Title	A learning analytics approach using clustering data mining for learners profiling to extrapolate e-learning behaviours
Author(s)	Khor Ean Teng

This is the author accepted manuscript of the following conference paper:

Khor, E.T. (2021). A learning analytics approach using clustering data mining for learners profiling to extrapolate e-learning behaviours. In T. Bastiaens (Ed.), *Proceedings of Innovate Learning Summit 2021 Online* (pp. 59-64). Association for the Advancement of Computing in Education (AACE). https://www.learntechlib.org/primary/p/220271/

A Learning Analytics Approach Using Clustering Data Mining for Learners Profiling to Extrapolate e-Learning Behaviours

Khor Ean Teng
National Institute of Education, Nanyang Technological University
Singapore
eanteng.khor@nie.edu.sg

Abstract: The study aims to gain insights into the patterns related to the diversity of learners and evaluate the relationship between learners' factors and their academic performance. The pattern discovery was performed by applying clustering data mining to obtain typologies of the learners based on the academic records and e-learning interaction behaviour feature category. In this study, the k-means clustering unsupervised machine learning algorithm was applied to obtain the clusters. The clustering analysis of learners' academic records and interaction behavioural patterns between learner sub-populations allows for a better understanding of how the learners behave and achieve. The clustering results identified similar group learners, and the learners could be provided with appropriate educational supports and approaches to enhance learners' learning experience. The findings of this study are also useful to understand the effects of different features on learners' academic performance specifically in an e-learning environment.

Introduction

Different educational approaches need to be designed to be effective for learners with diverse backgrounds (Ferguson & Sharples, 2014). Individual learners' differences such as behavioural factors affect the outcomes of a learning experience (Alamri et al., 2019; Cristea et al., 2018). The efforts of past research include those aiming at examining learners' completion, dropouts, engagement and or motivation (Alario-Hoyos, Estévez-Ayres, Pérez-Sanagustín, Kloos, & Fernández-Panadero, 2017; Gardner & Brooks, 2018; Gregori, Zhang, Galván-Fernández, & de Asís Fernández-Navarro, 2018; Sunar, White, Abdullah, & Davis, 2016). However, there is limited study of patterns related to learner diversity. Hence, this study aims to use learning analytics to reveal the hidden learning patterns of learners. The hidden patterns were investigated through the process of data mining.

This study worked on an anonymised publicly available education dataset. The source of the dataset is from Kuzilek, Hlosta, and Zdrahal (2017). The techniques of clustering data mining are adopted in this study to examine the effect of (1) learners' academic records and (2) e-learning interaction feature category on learners' academic performance. Learners' academic performance is determined by the result of the learner's effort and the final result awarded is in either of the four categories: (i) distinction, (ii) pass, (iii) fail and (iv) withdrawn. The learning process could be improved by gaining actionable insight into the learners' profiles so that they are provided with the appropriate intervention and learning supports as early as possible.

A clustering algorithm was applied to find groups of similar instances in the dataset. Amongst them, k-means non-hierarchical clustering algorithm is widely used and most popular. Besides, the k-means technique had been shown effective in producing promising clustering results (Alsabti, Ranka, & Singh, 1997). Hence, the choice of k-means in this study to perform clustering data mining.

Research Methods

This research work performed clustering data mining technique to assess the effect of (1) academic records and (2) e-learning interaction feature category on learners' academic performance. Figure 1 summarizes the main research steps carried out in this study.

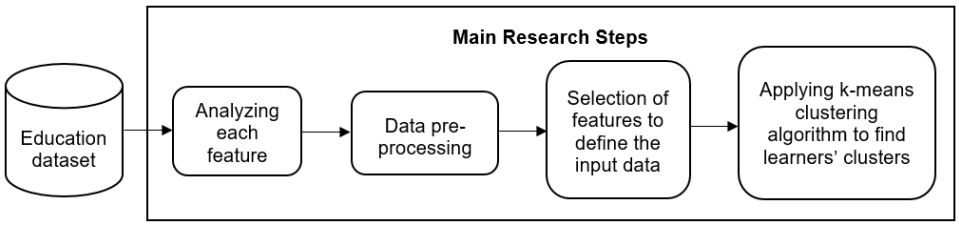

Figure 1. Summary of Main Research Steps

For the first step, each feature available in the dataset was analyzed. Data pre-processing steps were then carried out to apprehend the nature of the features. During the data-pre-processing process, data cleaning was performed to remove the missing values and irrelevant items of selected target data. The dataset after cleaning becomes 3,929 records. After data pre-processing, feature selection was conducted to reduce insignificant features and hence dimensionality. The selected features will then be used to define the input data. A correlation-based feature selection with the *CorrelationAttributeEval* technique was applied to obtain the weighted score of each feature. The higher the weighted score of the feature, the strongest its relationship with the output variable (Khor, 2020). Table 1 shows the selected features used in the study (based on the top five weighted scores) according to the feature category and the descriptive statistics of the features are presented in Table 2 and Table 3.

Table 1. *Depiction of Features*

Feature Category	Feature	Feature Description
Academic records	*highest_education*	The learner's highest level of education on entry to the module
	num_of_prev_attempts	The number of times the learner has enrolled on the module
	studied_credits	The learner's module credits he or she is currently studying
	assessment_score	The learner's assessment score (ranging from 0 to 100)
e-learning interaction	*sum_clicks*	The number of times the learner has e-learning interaction

Table 2. *Descriptive Statistic of Features (a)*

Features	Label	Value	Count	Percentage (%)
highest_education	Post graduate Qualification	5	15	0.38
	HE Qualification	4	539	13.72
	A-level or Equivalent	3	1786	45.46
	Lower than A level	2	1556	39.60
	No Formal quals	1	33	0.84

Next, the k-means machine learning algorithms were applied to cluster the learners and characterize the learners' segments. The k-means algorithms normalized the numerical attributes and used Euclidean distance (-R first last) measure to calculate the distances between the instances and clusters. The test mode was via classes to clusters evaluation on the training data. The k-means clustering algorithm was performed to calculate the number of instances with a label that has been clustered to each cluster based on most instances of some class in each cluster.

Table 3. *Descriptive Statistic of Features (b)*

Features	Minimum	Maximum	Mean	Standard Deviations
num_of_prev_attempts	0	5	0.223	0.566
studied_credits	30	420	84	39.953
assessment_score	0	100	71	13.99
sum_clicks	1	15,716	932	27.638

Analysis and Results

The four clusters (fail, withdrawn, pass and distinction) are analysed based on the top five features (*highest_education*, *num_of_prev_attempts*, *studied_credit*, *assessment_score* and *sum_clicks*). Figures 2 and 6 display the boxplot of the five features respectively. For the *highest_education* feature, the higher the score, the higher the level of the highest education. The academic status of distinction is around the score of 3.01 (eg: A-level or above). For the *num_of_prev_attempts* feature, the higher the score, the greater number of previous attempts. As observed from Figure 4 (*studied_credit*), the students who are enrolling with around 91 credits will tend to withdraw. For *assessment_score* and *sum_clicks* features, the higher the score, the higher the score of the assessment and e-learning interaction. The academic status of distinction is around 80.79 for assessment score and with 1567 clicks.

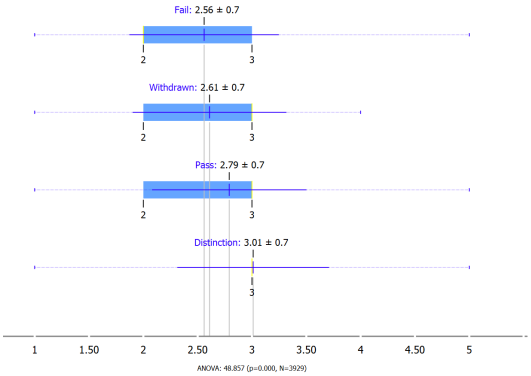

Figure 2. Boxplot (*highest_education*)

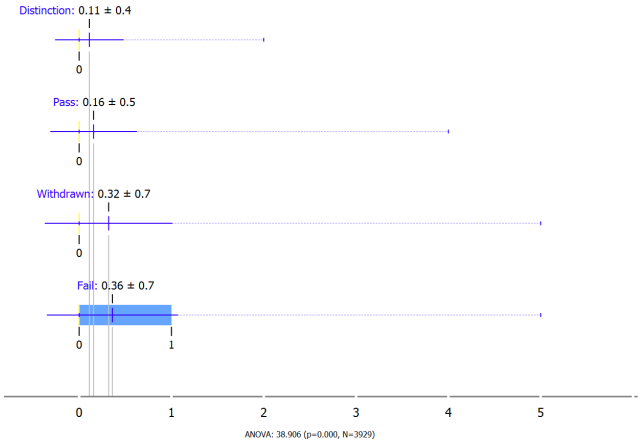

Figure 3. Boxplot (*num_of_prev_attempts*)

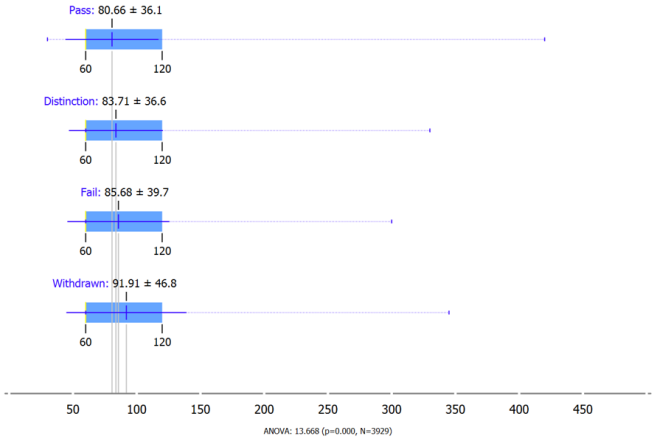

Figure 4. Feature (*studies_credits*)

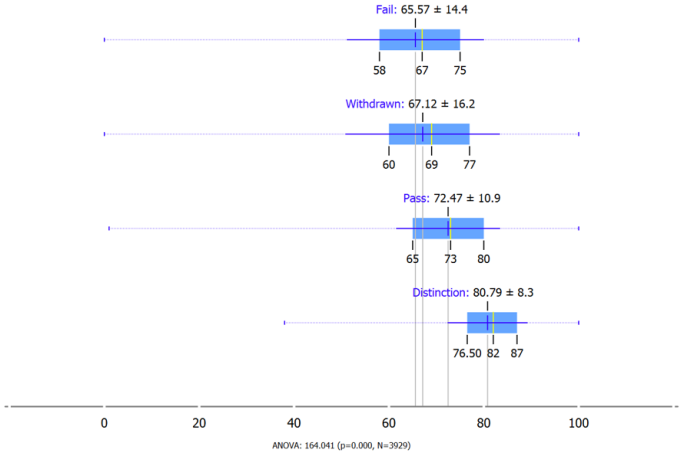

Figure 5. Feature (*assessment_score*)

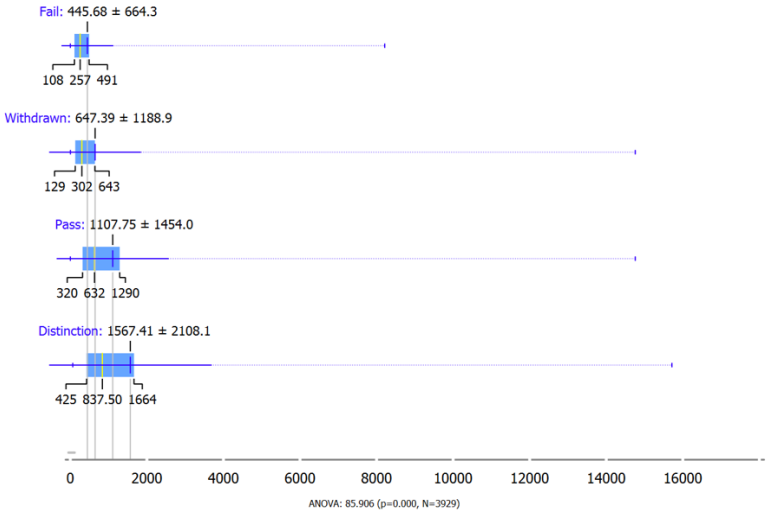

Figure 6. Feature (*sum_clicks*)

Figure 7 and Figure 8 show two of the samples cluster formation of the study. The yellow, blue, red and green colors represent cluster 1 (n=462), cluster 2 (n=406), cluster 3 (n=949) and cluster 4 (n=2112) respectively. For the level of highest education, the result indicates that the majority of the learners are 'A-level or equivalent' in clusters 1 and 4. The majority of the learners are 'HE qualification' in cluster 2 and 'Lower than A-level' in cluster 3. For the number of previous attempts, we observe that those who achieved distinction are mostly attempted just once. There are a few who has failed and withdrawn attempted up to a maximum of five times. For learners' total studied credits, the minimum and the maximum number of credits is 30 and 420 respectively with a mean, 84. It is observed that those with distinction has relatively low studied credits. The minimum and maximum assessment score is 0% and 100% respectively with a mean, 71%. It is also observed that a few of those who are withdrawn and fail has zero assessment score. The minimum and the maximum number of clicks is 1 and 15,716 respectively with a mean, 932. It is also shown that the one who has the most clicks achieved distinction status and those who obtained failed status have the lowest number of clicks.

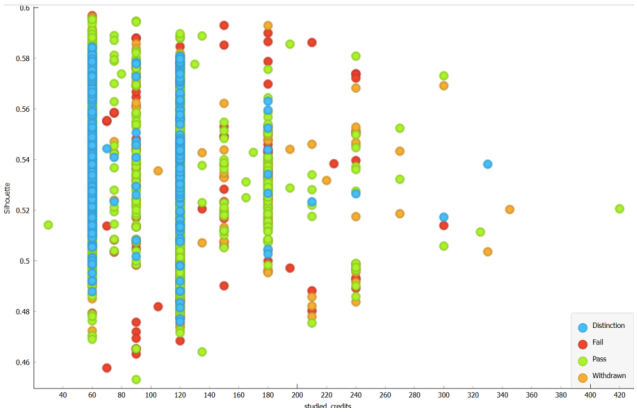

Figure 7. Clusters Formation (*studied_credits)*

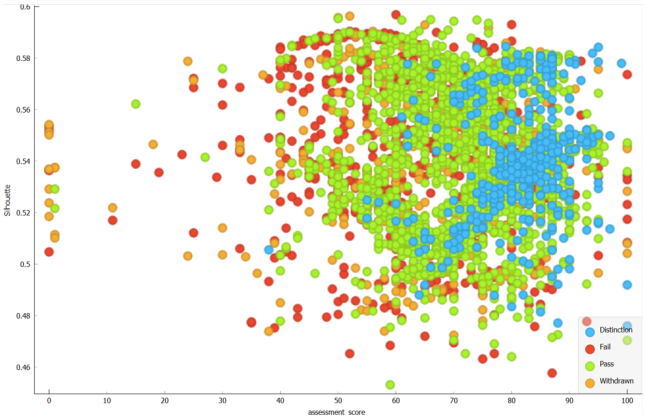

Figure 8. Clusters Formation (*assessment_score)*

Table 4. *Summary of Cluster Analysis*

Features	Cluster 1 Withdrawn n=462 (12%)	Cluster 2 Distinction n=406 (10%)	Cluster 3 Fail n=949 (24%)	Cluster 4 Pass n=2112 (54%)
highest_education	Low	Highest	Lowest	High
num_of_prev_attempts	High	Lowest	Highest	Low
studied_credits	Highest	Low	High	Lowest
assessment_score	Low	Highest	Lowest	High
sum_clicks	Low	Highest	Lowest	High

The cluster analysis based on the factors (*highest_education*, *num_of_prev_attempts*, *studied_credit*, *assessment_score* and *sum_clicks*) that affect the learners' academic performance is summarized in Table 4. It was observed that both academic record and e-learning interaction features influence learners' overall academic performance. The learners who achieved distinction has the lowest *number_of_prev_attempts* but with the highest education level, *assessment_score* and *sum_clicks*. On the other hand, those who fail have the highest *number_of_prev_attempts* but with the lowest education level, *assessment_score* and *sum_clicks*.

Conclusions

This study worked on data with 3,929 data points using data analytics and computing techniques. It presents an insight into academic records and e-learning interaction behavioural patterns of learners with diverse learner populations. The diversity affects how learners behave and achieve and the patterns might tell which learners are struggling. The findings show that academic records and e-learning interaction are significant features affecting learners' learning outcomes. The cluster-based approach to learning analytics can be used in practice and is beneficial to analyse heterogeneous data of learners and can help suggest possible improvement in curriculum design and learning support strategy. The clustering results allow instructors to identify similar group learners and they could provide the learners with appropriate learning support and intervention. The different types of materials (direct hyperlinks to specific learning resources and videos) could then be recommended to the individual cluster of learners. The instructors could also provide in-time or early intervention and feedback to the learners about their learning behaviour to better facilitate learner support. In future, various other clustering methods can be applied to improve the cluster data mining analysis. One of the limitations of the k-means clustering is it assigns empty clusters when no points are allocated to a cluster (Singh, Malik, & Sharma, 2011). In addition, more analysis can be conducted to study the inter-relationships between different features. The other features like learning behavioural features and administrative usage features could be considered in the extension work.

References

Alamri, A., Alshehri, M., Cristea, A., Pereira, F. D., Oliveira, E., Shi, L., & Stewart, C. (2019). *Predicting MOOCs dropout using only two easily obtainable features from the first week's activities.* Paper presented at the International Conference on Intelligent Tutoring Systems.

Alario-Hoyos, C., Estévez-Ayres, I., Pérez-Sanagustín, M., Kloos, C. D., & Fernández-Panadero, C. (2017). Understanding learners' motivation and learning strategies in MOOCs. *The International Review of Research in Open and Distributed Learning, 18*(3).

Alsabti, K., Ranka, S., & Singh, V. (1997). An efficient k-means clustering algorithm.

Cristea, A. I., Alamri, A., Alshehri, M., Kayama, M., Foss, J., Shi, L., & Stewart, C. D. (2018). *Can learner characteristics predict their behaviour on MOOCs?* Paper presented at the Proceedings of the 10th International Conference on Education Technology and Computers.

Ferguson, R., & Sharples, M. (2014). *Innovative pedagogy at massive scale: teaching and learning in MOOCs.* Paper presented at the European Conference on Technology Enhanced Learning.

Gardner, J., & Brooks, C. (2018). Student success prediction in MOOCs. *User Modeling and User-Adapted Interaction, 28*(2), 127-203.

Gregori, E. B., Zhang, J., Galván-Fernández, C., & de Asís Fernández-Navarro, F. (2018). Learner support in MOOCs: Identifying variables linked to completion. *Computers & Education, 122*, 153-168.

Khor, E. T. (2020). Features Identification and Classification of Discussion Threads in Coursera MOOC Forums. In *Transforming Teaching and Learning in Higher Education* (pp. 189-202): Springer.

Kuzilek, J., Hlosta, M., & Zdrahal, Z. (2017). Open university learning analytics dataset. *Scientific data, 4*, 170171.

Singh, K., Malik, D., & Sharma, N. (2011). Evolving limitations in K-means algorithm in data mining and their removal. *International Journal of Computational Engineering & Management, 12*(1), 105-109.

Sunar, A. S., White, S., Abdullah, N. A., & Davis, H. C. (2016). How learners' interactions sustain engagement: a MOOC case study. *IEEE Transactions on Learning Technologies, 10*(4), 475-487.