A data mining approach using unsupervised learning for profiling students

Khor Ean Teng

National Institute of Education, Singapore

Abstract

The paper presents a data mining approach using unsupervised learning for profiling students. Unsupervised learning specifically the K-means clustering algorithm is applied to obtain clusters with similar patterns and characteristics. The clustering experiments were performed using academic background, parental support, and learning behavioural features as attributes. The characteristics that distinguish students belonging to those different clusters were examined. The findings uncovered the key characteristics of students' performance, and it is helpful for future prediction. Appropriate learning support and intervention could be provided to tailor to the individual cluster of students to enhance their performance. The clustering algorithm also serves as a potential benchmark to monitor the progress of students' performance and helps teachers to improve the course success.

Introduction

In education, unsupervised learning clustering can be used to find students' clusters with similar patterns and characteristics. The key characteristics of the performance of students are discovered and they can be used for future prediction. For example, Oyelade et al. (2010) apply clustering algorithms to predict students' academic performance while Shovon et al. (2012) use clustering algorithms to improve students' academic performance by predicting students' learning activities.

Besides prediction, the clustering algorithm can also be leveraged to profile students. For example, Bovo et al. (2013) mine Moodle log data to cluster the students with similar behaviour patterns. Their research works aim to forecast the performance of students during an online curriculum on a learning management system to prevent students to fall behind. Bouchet et al. (2013) also apply clustering algorithms to cluster and profile students based on their interactions within Intelligent Tutoring System. The interactions

CONTACT Khor Ean Teng, eanteng.khor@nie.edu.sg

include mouse clicks and keyboard entries within the system, time spent taking notes, the number of times visiting each page and duration on the page. Three clusters were formed for characterization and profiling.

In this study, unsupervised learning was applied to assess the effect of students' academic background, parental support, and learning behavioural features on their academic performance. The study makes use of clustering data mining techniques to cluster students into groups according to their characteristics. Clustering data mining was applied in this to obtain clusters which were then mapped to identify the important attributes of a learning context. Data clustering involves the process of extracting hidden patterns and positional useful from large data sets (Shovon et al. 2012). The large data sets are then segmented into smaller sets which are known as clusters.

Methods

In this research work, unsupervised clustering data mining techniques were applied to Amrieh et al. (2016)'s dataset to assess the effect of (1) academic background, (2) parental support, and (3) learning behavioural feature category on the performance of students. The dataset is event log data collected from a learning management system through Experience API.

Figure 1 summarizes the main research processes carried out in this study. Data pre-processing was conducted to clean the missing data and convert raw data into an appropriate form for further processing. 20 records with missing values were removed from the 500 records. After the data pre-processing, exploratory data analysis was conducted to analyse the dataset and uncover patterns.

Next, feature selection was conducted to reduce the number of features that are insignificant based on the weighting score. Information Gain (IG) was applied in this study to conduct feature selection. An IG algorithm was employed by the feature ranking stage that is based on a filtering approach. It evaluates the gain of each independent category feature in the context of the dependent category feature. The higher the information gain score, the lower the entropy group(s) of samples. The feature that has the larger information gain was selected for the modelling. Table 1 shows the ranking of features according to the information gain scoring and Table 2 describes the feature and its data type.

Finally, clusters were obtained using the K-means clustering algorithm which was further mapped to find the significant feature. The relationship between these features was then identified to assess students' performance.

Figure 1

Main research processes

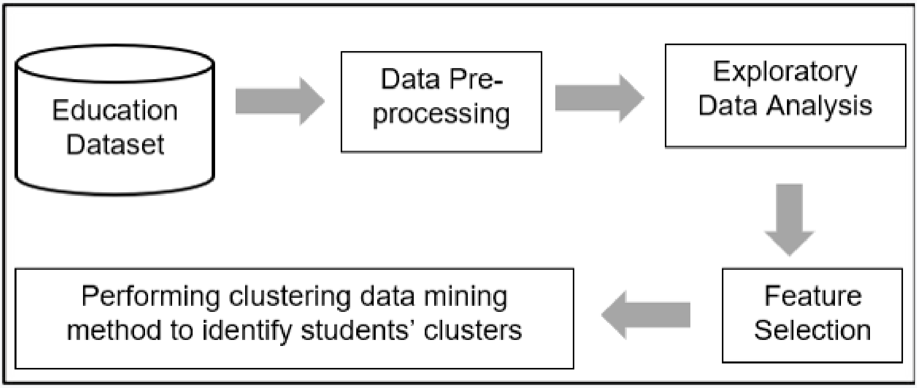
Education Dataset
Data Pre-processing
Exploratory Data Analysis
Feature Selection
Performing clustering data mining method to identify students' clusters

Table 1

Feature selection

Feature	Information Gain Score
student_absence_day	0.397
visited_resources	0.391
raised_hand	0.362
announcements_view	0.253
parent_answering_survey	0.150
relation	0.126
discussion	0.088
stage_id	0.011

Table 2

Description of features

Feature Category	Feature	Description	Data Type
Academic Background	stage_id	Level of education	Nominal
Parental Support	relation	Primary caregiver	Nominal
	parent_school_ satifaction	Parent satisfaction with the school	Nominal
Learning Behavioural	raised_hand	Question raised count	Numeric
	visited_resources	Course content visit count	Numeric
	announcements_view	Announcement view count	Numeric
	discussion	Discussion participation count	Numeric
	student_absence_day	Absence days count	Nominal

Data analysis and results

The features are visualized according to the clusters (Figures 2 to 9). The blue nodes represent "High-performer (H)"; green nodes represent "Moderate-performer (M)", and red nodes represent "Low-performer (L)".

Table 3 displays the number of students according to the clusters with the highest numbers of students in cluster 2. The cluster analysis results based on the 8 factors are summarized in Table 4. Cluster 1 represents low-performer, cluster 2 represents moderate-performer, and cluster 3 represents high-performer. For the education level, the result indicates that most of the students are 'lower level' in cluster 1. Most students are 'Middle school' in cluster 2 and cluster 3 respectively. Cluster 1 has the lowest numbers while cluster 3 has the highest numbers of visited_resources, raised_hand, announcement_view, and discussion.

Conclusions

Unsupervised learning, particularly clustering was applied to data with 480 data points to gain insights into students' academic background, parental support, and their learning behavioural patterns with diverse populations. Diversity influences the way students behave and learn. Students who are struggling or have low progress could be identified

Figure 2

Feature (stage_id)

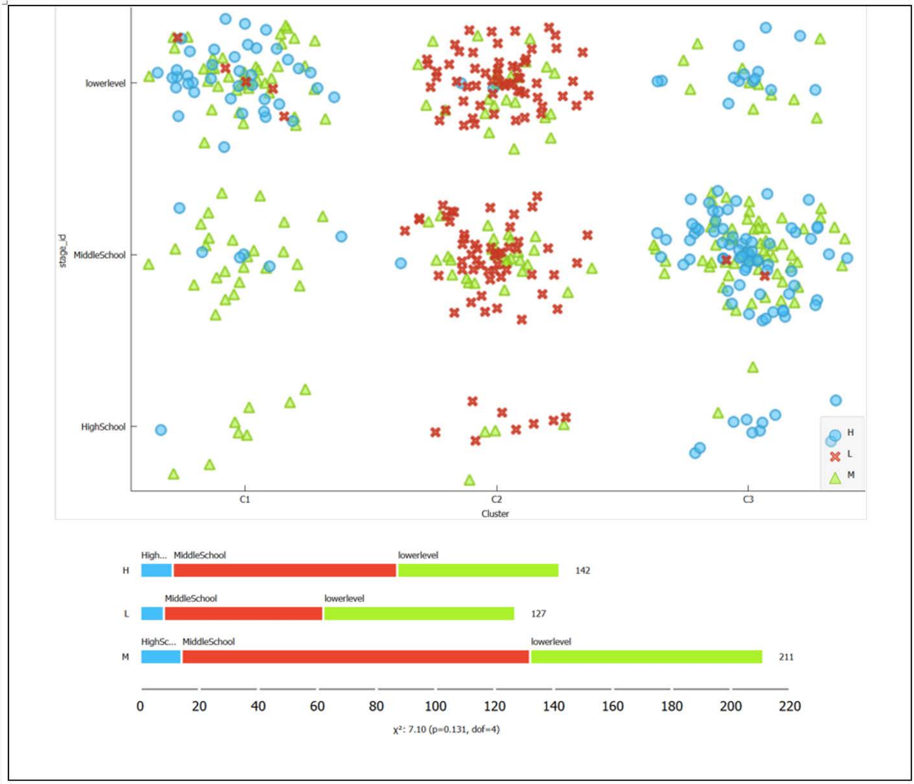

Figure 3

Feature (relation)

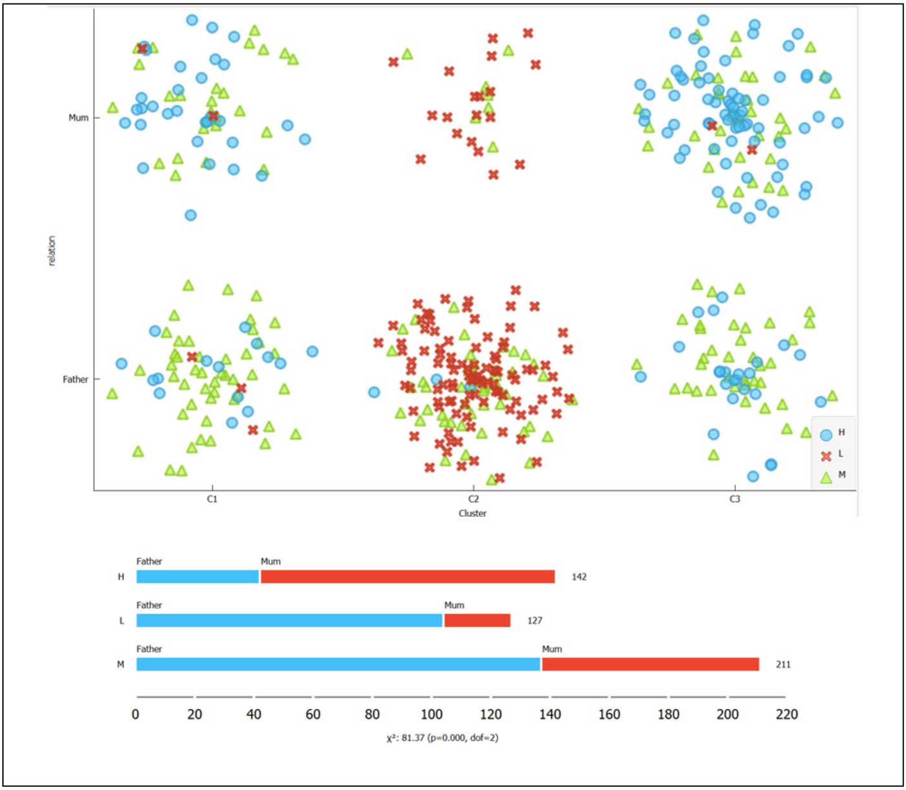

Figure 4

Feature (parent_school_satisfaction)

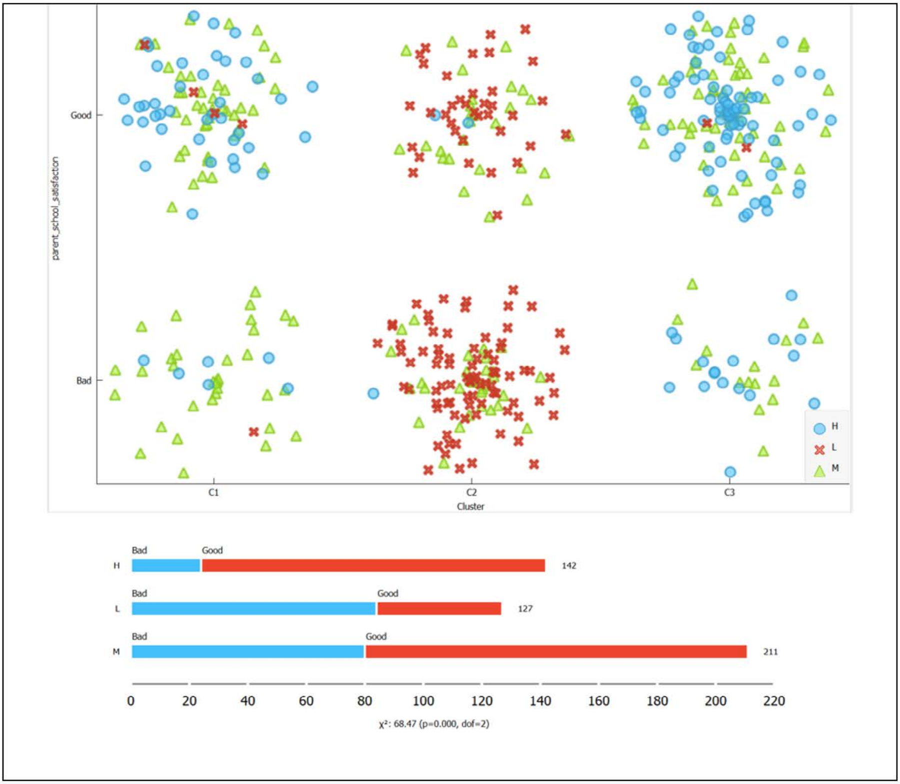

Figure 5

Feature (raised_hand)

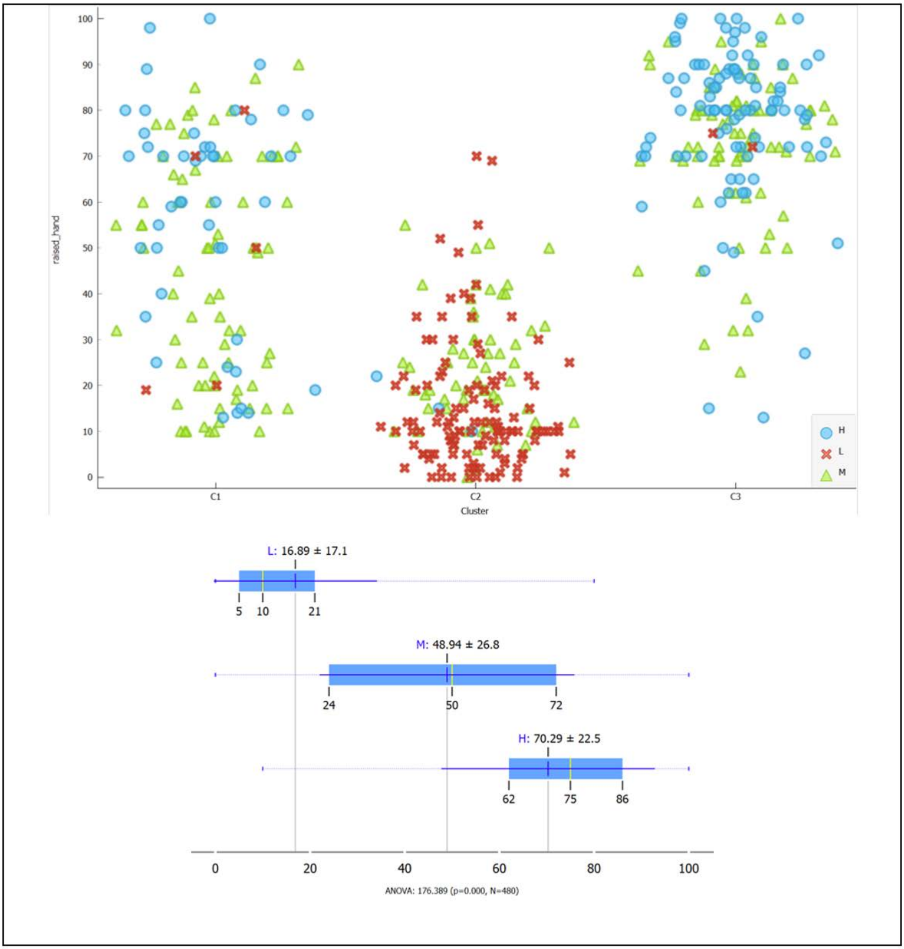

Figure 6

Feature (visited_resources)

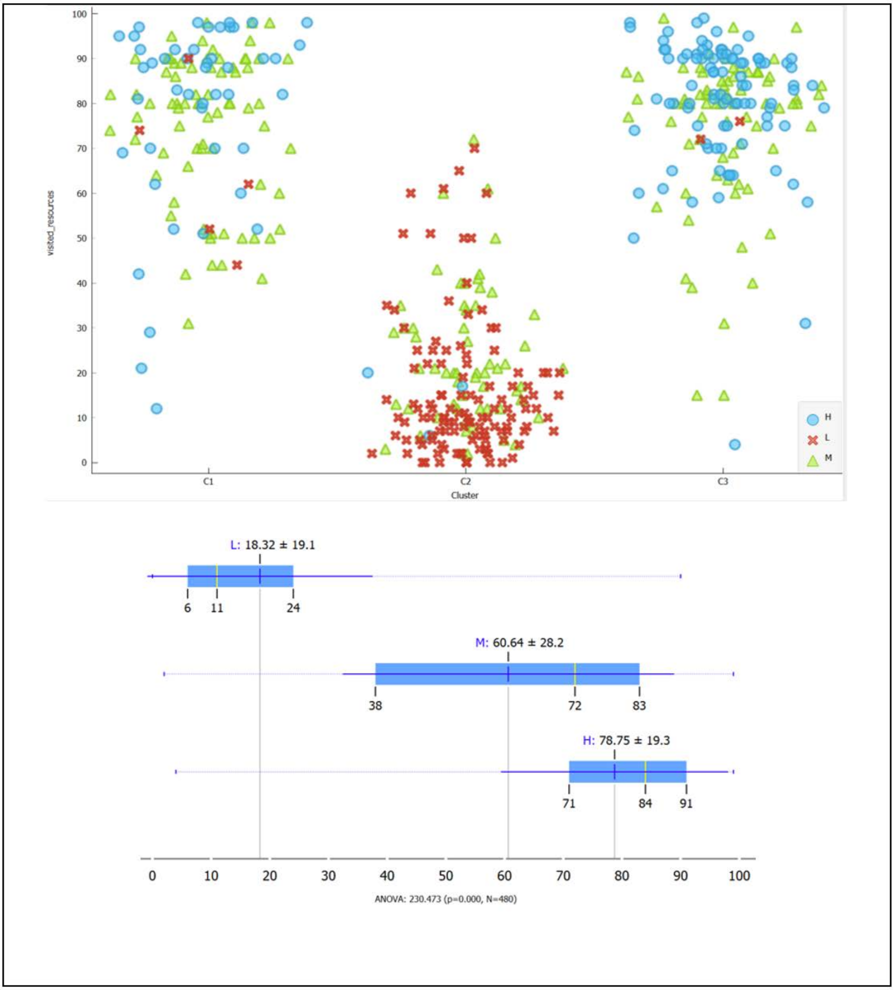

Figure 7

Feature (announcements_view)

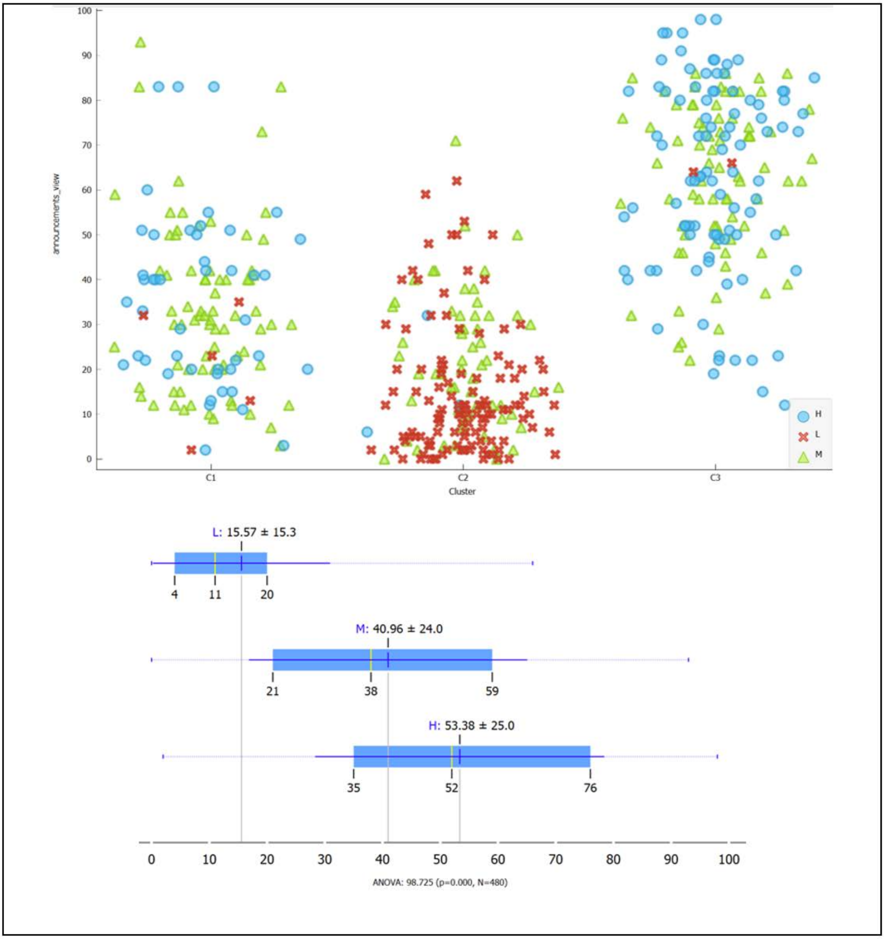

Figure 8

Feature (discussion)

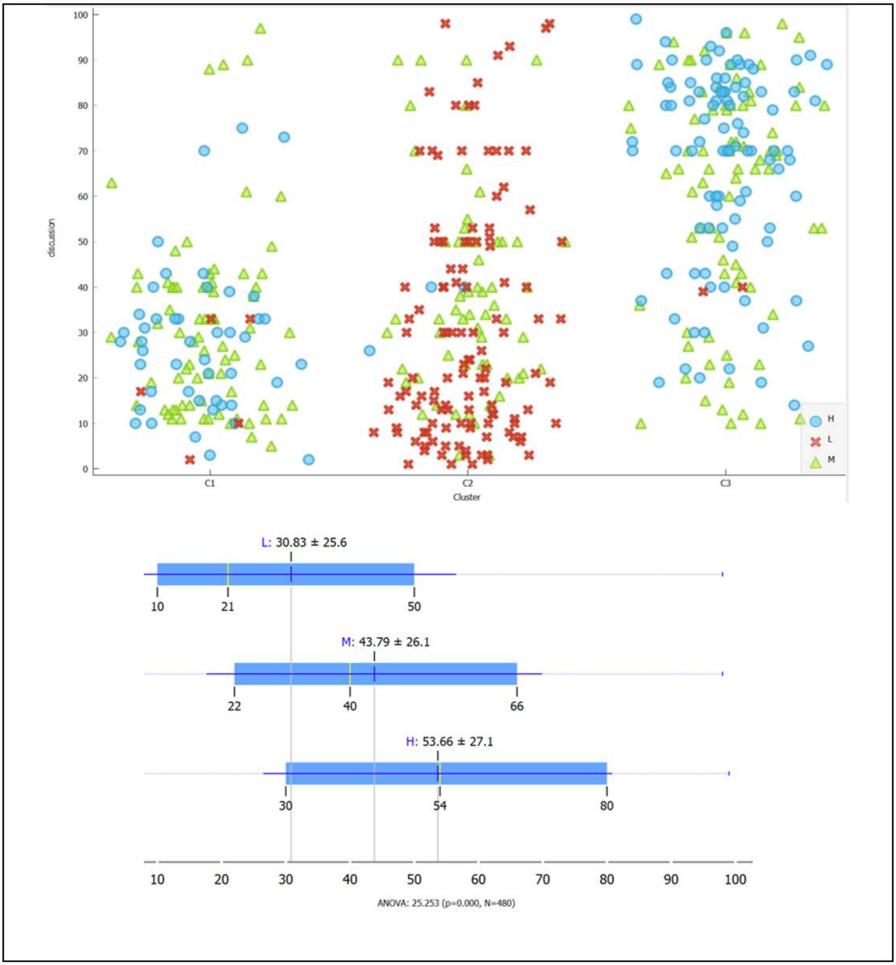

Figure 9

Feature (student_absence_day)

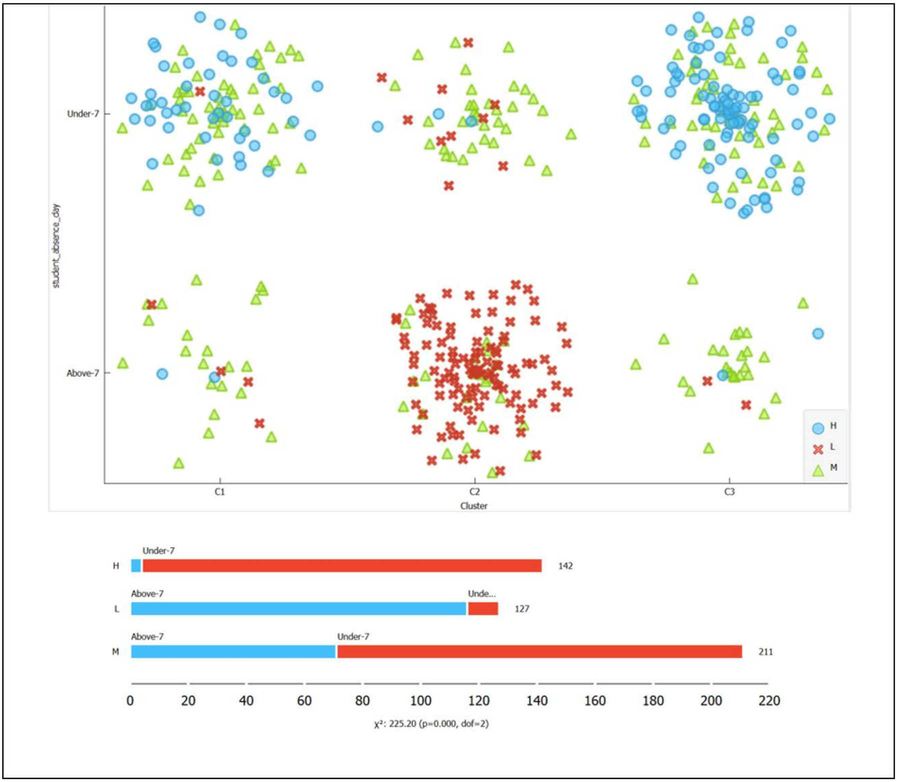

Table 3

Number of students according to cluster

Cluster	1	2	3	Total
1	125	0	0	125
2	0	185	0	185
3	0	0	170	170
Total	125	185	170	480

Table 4

Cluster analysis summary

Feature	Cluster 1 (L)	Cluster 2 (M)	Cluster 3 (H)
stage_id	Lower level	Middle school	Middle school
relation	Father	Father	Mother
parent_school_ satifaction	Bad	Good	Good
raised_hands	16.89±17.1	48.94±26.8	70.29±22.5
visited_resources	18.32±19.1	60.64±28.2	78.75±19.3
announcements_view	15.57±15.3	40.96±24	53.38±25
discussion	30.83±25.6	43.79±26.1	53.66±27.1
student_absence_days	Above-7	Under-7	Under-7

from the patterns. The results show that academic background, parental support, and learning behavioural are important features that affect learners' academic performance. Students' heterogeneous data were analyzed with cluster-based data mining. The clustering findings help teachers to identify similar (and different) groups of students. Appropriate learning support and intervention could then be provided to tailor to the individual cluster of students to enhance their performance. For example, different types of materials could be recommended to the individual cluster of students. Further works may include the application of other clustering methods (eg: Expectation-Maximization, Mean-shift) to improve the cluster data mining analysis. Besides, further analysis can be performed to examine the inter-relationships among the features.

References

Amrieh, E. A., Hamtini, T., & Aljarah, I. (2016). Mining educational data to predict Student's academic performance using ensemble methods. *International Journal of Database Theory and Application, 9*(8), 119-136.

Bouchet, F., Harley, J. M., Trevors, G. J., & Azevedo, R. (2013). Clustering and profiling students according to their interactions with an intelligent tutoring system fostering self-regulated learning. *Journal of Educational Data Mining, 5*(1), 104–146. https://doi.org/10.5281/zenodo.3554613

Bovo, A., Sanchez, S., Héguy, O., & Duthen, Y. Clustering moodle data as a tool for profiling students. In *2013 Second International Conference on E-Learning and E-Technologies in Education (ICEEE), 2013* (pp. 121-126): IEEE

Fahim, A., Salem, A., Torkey, F. A., & Ramadan, M. (2006). An efficient enhanced k-means clustering algorithm. *Journal of Zhejiang University-Science A, 7*(10), 1626-1633.

Oyelade, O. J., Oladipupo, O. O., & Obagbuwa, I. C. (2010). Application of k means clustering algorithm for prediction of students academic performance. *International Journal of Computer Science and Information Security, 7*(1), 292-295.

Shovon, M., Islam, H., & Haque, M. (2012). An approach of improving students academic performance by using k means clustering algorithm and decision tree. *International Journal of Advanced Computer Science and Applications, 3*(8), 146-149.