Determining Optimal Number of K for e-Learning Groups Clustered using K-Medoid

S. Anthony Philomen Raj[1]
Research Scholar
Department of Computer Science
Periyar University
Tamil Nadu
India

Vidyaathulasiraman[2]
Assistant Professor
Department of Computer Science
Government Arts and Science College for women
Tamil Nadu
India

***Abstract*—e-Learning is appropriate when the learners are grouped and facilitated to learn according to their learning style and at their own pace. Elaborate researches have been proposed to categorize learners based on various e-learning parameters. Most of these researches have deployed the clustering principles for grouping eLearners, and in particular, they have utilized K-Medoid principle for better clustering. In the classical K-Medoid algorithm, predicting or determining the value of K is critical, two methods namely the Elbow and Silhouette methods are widely applied. In this paper, we experiment with the application of both these methods to determine the value of K for clustering eLearners in K-Medoid and prove that Silhouette method best predicts the value of K.**

Keywords—Clustering; e-learning; elbow method; k-means; k-medoid; machine learning; silhouette method

I. Introduction

The educational systems nowadays are slightly moved from Traditional Teaching Method to Electronic Teaching Method. There are a variety of tasks that can be performed in e-learning, such as; assignments, quizzes, and so on. These activities are used to assess the learner's performance. To facilitate appropriate e-learning activities by grouping users into a possible number of groups, Clustering is the Machine Learning (ML) technique that is used to group the related objects. There are numerous existing methods available for cluster analysis in the field of Data Analytics. Determining the optimal number of clusters in a data set is a fundamental problem in partitioning clustering, such as K-means clustering, which allows the user to define the number of clusters K to be generated. The possible number of clusters is rather arbitrary and is determined by the method used for measuring similarities and the parameters used for partitioning [1]. There are many clustering algorithms used for the group the similar objects in many domains such as the Medical domain, Education domain, Governance domain, etc. The clustering algorithm is the most suitable one to group users based on the learners preferred learning activities in e-learning.

The existing methods mainly focus on the majority of learning activities based on the learner's style. This could be improved further by grouping the users based on their learning activities. The main objective of this paper is to identify the optimal number of groups by using cluster validation methods. If we identify the possible number of groups of learners, we can easily enhance their learning abilities according to their preferred learning activities.

The flow of organization of work is as follows: This paper introduces a different method for identifying the value of K. Then it elaborates two major method such as Elbow and Silhouette method. The paper experiments with data using both methods. It further denotes the best method for identifying K values in K-means along with the eLearners.

A. Choosing the Optimum Number of Cluster

The optimum number of clusters obtained by using the following two methods such as *Elbow* and *Silhouette methods* [2][3].

1) Elbow method: The number of clusters (K) in the Elbow method ranges from 1 to n. We calculate WCSS (Within-Cluster Sum of Square) for each value of K. In a cluster, WCSS is the number of squared distances between each point and the centroid. The plot looks like an Elbow when we plot the WCSS with the K meaning. The WCSS value will begin to decrease as the number of clusters grows. When K = 1, the WCSS value is the highest. When we examine the graph, we can see that it will shift rapidly at a point, forming an elbow shape. The graph begins to travel almost in the same direction as the X-axis [2][3]. The optimal K value or the optimum number of clusters corresponds to this point.

Algorithm

Step 1: Compute clustering algorithm (e.g., K-means clustering) for different values of K. For instance, by varying K from 1 to N clusters.

Step 2: For each K, calculate the total Within-Cluster Sum of Square (WCSS).

$$\sum_{k=1}^{K}\sum_{i\in S_k}\sum_{j=1}^{p}(x_{ij} - \bar{x}_{kj})^2 \qquad (1)$$

Where S_k is the set of observations in the K^{th} cluster and $\bar{x}_{kj}$ is the j^{th} variable of the cluster center for the K^{th} cluster.

Step 3: Plot the curve of WCSS according to the number of clusters K.

Step 4: The location of a bend (knee) in the plot is generally considered as an indicator of the appropriate number of clusters.

2) Silhouette method: This method calculates the similarity of an object to its own cluster called *cohesion*, when compared to other clusters is called *separation*. The Silhouette value, which is a value in the range [-1, 1], is the comparison's means; a value close to 1 indicates a close relationship with objects in its cluster, while a value close to -1 indicates the opposite [2][3]. A model that produces mostly high Silhouette values from a clustered collection of data is most likely acceptable and reasonable.

Algorithm

Step 1: Choose the K from 1 to n clusters.

Step 2: For each k, calculate the Silhouette value.

Let C(i), be the cluster to which the i^{th} *data point has been allocated.*

Let |C(i)|,be the number of data points allocated to the i^{th} *data point in the cluster.*

Let a(i), indicates how well the i^{th} *data point is allocated to its cluster.*

$$a(i) = \frac{1}{|C(i)|-1}\sum_{C(i),\ i\neq j} d(i,j) \qquad (2)$$

Let b(i), be the average dissimilarity to the cluster nearest to it, but not its own

$$b(i) = min_{i\neq j}\left(\frac{1}{|C(i)|}\sum_{j\in C(i)} d(i,j)\right) \qquad (3)$$

The Silhouette Coefficient S(i) is given by:

$$S(i) = \frac{b\ (i) - a\ (i)}{max\ (a\ (i), b\ (i))} \qquad (4)$$

Step 3: Plot the curve of Silhouette value according to the number of clusters K.

Step 4: The location of the highest point is taken as the suitable number of clusters.

B. Flow Chart of Elbow and Silhouette Method

Fig. 1 portrays the flow chart of *Elbow* and *Silhouette* method to identify the optimum number of clusters.

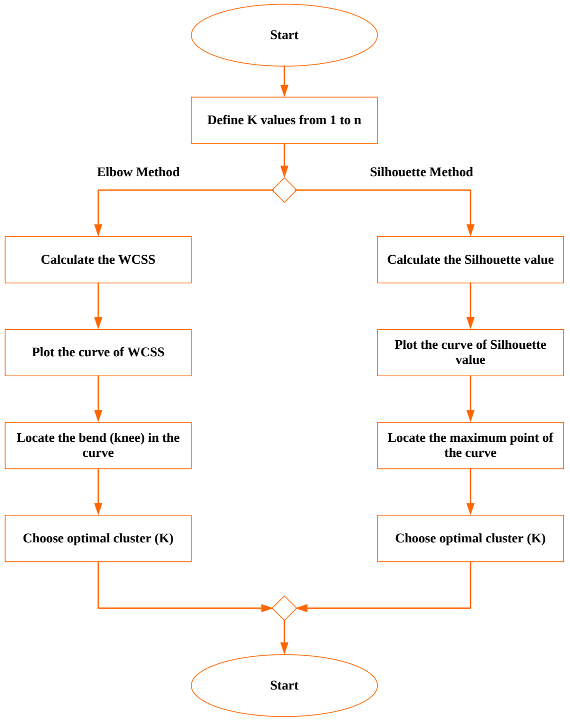

Start
Define K values from 1 to n
Elbow Method
Silhouette Method
Calculate the WCSS
Calculate the Silhouette value
Plot the curve of WCSS
Plot the curve of Silhouette value
Locate the bend (knee) in the curve
Locate the maximum point of the curve
Choose optimal cluster (K)
Choose optimal cluster (K)
Start

Fig. 1. Flow Chart of Elbow and Silhouette Method.

C. Comparison of Elbow and Silhouette Method

Table I contrasts the comparison between *Elbow* and *Silhouette* methods.

TABLE I. COMPARISON BETWEEN ELBOW AND SILHOUETTE METHOD

Elbow method	**Silhouette method**
It is more of a criterion for making decisions.	The Silhouette is a validation metric used in clustering.
The WCSS is a metric for clustering compactness, and it should be as low as possible.	This approach is useful for determining the consistency of clustering, or how well an object fits into its cluster.
The Elbow method is not computationally demanding.	The Silhouette method is the most computationally demanding.
Sometimes we don't get a clear elbow point on the plot, in such cases it's very hard to finalize the size of the cluster.	Based on the Silhouette we can identify the cluster size.

II. RELATED WORKS

In cluster analysis, especially in the field of Data Analytics determining the optimum cluster number is a major challenge. Many methods are used to find optimal clusters among Elbow and Silhouette methods which are frequently used.

H Humaira et al. [4] proposed the method of the identifying size of the clusters using the Elbow method for the K-Means Algorithm. However, this approach lacks comparison with another method called Silhouette.

Integration K-Means Clustering method and Elbow method for identification of the best Customer profile cluster were suggested by M A Syakur et al. [5]. This approach is used to determine the best number of clusters with the elbow method and it will be the default for characteristic process based on the case study.

Mohammad Khalil et al. [6] applied a Clustering pattern of engagement in Massive Open Online Courses (MOOCs): the use of learning analytics to identify student groups. This research is used to predict the learners' engagement in MOOCs by choosing optimal clusters.

Brahim Hmedna et al. [7] proposed a method, how does a learner prefer to process information in MOOCS? Using the K-Means clustering algorithm, this study discovered that the majority of learners favour active learning styles.

Towards an optimal personalization strategy, MOOCs were suggested by Alaa A.Qaffas et al. [8]. Using K-Means clustering, this approach was used to increase the retention rate and quality of learning in MOOCs.

MOOC Video Personalized Classification Based on Cluster Analysis and Process Mining was suggested by Feng Zhang et al. [9]. They suggest a process model for a group of students that represent the students' overall video-watching behaviour. Then, based on the video watching data of the students involved, it suggested using the process mining technique to mine the process model of each student cluster. Finally, the method is used to measure the difficulty and importance of a video based on a process model.

Mohammad KHALIL et al. [10] introduced a Portraying MOOCs Learners: a Clustering Experience Using Learning Analytics. The study used Clustering Analysis to group the students into suitable profiles based on their participation in a university-mandated MOOC that was also accessible to the public.

Delali Kwasi Dake et al. [11] applied a K-means clustering algorithm to analyze students' clusters for centered project-based learning. K clusters of 20 are used in this study. The findings show that the K-means clustering algorithm is good at grouping learners based on similar performance characteristics.

Analysis of University Students' Behavior Based on a Fusion K-Means Clustering Algorithm was proposed by Wenbing Chang et al. [12]. They proposed a new algorithm based on K-means and clustering by quick search and find the density peaks (K-CFSFDP), which improves data point distance and density.

Abdallah Moubayed et al. [13] proposed a model for Student Engagement Level in an e-learning Environment: Clustering Using K-means. This study recommends that students be clustered using the K-means algorithm based on 12 engagement metrics divided into two categories: interaction-related and effort-related.

Using Self-Organizing Map and Clustering to Investigate Problem-Solving Patterns in the Massive Open Online Course: An Exploratory Study proposed by Youngjin Lee et al. [14]. This study suggests that combining self-organizing map and hierarchical clustering algorithms in a clustering technique can be a useful exploratory data analysis method for MOOC instructors to classify related students based on a large number of variables and analyse their characteristics from multiple perspectives.

Prerna Joshi et al. [15] proposed a model for Prediction of Students Academic Performance Using K-Means and K-Medoids Unsupervised Machine Learning Clustering Technique. The K-mean and K-Medoids grouping algorithms were used in this study to examine students' consequence information.

Yaminee S. Patil et al. [16] suggested a technique K-means Clustering with Map Reduce Technique. This research article identified the implementation of the K-Means Clustering Algorithm over a distributed environment using Apache Hadoop.

Xin Lu et al. [17] proposed a method Improved K-means Distributed Clustering Algorithm based on Spark Parallel Computing Framework. This research identified, a density based initial clustering center selection method proposed to improve the K-means distributed clustering algorithm.

Literature review reveals that the authors have mostly focused on evaluating learner's performance by clustering techniques based on their learning styles, learning activities, and e-learning tools. The existing approaches lack in choosing an optimum number of cluster size K to group the learners. Hence, this research proposes an algorithm to identify optimum numbers of cluster size K to group the learners based on their preferred e-learning activities.

III. Research Objectives

1) To identify learners preferred e-learning activities using PCA.

2) To find correlation coefficient for selected e-learning activities using Pearson Correlation.

3) To identify an optimal number of clusters using Elbow and Silhouette method.

4) To select best method for choosing optimal cluster size.

5) To group the learners based on their preferred e-learning activities with optimal cluster size.

IV. Proposed Architecture of Choosing Optimal Cluster

The purpose of finding optimal number cluster is to extract the most possible number of groups with learner's preferred e-learning activities. The course teacher can implement those selected activities to learners based on the clusters, which helps the learners to enhance their learning abilities. Fig. 2 portrays the process to find an optimal number of groups with the help of *Elbow* and *Silhouette* method. It has been classified in the following stages:

Stage 1: Identify learners preferred e-learning activities.

Stage 2: Apply cluster validation to fix the cluster size.

Stage 3: Identify the possible number of clusters using Elbow and Silhouette method.

Stage 4: Select the method which is suitable to group the learners based on their preferred learning activities.

Stage 5: List the optimal cluster to group the learners.

Stage 1: Identify learner's preferred e-learning activities: In the first stage, identify the learners' preferred e-learning activities by using Principal Component Analysis (PCA) and compute the correlation coefficient using Pearson Correlation.

Stage 2: Apply cluster validation to fix the cluster size: In the second stage, validate cluster size by using appropriate cluster validation methods. Identifying possible number cluster size is a big challenge.

Stage 3: Identify the possible number of clusters using Elbow and Silhouette method: These two methods are used to identify the size of the cluster known as Elbow and Silhouette method. In the Elbow method by calculation value of WCSS the cluster size is fixed. Similarly, in the Silhouette by calculation of Silhouette Value the cluster size is fixed.

Stage 4: Select the method which is suitable to group the learners based on their preferred learning activities: Apply the data set in both methods and list the possible number of clusters. Let K1 be the number of cluster sizes identified by the Elbow method. Let K2 be the number of cluster sizes identified by the Silhouette method. Choose a suitable method by comparing both the cluster size (K1 and K2) and give the highest priority by choosing cluster size which one is bigger (either K1 or K2).

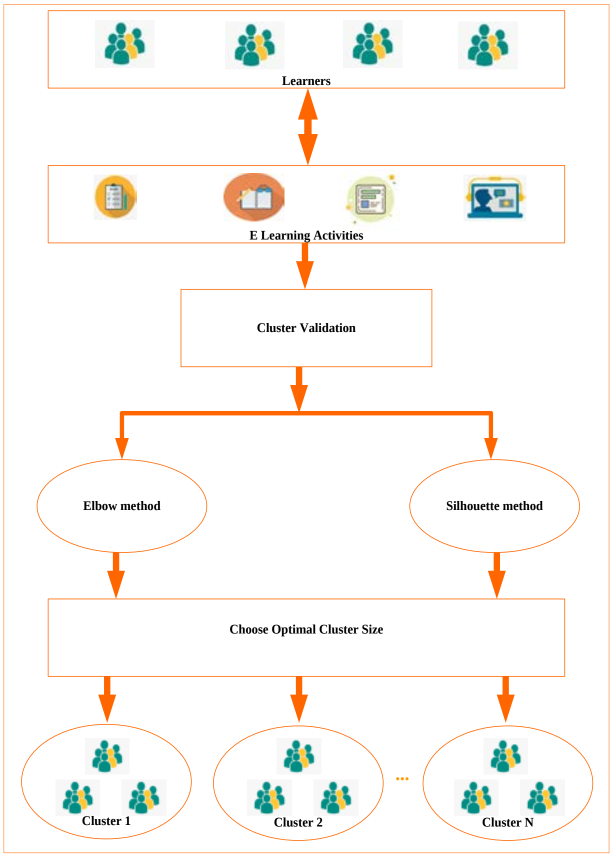


Fig. 2. Architecture of Choosing Optimal Cluster.

Stage 5: List the optimal cluster to groups the learners: Finally, the most possible cluster size (Kn) is identified by using Stage 4. Use this cluster size to identify the possible number of learners groups according to their preferred e-learning activities.

V. Proposed Algorithm of Identification of Cluster Size using Elbow and Silhouette Method

Input: *S: Data Set*

Output: *K: Number of cluster size*

1: Let S be the given set of preferred e-learning activities.

$$S = A_{ij} \tag{5}$$

2: Compute the sum of activities based on learner wise.

$$S = \sum_{i=1}^{n}\sum_{j=1}^{m} A_{ij} \tag{6}$$

3: Apply the cluster validation methods

Elbow method:

$$\sum_{k=1}^{K}\sum_{i\in S_k}\sum_{j=1}^{p}(x_{ij} - \bar{x}_{kj})^2 \tag{7}$$

Where S_k is the set of observations in the K^{th} cluster and $\bar{x}_{kj}$ is the j^{th} variable of the cluster center for the K^{th} cluster

Silhouette method:

$$S(i) = \frac{b\,(i) - a\,(i)}{max\,(a\,(i), b\,(i))} \tag{8}$$

4: Choose an appropriate method to fix cluster size

$$K = \left\{\begin{matrix} Choose\ Elbow\ Method\ if\ K1 > K2 \\ Choose\ Silhouette\ Method\ if\ K2\ > K1 \end{matrix}\right\} \tag{9}$$

5: List the optimal cluster size
6: End

VI. Results and Discussion

The optimal number of cluster size was implemented with the following e-learning activities such as Continuous Assessment (CA), Assignment, Test, Practical, Seminar, and Course Work.

Step 1: Preferred e-learning activities

Table II listed the preferred e-learning activities of 70 users and their performances. These preferred e-learning activities are identified through PCA and compared with the Pearson Correlation to find a correlation coefficient with each attribute.

Step 2: Compute the sum of activities based on learner wise.

Compute the sum of selected activities for each learner. Table III listed the sum of e-learning activities of 70 users and their performances.

Step 3: Apply the cluster validation methods

Graph 1 portrays dataset of 70 users and their performance with their preferred e-learning activities.

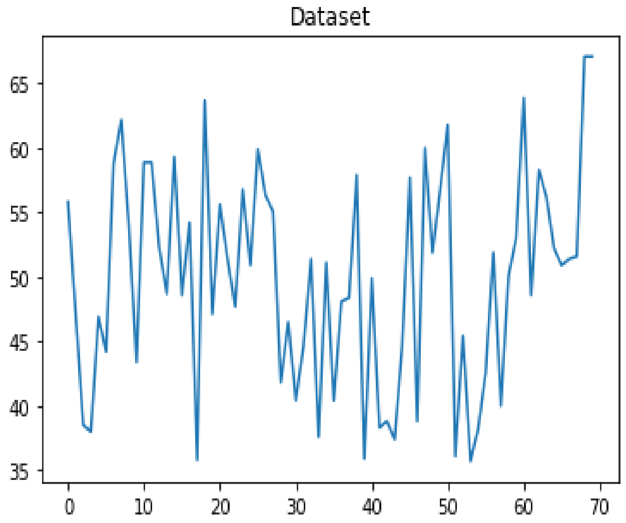
TABLE II. PREFERRED E-LEARNING ACTIVITIES OF 70 USERS. A1- CONTINUOUS ASSESSMENT (CA), A2 - ASSIGNMENT, A3 - TEST, A4 - PRACTICAL, A5 - SEMINAR, A6 - COURSE WORK.

Users/ Activities	**A1 (20)**	**A2 (5)**	**A3 (10)**	**A4 (20)**	**A5 (10)**	**A6 (10)**
BP191001L	14.5	4.5	7.8	13	8	8
BP191002L	9	3.5	8.6	12	7	7
BP191003L	3.1	2.5	4.9	10	10	8
BP191004L	4.5	3.3	6.2	8	8	8
BP191005L	7.5	4	6.4	13	8	8
BP191006L	7.5	3.8	5.9	11	8	8
BP191007L	9.3	4.2	7.3	20	10	8
BP191008L	12.3	4.5	7.4	20	10	8
BP191009L	13.2	4.3	8.6	14	7	7
BP191010L	8.3	3.3	5.8	8	10	8
BP191011L	13	3.5	4.4	20	10	8
BP191012L	11.1	3.4	6.4	20	10	8
BP191013L	12.5	3.9	7.9	14	7	7
BP191014L	11.1	3.4	9.2	11	7	7
BP191015L	8.9	4	8.4	20	10	8
BP191016L	8	3	7.6	15	7	8
BP191017L	12	4	8.2	15	7	8
BP191018L	6.5	2.9	7.4	5	7	7
BP191019L	13.2	3.6	8.9	20	10	8
BP191020L	5.8	3.2	7.1	15	8	8
BP191021L	6.5	3.8	7.3	20	10	8
BP191023L	12	4	7.4	10	10	8
BP191024L	9	4.3	6.4	10	10	8
BP191025L	13.5	4.4	7.9	15	8	8
BP191026L	9.6	3.4	6.9	15	8	8
BP191027L	10.5	3.5	7.9	20	10	8
BP191028L	7	3	8.3	20	10	8
BP191029L	12.5	4	7.6	15	8	8
BP191030L	10.2	3	6.6	8	7	7
BP191031L	8.8	2.8	6.9	10	10	8
BP191032L	6.8	2.8	6.8	8	8	8
BP191033L	6.5	2.5	8.6	13	7	7
BP191034L	13	3.3	7.1	14	7	7
BP191035L	4.2	2	6.4	11	7	7
BP191036L	11.3	4	7.8	14	7	7
BP191037L	5.3	2.5	4.6	14	7	7
BP191038L	14	3.5	8.6	8	7	7
BP191039L	10.8	3	6.6	14	7	7
BP191040L	8.8	2.9	8.2	20	10	8
BP191041L	6.5	2.5	4.9	8	7	7
BP191042L	10.9	3.3	6.7	15	7	7
BP191043L	4.5	2.5	6.3	11	7	7
BP191044L	5.5	2.6	6.7	10	7	7
BP191045L	7.5	3	4.9	8	7	7
BP191046L	8.3	3.4	7.9	11	7	7
BP191047L	7.8	3.5	8.4	20	10	8
BP191049L	6.6	2.5	5.7	10	7	7
BP191050L	10.1	3.3	8.6	20	10	8
BP191051L	11.8	3.5	9.6	13	7	7
BP191052L	8	3.5	7.3	20	10	8
BP191053L	11	4	8.8	20	10	8
BP191054L	6	2.5	5.6	8	7	7
BP191055L	9.1	3.1	6.2	13	7	7
BP191056L	6.5	2.4	4.8	8	7	7
BP191057L	6.5	2.5	7.2	8	7	7
BP191058L	9.5	3.5	7.6	8	7	7
BP191059L	12.5	4	7.4	14	7	7
BP191060L	7.3	2.3	5.4	11	7	7
BP191061L	7.8	2.5	5.8	18	8	8
BP191001	8	5	8	16	8	8
BP191002	14.9	5	8	18	9	9
BP191003	5.6	5	6	16	8	8
BP191004	9.3	5	8	18	9	9
BP191005	7.2	5	8	18	9	9
BP191006	7.2	5	8	16	8	8
BP191007	5.9	5	8	16	8	8
BP191008	6.4	5	8	16	8	8
BP191009	6.6	5	8	16	8	8
BP191010	12.1	5	10	20	10	10
BP191011	12.1	5	10	20	10	10

TABLE III. SUM OF E-LEARNING ACTIVITIES OF 70 USERS

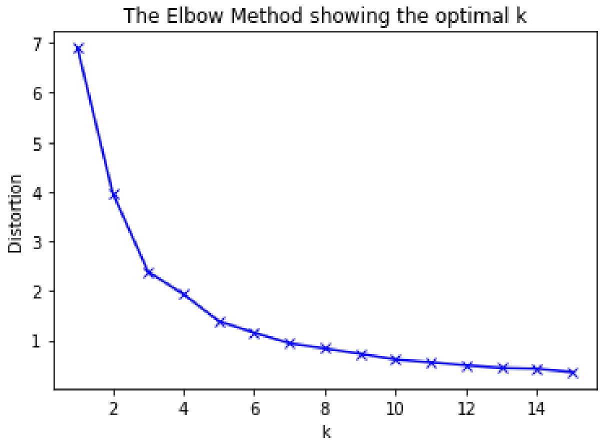
Users	**Total Score**
BP191001L	55.8
BP191002L	47.1
BP191003L	38.5
BP191004L	38
BP191005L	46.9
BP191006L	44.2
BP191007L	58.8
BP191008L	62.2
BP191009L	54.1
BP191010L	43.4
BP191011L	58.9
BP191012L	58.9

BP191013L	52.3
BP191014L	48.7
BP191015L	59.3
BP191016L	48.6
BP191017L	54.2
BP191018L	35.8
BP191019L	63.7
BP191020L	47.1
BP191021L	55.6
BP191023L	51.4
BP191024L	47.7
BP191025L	56.8
BP191026L	50.9
BP191027L	59.9
BP191028L	56.3
BP191029L	55.1
BP191030L	41.8
BP191031L	46.5
BP191032L	40.4
BP191033L	44.6
BP191034L	51.4
BP191035L	37.6
BP191036L	51.1
BP191037L	40.4
BP191038L	48.1
BP191039L	48.4
BP191040L	57.9
BP191041L	35.9
BP191042L	49.9
BP191043L	38.3
BP191044L	38.8
BP191045L	37.4
BP191046L	44.6
BP191047L	57.7
BP191049L	38.8
BP191050L	60
BP191051L	51.9
BP191052L	56.8
BP191053L	61.8
BP191054L	36.1
BP191055L	45.4
BP191056L	35.7
BP191057L	38.2
BP191058L	42.6
BP191059L	51.9
BP191060L	40
BP191061L	50.1
BP191001	53
BP191002	63.9
BP191003	48.6
BP191004	58.3
BP191005	56.2
BP191006	52.2
BP191007	50.9
BP191008	51.4
BP191009	51.6
BP191010	67.1
BP191011	67.1



Graph 1. Dataset of 70 users and their Performance.

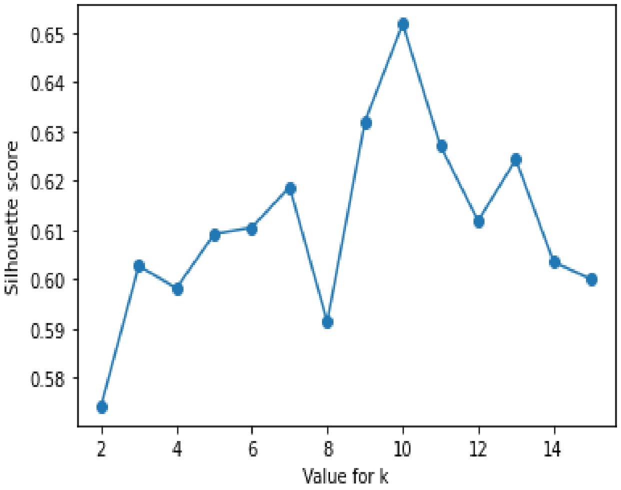
Elbow method – Graph 2 portrays optimal numbers of cluster size identified by using Elbow method. The graph shows the possible optimal number of cluster K = 5.



Graph 2. Elbow Method: Optimal Number of Cluster.

Silhouette method – Graph 3 portrays optimal numbers of cluster size identified by using Silhouette method. The graph shows the possible optimal number of cluster K = 10. The Silhouette Score for cluster size K = 2 to 15 are as follows:

Silhouette Score for 2 Clusters: 0.5742
Silhouette Score for 3 Clusters: 0.6004
Silhouette Score for 4 Clusters: 0.5982
Silhouette Score for 5 Clusters: 0.6091
Silhouette Score for 6 Clusters: 0.6034
Silhouette Score for 7 Clusters: 0.5960
Silhouette Score for 8 Clusters: 0.6094
Silhouette Score for 9 Clusters: 0.6185
Silhouette Score for 10 Clusters: 0.6518
Silhouette Score for 11 Clusters: 0.6248
Silhouette Score for 12 Clusters: 0.5981
Silhouette Score for 13 Clusters: 0.5995
Silhouette Score for 14 Clusters: 0.6061
Silhouette Score for 15 Clusters: 0.6047



Graph 3. Silhouette Method: Optimal Number of Cluster.

Step 4: Choose appropriate method to fix cluster size

From Step 3, the appropriate method for validating cluster size for given data set is Silhouette method. This can be achieved by comparing cluster size of both the methods.

Step 5: List the optimal cluster size

The most possible number of cluster size is K = 10. The possible number of learners' groups is also 10, according to their preferred e-learning activities.

The paper experiments data with two cluster validation methods such as Elbow and Silhouette method. These two methods are frequently used to validate cluster size. The cluster size return by Elbow method for the given data set is 5 (K=5). The cluster size return by Silhouette method for the given data set is 10 (K=10). Naturally when the cluster size increases, the learning abilities of the each learner is identified in depth. Based on their learning abilities, we can optimize and predict e-learning activities for each user. Finally this paper concludes that Silhouette method is the optimal method for validating cluster size for the given data set.

VII. Conclusion and Future Work

In this paper, we have illustrated that determining optimal value for K to cluster eLearners using K-Medoid algorithm is essential. Further, we have experimented the Elbow and Silhouette methods to compute the optimal value for K. To decide the suitable method among these two, sufficient experiments were conducted, and the results of the experiments were investigated carefully. The results were indicative that the Silhouette method best suits to fix the optimal value for K. This paper is focused in estimating K value for K-Medoid based clustering of eLearners, computing the optimum value for k, number of clusters to be formed can also be done in other clustering methods which are applied in eLearners groupification, which is suggested by the authors as a future enhancement.

References

[1] Channamma Patil and Ishwar Baidari, "Estimating the Optimal Number of Clusters *k* in a Dataset Using Data Depth", Data Science and Engineering, vol 4, pp. 132–140, June 2019.

[2] Sukavanan Nanjundan, Shreeviknesh Sankaran, C.R. Arjun and G. Paavai Anand, "Identifying the number of clusters for K-Means: A hypersphere density based approach", International Conference on Computers, Communication and Signal Processing, December 2019.

[3] Chunhui Yuan and Haitao Yang, "Research on K-Value Selection Method of K-Means Clustering Algorithm", Multidisciplinary Scientific Journal, vol 2, pp. 226-235, June 2019.

[4] H Humaira and R Rasyidah, "Determining The Appropiate Cluster Number Using Elbow Method for K-Means Algorithm", WMA-2, January 2020.

[5] M A Syakur, B K Khotimah, E M S Rochman and B D Satoto, "Integration K-Means Clustering Method and Elbow Method For Identification of The Best Customer Profile Cluster". IOP Conference Series: Materials Science and Engineering, vol 336, 2018.

[6] Mohammad Khalil and Martin Ebner, "Clustering patterns of engagement in Massive Open Online Courses (MOOCs): the use of learning analytics to reveal student categories", Journal of Computing in Higher Education, vol 29, pp.114–132, October 2016.

[7] Brahim Hmedna, Ali El Mezouary and OmarBaz, "How Does Learners' Prefer to Process Information in MOOCs? A Data-driven Study", Procedia Computer Science, vol 148, pp. 371-379, 2019.

[8] Alaa A.Qaffas, Kaouther Kaabi, Rustam Shadiev and Fathi Essalmi, "Towards an optimal personalization strategy in MOOCs", Smart Learning Environments, pp. 7-14, April 2020.

[9] Feng Zhang, Di Liu and Cong Liu, "MOOC Video Personalized Classification Based on Cluster Analysis and Process Mining", Sustainability, vol 12, 2020.

[10] Mohammad KHALIL, Christian KASTL and Martin EBNER, "Portraying MOOCs Learners: a Clustering Experience Using Learning Analytics", Proceedings of the European MOOC Stakeholder Summit, 2016.

[11] Delali Kwasi Dake and Esther Gyimah, "Using K-Means to Determine Learner Typologies for Project-based Learning: A Case Study of the University of Education, Winneba", International Journal of Computer Applications (0975 – 8887), vol 178, pp. 29-34, August 2019.

[12] Wenbing Chang, Xinpeng Ji, Yinglai Liu, Yiyong Xiao, Bang Chen, Houxiang Liu and Shenghan Zhou, "Analysis of University Students' Behavior Based on a Fusion K-Means Clustering Algorithm", Applied Sciences, vol 10, September 2020.

[13] Abdallah Moubayed, Mohammadnoor Injadat, Abdallah Shami and Hanan Lutfiyya, "Student Engagement Level in an eLearning Environment: Clustering Using K-means", American Journal of Distance Education, vol 34, pp. 137-156, 2020.

[14] Youngjin Lee, "Using Self-Organizing Map and Clustering to Investigate Problem-Solving Patterns in the Massive Open Online

Course: An Exploratory Study", Journal of Educational Computing Research, vol 57, pp. 471-490, January 2018.

[15] Prerna Joshi and Pritesh Jain, "Prediction of Students Academic Performance Using K-Means and K-Medoids Unsupervised Machine Learning Clustering Technique", International Journal of Scientific Development and Research (IJSDR), vol 3, pp. 162-171, June 2018.

[16] Yaminee S. Patil and M. B. Vaidya, "K-means Clustering with MapReduce Technique", International Journal of Advanced Research in Computer and Communication Engineering, vol 4, pp. 349-352, November 2015.

[17] Xin Lu, Huanghuang Lu, Jiao Yuan and Xun Wang, "An Improved K-means Distributed Clustering Algorithm Based on Spark Parallel Computing Framework", Journal of Physics: Conference Series, vol 1616, 2020.