, Jaunpur (Uttar Pradesh) enrolled in a Master of Computer Applications course between 2007 and 2010. The study confirmed the effectiveness of using a decision tree model to predict student performance. On the other hand, the limitations of EDM have also been studied. For example, Nasiri and Minaei [18] used data mining to study the use of educational attributes to predict student dropout rates and results. The study revealed the ineffectiveness of certain data mining methods due to their strong dependence on data distribution. Among the various data mining algorithms, clustering is one of the most common. For instance, based on the study in [19], clustering provides a useful method to study student learning style, with considerations placed on such factors as time spent on learning tasks, learner behavior in class, and student learning motivation. A previous study has also applied K-means clustering to analyze student learning behaviors based on quiz, assignment and exam results (i.e., to classify students into three clusters: high, medium and low performance) [20]. The study provided good evidence of the effectiveness of using K-means clustering for student classification. Compared to their work, our focus is to support online analysis and customized processing.

This paper presents an effective online system to analyze, cluster and visualize e-learning data in order to understand students better. The proposed system gears toward addressing differences in student learning experiences, and hence student learning outcomes. In particular, the system has the capacity to cluster students in terms of their learning outcomes (or scores) with respect to potential factors that reflect hindrances in learning in certain students. These factors can include reduced participation rates in learning activities and difficulties in completing exercises. The results can help improve course design and enrich the learning experience. The

Manuscript received January 8, 2019; revised February 11, 2019. This work was supported by the Innovation and Technology Fund of the Hong Kong Special Administrative Region Government.

Zongsheng Zhao, Yi Dou, Yik Him Ho, and Henry C.B. Chan are with the Department of Computing, The Hong Kong Polytechnic University (e-mail: csyhho@comp.polyu.edu.hk).

Yating Lei and Chetwyn C.H. Chan are with the Department of Rehabilitation Sciences, The Hong Kong Polytechnic University.

results can also help identify students who may require special attention (e.g., with very different behavior), so that advice/counseling can be provided.

The remaining sections are outlined as follows. Section II presents the Studentlyzer system. Section III discusses the test results based on two open datasets. Section IV concludes the paper.

II. STUDENTLYZER SYSTEM

Studentlyzer (<https://studentlyzer.comp.polyu.edu.hk>) is a web-based platform for analyzing student data. It allows users to upload CSV-formatted student data so that data can be clustered using Agglomerative, Gaussian Mixture or K-means clustering algorithms. Agglomerative clustering is a hierarchical clustering method, seeking to group data recursively into clusters based on a “bottom-up” approach. Gaussian Mixture clusters data based on the assumption of Gaussian distribution. K-means clusters data into a predefined number of clusters based on the nearest centroids. In this paper, as an example, K-means clustering was used for the data analysis. After clustering, the website will show the clustering results and data with graphs and tables. Users can perform further analysis based on the figures and graphs on the website.

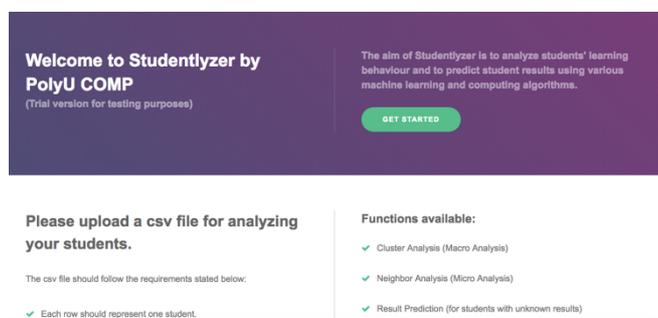


Fig. 1. Studentlyzer website

After uploading a CSV file, the user can choose the clustering algorithm, the number of clusters and column of student ID, as well as student academic performance indicator (student score). ID and score columns will not be considered using the clustering algorithm. The platform is also capable of selecting an arbitrary subset of columns other than ID and score as student behavior data, which are used for clustering (see Fig. 2). After clicking the “Cluster” button, the server will perform clustering for the data and return the results through the website. The clustering results are shown through graphs. The website also provides a sample CSV file with randomly generated student data. The following shows the basic functions of Studentlyzer using the sample data.

A. User Behavior Data and Correlation Coefficients with Score

Consider that we make the following settings after uploading the sample CSV data file: setting the number of clusters to “4”, the student academic performance indicator as “score”, and column ID as “student ID”. Furthermore, the following behavior data were chosen for clustering: `read_time`, `mouse_scroll_frequency`, `switch_window_frequency`.

After clicking the “Cluster” button to initiate the clustering process, a Pearson product-moment correlation coefficient will be provided for each behavior attribute. This seeks to evaluate the correlation/relationship between student score and each selected behavior data. In this case, `read_time` was positively correlated with the score with the highest correlation coefficient ($r = 0.085$) among the three parameters.

Clustering

sample.csv

Please select one of the following methods to generate clusters:

Algorithm:

Agglomerative Clustering

Gaussian Mixture

K-means

Weighted/Non-weighted:

Weighted

Non-weighted

Number of clusters:

Choose ID column:

id

Choose score column:

score

Choose columns to be used for clustering:

`read_time`

`mouse_scroll_frequency`

`login_count`

`switch_window_frequency`

Fig. 2. Interface after CSV file is selected

It also indicates a positive correlation between `switch_window_frequency` and score ($r = 0.063$) and a negative correlation between `mouse_scroll_frequency` and score ($r = 0.046$). Based on these correlation coefficients, a user can evaluate the effect of each parameter on students’ academic performance. Additionally, the user can also click a radio button to select any one of the behavior parameters as the x-axis of behavior – score graph, as shown in Fig. 3.

B. Behavior – Score Graph

Based on the selected parameter, the website will generate an interactive scatter plot graph, as shown in Fig. 3. The x-axis is the selected student behavior (i.e., based on the radio buttons above), and the y-axis is the student academic performance indicator. Therefore, it is called a behavior – score graph. When pointing to a certain point with the mouse, a tooltip box will be shown, which shows the student ID, coordinate, and Euclidean distance from the cluster centroid. The user can also deselect and hide certain points from the clusters (e.g., see Fig. 4). Users can use a behavior – score graph to see how selected student behavior data are distributed and their relationship with academic performance. Users can also see the relationship between the selected student behavior and the clusters.

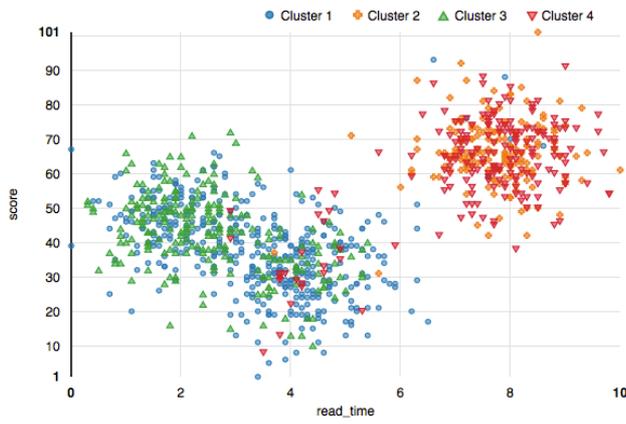


Fig. 3. Behavior – score graph

C. Distance – Score Graph

The next graph under behavior – score graph is the distance – score graph, which shows the Euclidean distance of each student to the cluster centroid and the corresponding academic performance indicator (see Fig. 5). This graph also indicates how representative a cluster is. If a cluster is more representative, the points will be closer to the centroid (i.e., shorter distance). Again, users can hover a mouse on a point to see that point’s information and only display certain clusters.

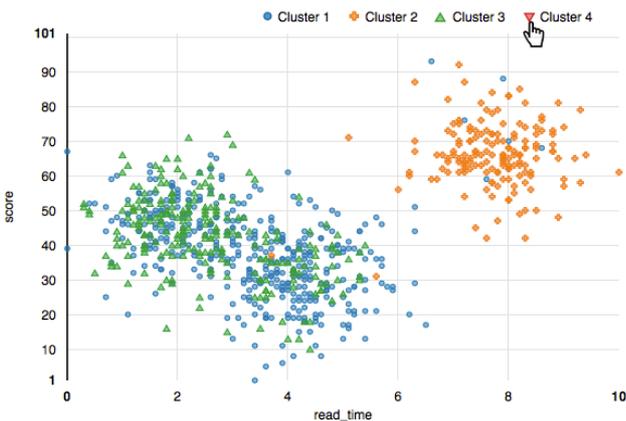


Fig. 4. Click legend to hide all points from clusters

D. Cluster Information Table

In addition, a cluster information table is shown after the graphs (see Table 1). This table displays the statistical data of each cluster, including the mean value of student academic performance, standard deviation of student academic performance, minimum distance to the centroid, and maximum distance to the centroid.

III. RESULTS AND DISCUSSION

In this section, we evaluate the practicality of our Studentlyzer system by analyzing two real world student e-learning datasets: Open University Learning Analytics Dataset (OULAD) [22] and Educational Processing Mining (EPM) dataset [23]. Based on the analyzed results generated from Studentlyzer, the aim is to explore the relationship between student behavior and academic performance.

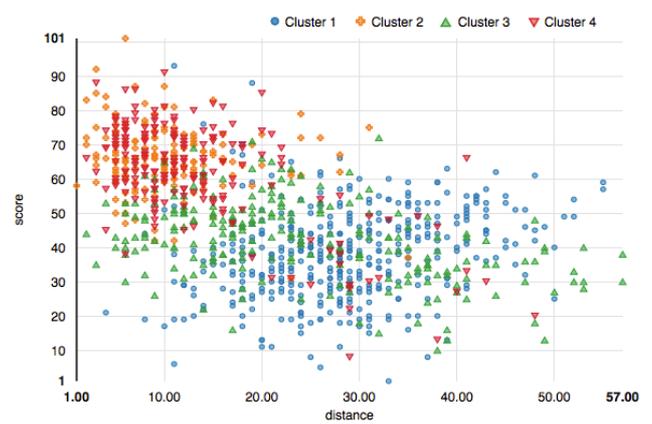


Fig. 5. Distance – score graph

TABLE I. CLUSTER INFORMATION TABLE

	Mean	Standard Deviation	Min. Distance to Centroid	Max. Distance to Centroid
Cluster 1	38.18	12.81	4.00	55.00
Cluster 2	66.11	9.77	1.00	35.00
Cluster 3	42.43	11.54	2.00	57.00
Cluster 4	61.55	14.41	2.00	48.99

A. Open University Learning Analytics Dataset

Open University Learning Analytics Dataset (OULAD) [22] is the first dataset used in our evaluation, containing students’ interaction behaviors with the Open University’s Virtual Learning Environment (VLE) and their assessment results in seven selected courses. There are seven tables in the OULAD dataset, each of which provides different information about the students. During the Studentlyzer evaluation, we eliminated the records of students who withdrew from their courses, and then randomly chose 1,500 student records.

1) Parameter settings

For the students’ behavior records, we chose the average value, standard deviation and total sum of the “sum_click” field in the “studentVLE” table. This field records the number of times a student interacts with the material. We renamed them “avg_click” (average number of clicks that a student pressed for each piece of material), “stdev_click” (standard deviation of number of clicks that a student pressed for each piece of material) and “sum_click” (total sum of clicks that a student pressed during interaction with VLE), respectively. Additionally, we also counted the number of records in the “studentVle” table associated with each student as another type of student behavior, called “num_of_day_used” (number of days student interacted with VLE). To evaluate student performance, we computed the average value of “score” field in the “studentAssessment” table, which records the student scores in an assessment. This indicates academic performance, which is labeled “avg_score”. During the analysis, Studentlyzer classified students based on the above four behaviors (i.e., avg_click, stdev_click, sum_click and num_of_day_used). In other words, each student’s behavior is represented as a four-dimensional vector. Based on student e-learning behavior, Studentlyzer uses a clustering algorithm to group students into a required number of clusters. It is of

interest to evaluate the relationship between student behavior and academic performance.

TABLE II. CLUSTER INFORMATION TABLE OF OULAD

	Mean	Standard Deviation	Min. Distance to Centroid	Max. Distance to Centroid
Cluster 1	73.26	14.42	7.42	2374.72
Cluster 2	82.05	10.36	159.78	17762.89
Cluster 3	78.61	11.25	144.08	5569.21

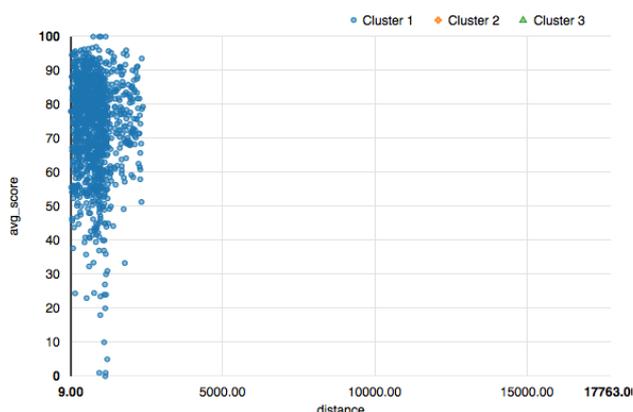


Fig. 6. Distance distribution of cluster 1 with the lowest mean score. Only cluster 1 is shown.

2) Clustering results

Table 2 summarizes the clustering results of the OULAD dataset analyzed by Studentlyzer. It can be seen that students in cluster 2 have the highest mean score value, and students in cluster 1 have the lowest mean score value. Furthermore, students' behavior vectors in cluster 1 have the smallest maximum distance to the cluster centroid, and students' behavior vectors in cluster 2 have the largest maximum distance to the cluster centroid.

The detailed distributions of each cluster are shown in Fig. 6, Fig. 7 and Fig. 8. Comparing Fig. 6 and Fig. 7, it can be seen that points in cluster 2 (i.e., orange points) have a larger variation of distance to centroid than the points in cluster 1 (i.e., blue points). This indicates that students in the cluster with lower academic performance usually have similar behaviors. On the other hand, students in the cluster with higher academic performance tend to have varied behaviors. This may be because some top students may skip certain materials.

3) Analyze connection between behaviors and clusters

Fig. 9 and Fig. 10 show the connection between each student behavior and the cluster results. In Fig. 9 (i.e., sum_click selected as x-axis), clusters do not overlap (i.e., clearly distinct from one another). In Fig. 10 (i.e., avg_click selected as x-axis), the three clusters overlap significantly with one another. This result indicates that the total sum of clicks plays a more important role in the clustering result than the average number of clicks. In Fig. 9, it can be seen that students in cluster 2 have larger sum_click compared to students from other clusters. Furthermore, the avg_score of each student in cluster 2 is higher than 60. This indicates that students with higher values of sum_click tend to achieve a higher academic performance. Finally, we also notice that almost all students with an avg_score lower than 40 have a

small sum_click, which could indicate that a very low sum_click might be linked with a lower avg_score. Therefore, if a student's sum_click is very low (less interaction with the materials), he/she is more likely to perform at a lower level academically. In this case, additional attention and advice may be required.

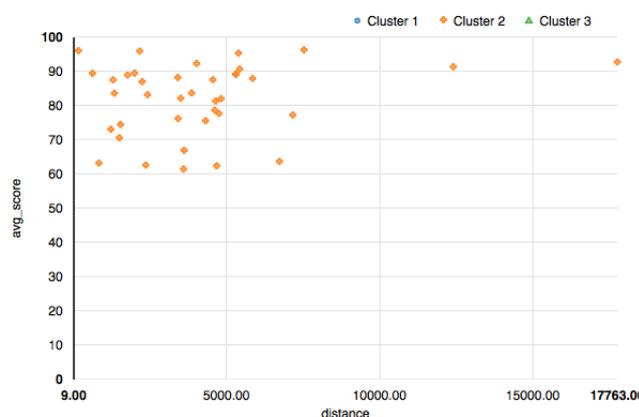


Fig. 7. Distance distribution of cluster 2 with the highest mean score. Only cluster 2 is shown.

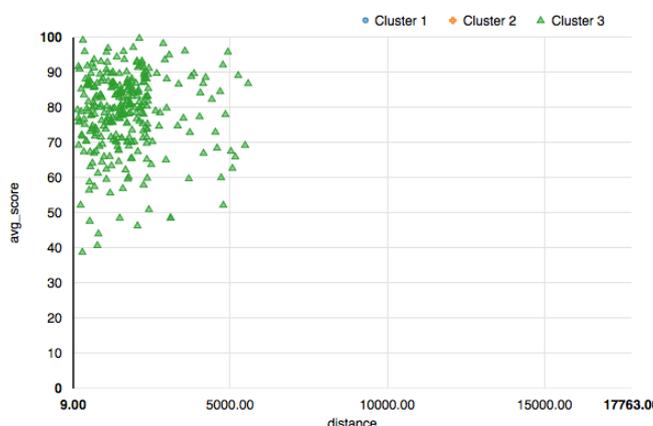


Fig. 8. Distance distribution of cluster 3 with the moderate mean score. Only cluster 3 is shown.



Fig. 9. Connection between the total sum of clicks that a student pressed and different clusters analyzed from the OULAD dataset

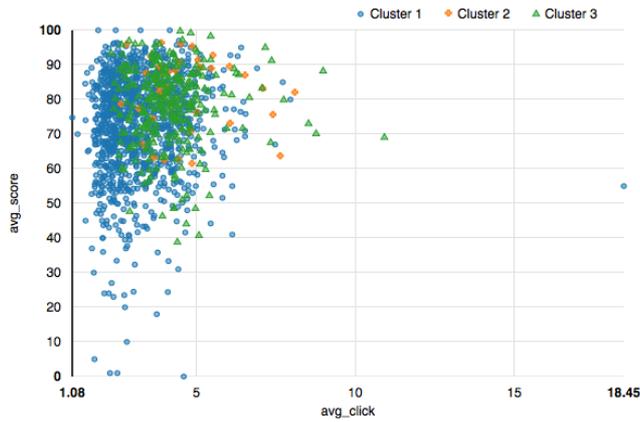


Fig. 10. Connection between the average number of clicks that a student pressed and different clusters analyzed from the OULAD dataset

B. Educational Processing Mining

Educational Processing Mining (EPM) dataset [23] is another dataset used for evaluation. Students are required to learn materials in a browser-like e-learning system called Deeds, and the learning behaviors and final scores are recorded in the dataset.

1) Parameter settings

We used students' total score in the second exam as their academic performance indicator, named exam2_score. The sum of several data fields in all attended sessions was used as students' behavior data, which included the following for each activity: "idle_time" (the duration (milliseconds) of idle time), "mouse_wheel" (the number of mouse wheel events), "mouse_wheel_click" (the number of mouse wheel clicks), "mouse_click_left" (the number of mouse left clicks), "mouse_click_right" (the number of mouse right clicks), "mouse_movement" (the distance covered by the mouse movements), and "keystroke" (the number of keystrokes), according to the dataset description in [23].

2) Clustering results

Table 3 summarizes the clustering results of the EPM dataset, analyzed by Studentlyzer. We find that the mean and standard deviation of these two clusters are very close, even if the maximum distance to cluster centroid and minimum distance to cluster centroid vary widely. This indicates that mean and standard deviation of the score within a cluster do not have a connection with the minimum and maximum distance to centroid.

TABLE III. CLUSTER INFORMATION TABLE OF EPM

	Mean	Standard Deviation	Min. Distance to Centroid	Max. Distance to Centroid
Cluster 1	45.67	20.03	566685313.69	3852312034.95
Cluster 2	50.20	26.39	52617017.72	1969937955.92

3) Analyzing relationship between sum of idle time and second exam score

As shown in Fig. 11, the majority of students are located in cluster 2. The students in cluster 1 have a relatively longer idle time. This indicates that they did not concentrate enough during their studies or the learning materials might have been difficult to understand. This observation is supported by the lower mean score of cluster 1, as shown in Table 3. Within cluster 2, there is an interesting finding. Students with a shorter idle time may not achieve high academic performance.

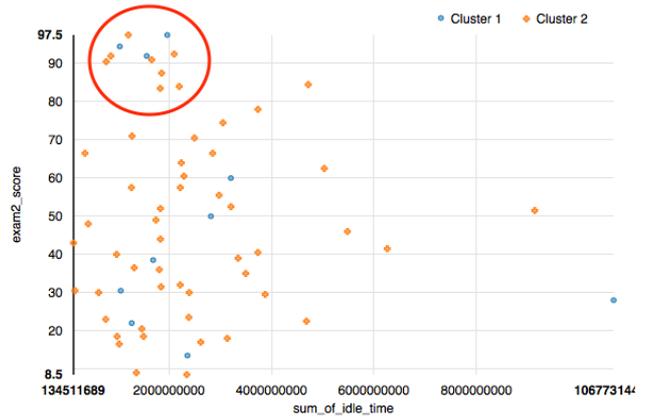


Fig. 11. Connections between sum of students' idle time and second exam score (clustered into two clusters)

However, top students, marked by a circle at the top left in Fig. 11, have a relatively short cumulative idle time. Since most students are placed in cluster 2, we conducted further analysis by increasing the number of clusters from two to six.

4) Comparison between clusters of original data and normalized data

When we increase the number of clusters to six, more interesting patterns can be found, as shown in Fig. 12. In this case, all clusters were spread by the different sum of idle time. This is because the K-mean clustering algorithm uses Euclidean distance to sort different behavior vectors into different clusters. As the value of the sum of idle time field is much larger than those of other fields/parameters, the sum of idle time has more impact on the Euclidean distance, and hence, it has more effect on the clustering result.

To eliminate the aforementioned effect, normalized data can be used for clustering. Essentially, for each behavior data field, we uniformly mapped all data to values between 0 and 1. The following equation shows the mapping method:

$$D'_{ij} = \frac{D_{ij} - \min_{1 \leq k \leq n} D_{kj}}{\max_{1 \leq k \leq n} D_{kj} - \min_{1 \leq k \leq n} D_{kj}} \quad (1)$$

The behavior – score graph generated from the normalized data is shown in Fig. 13. All students are pointed at the same position in Fig. 12 and Fig. 13, but students from the same cluster in Fig. 12 are very likely to belong to different clusters in Fig. 13. This means that the cluster result is not affected solely by sum_of_idle_time.

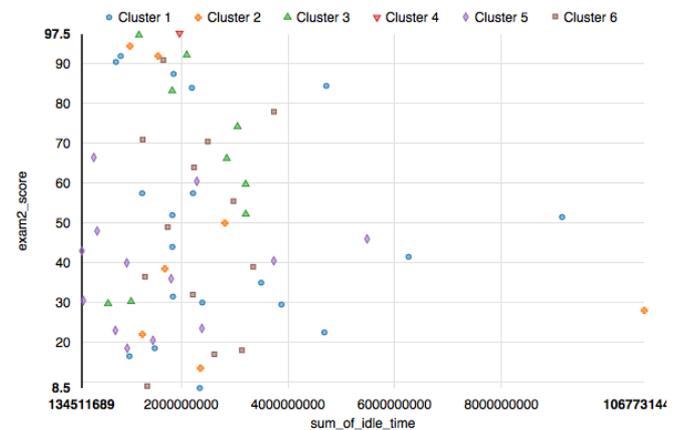


Fig. 12. Cluster distributions of original data (clustered into six clusters)

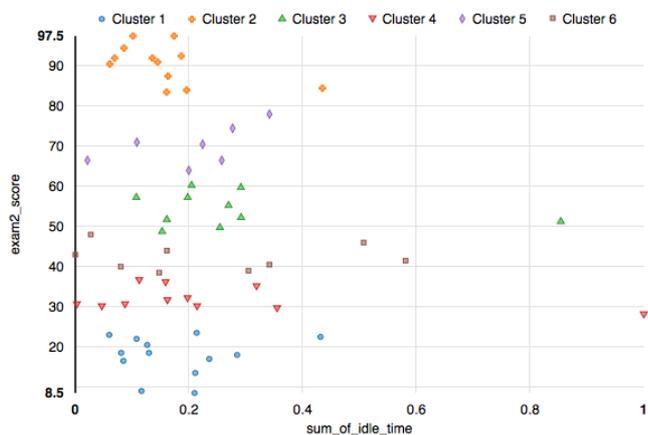


Fig. 13. Cluster distributions of normalized data (clustered into six clusters)

IV. CONCLUSION

With the advent of e-learning, there is a need to develop specialized online tools for analyzing and visualizing e-learning data. In this paper, we present a useful tool called Studentlyzer for this purpose. By means of a clustering algorithm, Studentlyzer clusters e-learning data in a customized way. It provides useful statistical tools and auxiliary graphs for customized analysis. We have tested Studentlyzer with two open e-learning datasets. The test results provide valuable insights into the relationship between student e-learning behavior and academic performance. Ongoing work is being conducted to enhance Studentlyzer, such as comparing the use of different clustering algorithms and providing more advanced features, such as predictive analytics functions.

REFERENCES

[1] D. T. Ku and C.-S. Chang, "The effect of academic discipline and gender difference on taiwanese college students' learning styles and strategies in web-based learning environments." *Turkish Online Journal of Educational Technology-TOJET*, vol. 10, no. 3, pp. 265–272, 2011.

[2] P. Baepler, J. Walker, and M. Driessen, "It's not about seat time: Blending, flipping, and efficiency in active learning classrooms," *Computers & Education*, vol. 78, pp. 227–236, 2014.

[3] F. Gru newaldandC.Meinel, "Implementation and evaluation of digital e-lecture annotation in learning groups to foster active learning," *IEEE Transactions on Learning Technologies*, vol. 8, no. 3, pp. 286–298, 2015.

[4] S. R. Harandi, "Effects of e-learning on students' motivation," *Procedia-Social and Behavioral Sciences*, vol. 181, pp. 423–430, 2015.

[5] H.-T. Hung, "Flipping the classroom for english language learners to foster active learning," *Computer Assisted Language Learning*, vol. 28, no. 1, pp. 81–96, 2015.

[6] L. K. Wright, D. L. Newman, J. A. Cardinale, and R. Teese, "Web-based interactive video vignettes create a personalized active learning classroom for introducing big ideas in introductory biology," *Bioscience: Journal of College Biology Teaching*, vol. 42, no. 2, pp. 32–43, 2016.

[7] (2019, Jan.) International educational data mining society. [Online]. Available: <http://educationaldatamining.org>

[8] A. Peña-Ayala, "Educational data mining: A survey and a data mining-based analysis of recent works," *Expert Systems with Applications*, vol. 41, no. 4, pp. 1432–1462, 2014.

[9] M. Vahdat, A. Ghio, L. Oneto, D. Anguita, M. Funk, and M. Rauterberg, "Advances in learning analytics and educational data mining," *Proc. of ESANN2015*, pp. 297–306, 2015.

[10] R. S. Baker and P. S. Inventado, "Educational data mining and learning analytics," *Learning Analytics*, Springer, pp. 61–75, 2014.

[11] F. Bouchet, R. Azevedo, J. S. Kinnebrew, and G. Biswas, "Identifying students' characteristic learning behaviors in an intelligent tutoring system fostering self-regulated learning," *International Educational Data Mining Society*, ERIC, pp. 65–72, 2012.

[12] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, 2016.

[13] H.-T. Hou, "A case study of online instructional collaborative discussion activities for problem-solving using situated scenarios: An examination of content and behavior cluster analysis," *Computers & Education*, vol. 56, no. 3, pp. 712–719, 2011.

[14] J. McCuaig and J. Baldwin, "Identifying successful learners from interaction behaviour." *International Educational Data Mining Society*, 2012.

[15] R. L. Sparks, J. Patton, and L. Ganschow, "Profiles of more and less successful 12 learners: A cluster analysis study," *Learning and Individual Differences*, vol. 22, no. 4, pp. 463–472, 2012.

[16] M. M. Quadri and N. Kalyankar, "Drop out feature of student data for academic performance using decision tree techniques," *Global Journal of Computer Science and Technology*, 2010.

[17] B. K. Baradwaj and S. Pal, "Mining educational data to analyze students' performance," *Computing Research Repository - arXiv*, vol. abs/1201.3417, 2012.

[18] M. Nasiri, B. Minaei, and F. Vafaei, "Predicting gpa and academic dismissal in lms using educational data mining: A case mining," *Proc. 6th National and 3rd International Conference of e-Learning and e-Teaching*, pp. 53–58, 2012.

[19] A. Dutt, M. A. Ismail, and T. Herawan, "A systematic review on educational data mining," *IEEE Access*, vol. 5, pp. 15991–16005, 2017.

[20] M. H. I. Shovon and M. Haque, "Prediction of student academic performance by an application of k-means clustering algorithm," *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 2, no. 7, 2012.

[21] J. Han, J. Pei, and M. Kamber, *Data Mining: Concepts and Techniques*. Elsevier, 2011.

[22] (2019, Jan.) Open university learning analytics dataset. [Online]. Available: <http://dx.doi.org/10.1038/sdata.2017.171>.

[23] M. Vahdat, L. Oneto, D. Anguita, M. Funk, and M. Rauterberg, "A learning analytics approach to correlate the academic achievements of students with interaction data from an educational simulator," *Design for Teaching and Learning in a Networked World*, Springer, pp. 352–366, 2015.