International Journal of Advanced Engineering, Management and Science (IJAEMS)
Peer-Reviewed Journal
ISSN: 2454-1311 | Vol-7, Issue-9; Sep, 2021
Journal Home Page: https://ijaems.com/
Article DOI: https://dx.doi.org/10.22161/ijaems.79.1

Clustering of Learners based on Readiness to Online Modality using K-Means Algorithm

Daryl B. Valdez, Rey Anthony G. Godmalin

[1]BSCS Department, Bohol Island State University – Clarin Campus, Philippines

Received: 22 Jul 2021; Received in revised form: 22 Aug 2021; Accepted: 01 Sep 2021; Available online: 08 Sep 2021

Abstract— *Clustering is one of the important techniques in data mining. It is an unsupervised task of grouping similar data. It has been applied in various fields with high degree of success. This study aimed to determine the learner segments based on readiness to online learning modality using K-means algorithm. A dataset was collected, tabulated and pre-processed. Further, the values were scaled and transformed using t-distributed Stochastic Neighbor Embedding. Using elbow method and determining the silhouette score, the best K value was determined. Then clustering was conducted using the selected number of clusters. Results revealed three groups of learners; Moderate-signal mobile users, Low-signal mobile users, and mixed group of Low/moderate-signal mobile/broadband users. Students from the different clusters are more suited for flexible learning as opposed to online learning. Varied learning modalities can be catered for students from the different learner segments. Formulation and adoption of new policies are needed to offset the effect of the pandemic towards the students.*

Keywords— ***Clustering, K-means algorithm, data mining, online learning modality, learner's segmentation.***

I. INTRODUCTION

Clustering is an unsupervised task of dividing data points into a fixed number of groups wherein the data points of a group bears close similarity and are different from those in other groups (Syakur et al, 2018). K-means algorithm is one of the methods of clustering data. It is the most commonly used clustering method due to its speed and simplicity (Yuan et al, 2019). Clustering has a variety of applications in various fields including; market segmentation, medical imaging, social network analysis, image segmentation and anomaly detection. Not only that, recent studies revealed that it can also be useful in the field of academe.

A study was conducted and used clustering to classify learners according to learning style preferences (Pasina et al, 2019). Results of the study revealed student outliers which have different learning style from the rest allows instructors to properly address their concerns. Further, clusters of students with similar learning styles allows ease of work on class assignments.

Another study was also conducted using hierarchical clustering in grouping students according to learning style (Yotaman et al, 2020). The experimental results show that grouping students into seven clusters using the Euclidean distance function and the ward linkage criteria yields the highest efficiency in clustering. The resulting clusters can help identify the behaviors and learning skills of students which will enable teachers more options in selecting and using appropriate methods and teaching strategies.

Aside from segmenting learners, cluster analysis can also be used in the other aspects of the student-learning environment such as in determining groups of teachers according to some factors. In fact, a study was successfully conducted using clustering to group teachers. Further, the results were used as basis for evaluating teaching quality (Sangita et al, 2011).

Other studies involve clustering of educational aspects in the case of online learning. Studies were conducted using clustering algorithms in determining user groups and personalized intelligent tutoring. Clustering algorithm was modified by exploiting the use of minimum spanning tree. Results revealed increased performance over traditional clustering algorithms when used in online learning resources (Wu et al, 2016).

Another study was made to understand behaviors of learners in the context of online learning (Peach et al, 2019). The study made use of mathematical framework for the analysis of time-series of online learner engagement, which allows the identification of clusters of learners with similar online temporal behavior directly from the raw data without prescribing a priori subjective reference behaviors. The study revealed outliers and other significantly distinct patterns of student engagements between high-performing learners and low-performing learners.

A similar study was conducted in order to determine institutional blended learning adoption using data extracted from universities (Park et al, 2016). Latent Class Analysis was used. Results of the studies revealed four clusters out from 612 courses. The results were used as basis for developing a Learning Management System which served as a strategic tool.

The onset of the pandemic brought great impact and many changes in our society today. In the educational setting, both the teachers and students were greatly affected. The traditional learning process cannot be utilized as of late and most especially in areas under community quarantine. Therefore, other modes of learning should be adopted.

Further, it is important for learners to continue education despite the current situation. To this end, this study attempts to determine groups of learners and their readiness in accessing online and/or flexible learning. The significance of this study will help in the formulation and adoption of policies in the educational setting relevant to the current times.

II. METHODOLOGY

This study was conducted in Bohol Island State University – Clarin Campus located at Poblacion Norte, Clarin, Bohol. The Institution offers the following degree programs; Bachelor in Technology and Livelihood Education (BTLEd), Bachelor of Secondary Education (BSEd), Bachelor in Elementary Education (BEEd), Bachelor of Science in Computer Science (BSCS), Bachelor of Science in Environmental Science (BSES), Bachelor of Science in Hospitality Management (BSHM) and Bachelor of Hotel and Restaurant Service Technology (BHRST).

In order to achieve the objective of this study, clustering using K-means algorithm was conducted. Clustering is a process of grouping the data into clusters (Jain, 2010). It classifies the data instances into subsets that has the same characteristics and similarities (Celebi et al, 2013). The workflow used in this study is presented following the figure below.

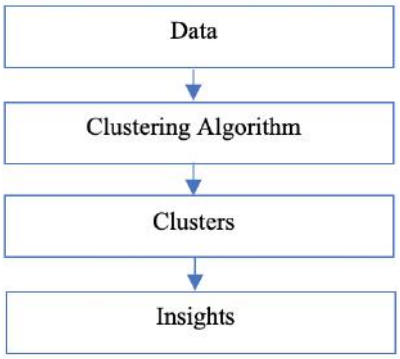

Fig. 1: Stages of the clustering pipeline

First, data was collected from the students using survey method through the use of online media such as Google Form and Facebook messaging Then, the data were recorded and stored as a dataset in a comma-separated value file. Next, data processing and cleaning stage was conducted to ensure the relevance and validity of the dataset. The table below describes the dataset.

Table 1. Dataset metadata

Data	**Description**
Degree Program	Current enrolled degree program
Address	Permanent address of the student
Gadgets Used	Device used to access the internet
Internet Access	Preferred mode of access
Signal Efficiency	Strength of the internet

Prior to the clustering task, the dataset was loaded as a pandas dataframe and null value checks and outlier detection were then conducted. After ensuring that there were no outliers and null values, each column was assigned numeric values and scaled. Then the resulting scaled values was then fed to t-distributed Stochastic Neighbor Embedding (t-SNE) algorithm. Since the dataset values are nonlinear, the t-SNE algorithm was used as a non-linear dimensionality reduction to fit and transform the scaled dataset values.

Using the transformed dataset values, clustering was then performed. We conducted experiments using different K-values ranged from 2 to 9. Using each K-value, K-means algorithm was performed and the inertia and silhouette coefficient were measured and recorded. The results were then plotted and analyzed to select the optimal K-value that best fits the given dataset using the Elbow

method. Then using the selected optimal K-value, K-means algorithm was re-run and the findings were analyzed and interpreted. The presentation of results and insights are presented in the next section.

III. RESULTS AND DISCUSSIONS

We collected and recorded the survey data. Then, the data was pre-processed and scaled to fit before feeding it to the K-means clustering algorithm. Then, experiments were conducted to select the best value for K. Results of the experiment are presented and discussed herein.

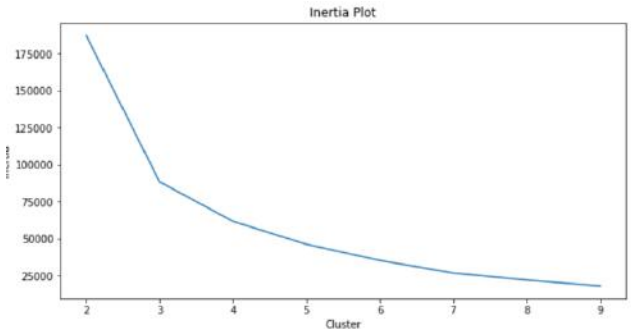

Fig. 2: Inertia Plot

The K-means algorithm clusters data by trying to separate samples into *k* groups of equal variances which optimizes a criterion known as the inertia (within-cluster sum-of-squares). Inertia can be recognized as a measure of how internally coherent clusters are. However, choosing the value for K will affect the inertia. Thus, this requires, careful observation and analysis.

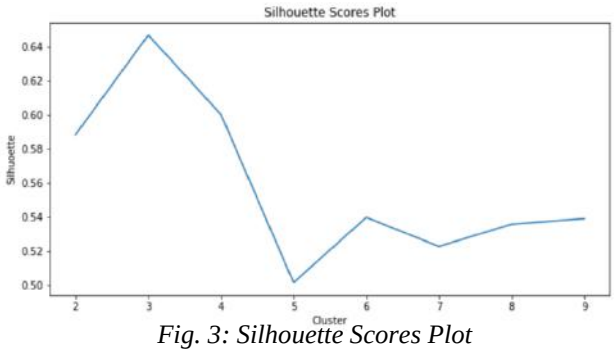

Fig. 3: Silhouette Scores Plot

Based on Fig. 2, the with-in cluster sum of squares (inertia) from K = 2 to 9 follows a downward trend. Starting at K = 2 with a value of 187355, it steadily decreased to 18122 at K = 9. Based on the analysis, the best K value is found at K =3, forming the elbow where there is sharp decline of inertia. However, using the inertia plot is not enough. Therefore, we also recorded and checked the Silhouette coefficient during the experiments. Results of the experiments are displayed in the next figure.

Silhouette score was used to evaluate the quality of clusters. This score describes how similar a sample is to its cluster as compared to samples from other clusters. This value is ranged from -1 to 1, but it is understood that the closer the score to 1, the better the clustered data points are in terms of cluster cohesion and separation.

Fig. 3 shows the results of the silhouettes scores for every value of K during the experiment. Based on the plot, the K where the silhouette score capped the highest was determined to be the best K value for the number of clusters in the dataset. The K value found in the figure is also the same K value found using the Elbow method discussed earlier. Thus, the K value of 3 was used as the number of clusters in the K-means algorithm. After clustering, the results are visualized and shown in the next figure.

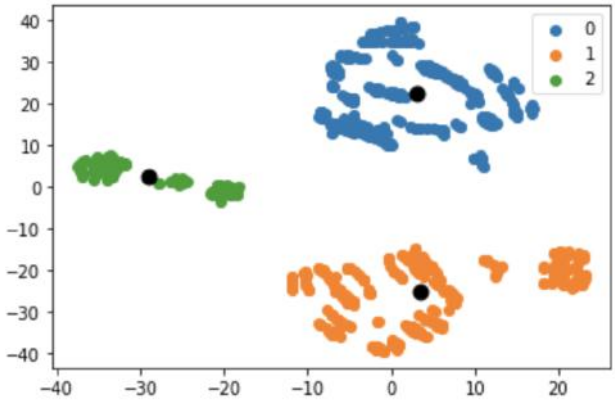

Fig. 4: Clusters Visualization

Based on the figure, the algorithm was able to distinguish three distinct groups of students. The centroid for each cluster is also plotted in the figure, which shows a healthy distance between data points within each cluster. On the other hand, the distance between each cluster are also far, thus, indicating good clusters. However, this does not describe the insights for each cluster. Thus, we inspected the data points belonging to each cluster and indicators were observed. The summary is shown in the following table.

Table 2. Cluster Descriptions

Cluster	Name
1	Moderate-signal mobile users
2	Low-signal mobile users
3	Low/moderate-signal mobile/broadband users

Cluster 1 refers to students having smartphones with access to moderate signal and prepaid data. This cluster has 306 students. Since majority of the students in this cluster own smartphones and have access to moderate prepaid data, they are most likely to have easy access to online learning resources (Fig. 5).

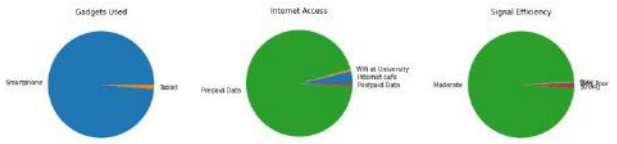

Fig. 5: Cluster 1

Students from Cluster 2 are mostly situated in areas where cell sites are available. On the other hand, Cluster 2 refers to students having smartphones with access to low-very low signal and prepaid mobile data. This cluster has 308 students located in the outskirts of the different municipalities.

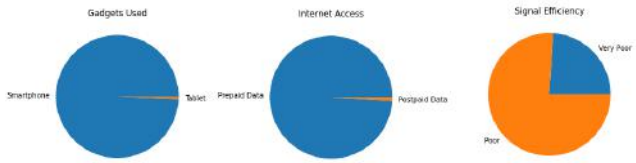

Fig. 6: Cluster 2

As can be seen in Fig. 6, students belonging to this cluster will most likely to have difficulty in accessing online learning resources. Moreover, access to learning management systems, knowledge databases and search engines are severely limited barring interventions from the Institution.

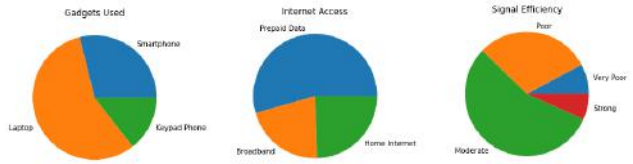

Fig. 7: Cluster 3

Lastly, Cluster 3 is a mixed group of students which own smartphones and/or laptops with access to low to high internet speeds using prepaid data or broadband home internet. This group is comprised of 90 students typically located near the municipalities where there is availability of home/wired broadband internet.

Based on the discovered clusters, three different groups of students were found to have varied usage of devices for learning, as well as access to internet, and signal efficiency. However, in order to get a much better understanding of the clusters, we also analyzed the degree program composition per cluster in an attempt to discover patterns that would be helpful in tailor-fitting the learning modalities.

Table 3 Cluster composition by degree

CLUSTER	BTLED (%)	BSED (%)	BEED (%)	CS (%)	ES (%)	HM (%)	HRST (%)
1	9	6	12	15	27	19	10
2	6	10	14	15	23	19	14
3	19	3	16	34	16	10	2

As shown in the table, for each cluster, there is no uniform distribution of students per degree program. However, Cluster 1 and 2 are predominantly consisted of BSES and BSHM students while the rest of the percentages of students from other courses are scattered across the results. Cluster 3 on the other hand, are predominantly comprised of CS students. As can be seen for each cluster, all courses cut across all programs have been represented. Policy interventions can be formulated for each learner segments which are helpful on the part of the students.

IV. CONCLUSIONS

Using K-means algorithm, we were able to successfully determine the different learner segments from the dataset according to gadgets used, internet access and signal efficiency. There were three clusters obtained; Moderate-signal mobile phone users, Poor-signal mobile phone users, and mixed group of Low-Strong mobile/broadband users. Overall, students from the different clusters are more suited for flexible learning modalities rather than online learning. This confirms that some interventions have to be formulated and implemented.

V. RECOMMENDATIONS

A Based on the conclusions, we recommend the adoption of and creation of new normal policies to cater the needs of the different learner segments. The University may consider allocating pocket WIFI and load allowances to the students. Further, flexible learning methods may be adopted for students in Cluster 1 and 3, while printed modules are recommended for students in Cluster 2. Lastly, creation of new policies such as a new grading system, student monitoring/advising mechanisms and online services are also recommended. This is to ensure

that a holistic approach to education is provided to the students to offset the effect of the pandemic.

REFERENCES

[1] Syakur, M. A., Khotimah, B. K., Rochman, E. M. S., & Satoto, B. D. (2018, April). Integration k-means clustering method and elbow method for identification of the best customer profile cluster. In IOP Conference Series: Materials Science and Engineering (Vol. 336, No. 1, p. 012017). IOP Publishing.

[2] Yuan, C., & Yang, H. (2019). Research on K-value selection method of K-means clustering algorithm. J—Multidisciplinary Scientific Journal, 2(2), 226-235.

[3] Pasina, I., Bayram, G., Labib, W., Abdelhadi, A., & Nurunnabi, M. (2019). Clustering students into groups according to their learning style. MethodsX, 6, 2189–2197. https://doi.org/10.1016/j.mex.2019.09.026.

[4] Steinbach, M., Kumar, V., & Tan, P. (2005). Cluster analysis: basic concepts and algorithms. *Introduction to data mining, 1st edn. Pearson Addison Wesley*.

[5] Rokach, L., & Maimon, O. (2005). Clustering methods. In Data mining and knowledge discovery handbook (pp. 321-352). Springer, Boston, MA.

[6] Bain K K, Firli I, And Tri S. (2016). Genetic Algorithm For Optimized Initial Centers K-Means Clustering In SMEs, Journal of Theoretical and Applied Information Technology (JATIT) 90 p 23.

[7] Jain, A., 2010. Data clustering: 50 years beyond k-means. Pattern Recognition Letters 31, 651–666.

[8] Celebi M E, Kingravi H A, & Vela, PA. (2013). A comparative study of efficient initialization methods for the k-means clustering algorithm, Expert Systems with Applications 40 p 200.

[9] Yotaman, N., Osathanunkul, K., Khoenkaw, P., & Pramokchon, P. (2020). Teaching Support System by Clustering Students According to Learning Styles. 2020 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering (ECTI DAMT & NCON). https://doi.org/10.1109/ectidamtncon48261.2020.9090729.

[10] Sangita O., Dhanamma J. (2011). An Improved K-Means Clustering Approach for Teaching Evaluation. In: Unnikrishnan S., Surve S., Bhoir D. (eds) Advances in Computing, Communication and Control. ICAC3 2011. Communications in Computer and Information Science, vol 125. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-18440-6_13.

[11] Wu, Q., Zhan, C., Wang, F. L., Wang, S., & Tang, Z. (2016). Clustering of online learning resources via minimum spanning tree. Asian Association of Open Universities Journal, 11(2), 197–215. https://doi.org/10.1108/aaouj-09-2016-0036.

[12] Peach, R.L., Yaliraki, S.N., Lefevre, D. (2019). Data-driven unsupervised clustering of online learner behaviour. npj Sci. Learn. 4, 14 (2019). https://doi.org/10.1038/s41539-019-0054-0.

[13] Zakrzewska, D. (2009) Cluster Analysis in Personalized E-Learning Systems. In: Nguyen N.T., Szczerbicki E. (eds) Intelligent Systems for Knowledge Management. Studies in Computational Intelligence, vol 252. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-04170-9_10.

[14] Park, Y., Yu, J. H., & Jo, I.-H. (2016). Clustering blended learning courses by online behavior data: A case study in a Korean higher education institute. 1–11. https://doi.org/10.1016/j.iheduc.2015.11.001.