Journal of Data Analysis and Information Processing, 2020, 8, 171-182
https://www.scirp.org/journal/jdaip
ISSN Online: 2327-7203
ISSN Print: 2327-7211

Clustering Approach for Analyzing the Student's Efficiency and Performance Based on Data

Tallal Omar[1], Abdullah Alzahrani[2], Mohamed Zohdy[2]

[1]Department of Computer Science and Engineering, Oakland University (OU), Rochester, MI, USA
[2]Department of Electrical and Computer Engineering, Oakland University (OU), Rochester, MI, USA
Email: tallalomar@oakland.edu, alzahrani2@oakland.edu, zohdyma@oakland.edu

How to cite this paper: Omar, T., Alzahrani, A. and Zohdy, M. (2020) Clustering Approach for Analyzing the Student's Efficiency and Performance Based on Data. *Journal of Data Analysis and Information Processing*, **8**, 171-182.
https://doi.org/10.4236/jdaip.2020.83010

Received: June 17, 2020
Accepted: August 15, 2020
Published: August 18, 2020

Abstract

The academic community is currently confronting some challenges in terms of analyzing and evaluating the progress of a student's academic performance. In the real world, classifying the performance of the students is a scientifically challenging task. Recently, some studies apply cluster analysis for evaluating the students' results and utilize statistical techniques to part their score in regard to student's performance. This approach, however, is not efficient. In this study, we combine two techniques, namely, k-mean and elbow clustering algorithm to evaluate the student's performance. Based on this combination, the results of performance will be more accurate in analyzing and evaluating the progress of the student's performance. In this study, the methodology has been implemented to define the diverse fascinating model taking the student test scores.

Keywords

K-Means Technique, Elbow Technique, Clustering Technique, Data Mining, Academic Performance

1. Introduction

Clustering is one of the most significant techniques in data mining that explores data sets [1]. In the last decades, several clustering approaches with better performance have been applied to a broad range of applications [2]. The clustering techniques could be divided into many methods such as partitioning and hierarchical clustering, [3] situation awareness in online learning [4], density [5], model building, and [6] others. Recently, the K-means technique has been successfully applied [7]. K-means is a partitioning-based technique that splits data

into sets based on their proximity to each other. The original K-means technique [8] applied Euclidean distance to duplicate an estimated pattern of resemblance between data points: a method that did not fit many applications. Research conducted by Unnati Raval *et al.* described the advantages and disadvantages of the original K-means algorithm and proposed techniques to improve the accuracy of clustering and reduce computational time [1]. Bo Yang *et al.* presented a nonlinear function that combined a dimensionality reduction (DR) and K-means partitioning to cluster latent data representations. The results jointly demonstrated performance improvement and advantages of joining the two tasks [2]. Dhendra Marutho *et al.* combined the elbow method and traditional K-means to determine the optimal number of clusters in the K-means algorithm on news headline data. The researchers then applied the purity method to evaluate news title clustering as an internal evaluation. The results then produced the best number of clusters gained using the Elbow method [3]. Jasser *et al.* applied the K-means algorithm to statistically cluster students online learning course grades and GPA, grouping them based on similarities of their history. Where similar student history and activity led to similar performance, the results demonstrated that each group of students assigned different homework to retake, quizzes and dedicated more learning effort to the assigned chapters [4].

The Elbow method is one of the most common approaches in identifying the optimal value of K-clusters in a data set. Syakur *et al.* took advantage of this approach by combining the traditional K-means algorithm with the Elbow method to enable an optimal way of counting clusters of segmented performance profiles [9]. Specifically, the Elbow approach consists of plotting the demonstrated difference as a function of the number of clusters, then selecting the elbow of the curve as the optimal number of clusters to apply. Purnima *et al.* streamlined the clustering process by creating groups of sensor nodes that in turn, reduced the routing computations of the smaller routing data size [10]. Computing the mean score for all clusters was achieved by running the K-means clustering on the set of data for a K-domain of values from 1 - 10, and then for every value of K [11].

The primary objective of this research was to introduce a new clustering model which combines the K-means algorithm with four functionalities: the Elbow method, scaling, and normalization/standardization on a dataset. For computer science students from Oakland University, the dataset was based on the following classification: course name, course grade, cumulative GPA, and number learning segments requiring more attention. Each of these classifications was referred to as the student's performance. After clustering students in groups, an improvement plan was structured for each group of students emphasizing the areas where each student was not performing well, recommending chapters for review, homework to retake, and topics to dedicate more attention to.

The rest of this paper is organized as follows: Section Two the methodology describes the broad theoretical that is applied in the study. In Section Three includes results and discussions and finally, Section Four concludes the paper with some future work suggestions.

2. Methodology

In the proposed system, optimizing the number of clusters of K-Mean approaches was the main intent of this research study considering the Elbow approach to define the number of clusters during the processes of evaluation. The proposed method could be described by the following flowchart in Figure 1.

One semester of study datasets for computer science students were selected to analyze and make predictions about future student performance. The data sets were passed through four stages of processing. The first step was converting grade/course data type from a string of data to just one number. The next step was scaling that applies standardization/normalization to dataset features that have a magnitude variance larger than others. The third step was applying the Elbow method to the dataset to define the optimal number of K-clusters. The fourth step was running the K-means algorithm on the dataset to partition or group students in clusters based on their performance while the SSE (the Error Sum of Squares, the sum of the squared differences between each observation and its group's mean) was calculated and recorded to determine the number of clusters.

2.1. Preprocessing

Scaling of datasets is a method of standardizing/normalizing the range of independent variables and a common requirement for many machine learning estimators implemented in Scikit-Learn (for the Python programming language) and it

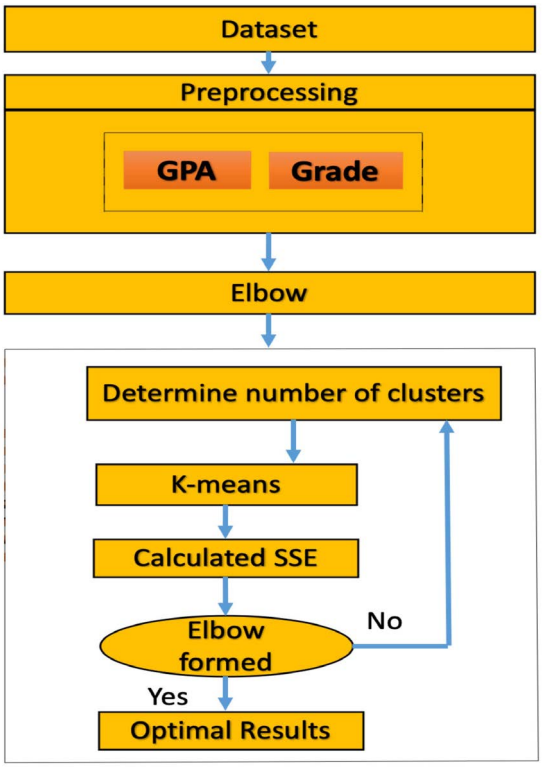

Figure 1. Process flowchart for the proposed system.

might dominate the objective function and make the estimator unable to learn from other features as correctly as expected [11]. Therefore, the estimator might behave badly if the individual features do not more-or-less resemble standard normally distributed data, Gaussian processing with a mean of zero and no unit variance. In practice, the shape of the distribution and centralize the data is ignored by removing the mean value of each feature, then scale it by dividing non-constant features by their standard deviation. In this process, the standard deviation is represented as follows:

$$z = \frac{x - \mu}{\sigma} \tag{1}$$

This formulation is a specification that has z as the standardized/normalized value, x is the raw value of the data point, μ is the population mean, and σ is the population standard divisor for the dataset.

2.2. K-Means Method

The K-means technique is a type of partitioning/clustering method that was first established by J. B. MacQueen [12]. This technique has been generally applied in data mining and pattern recognition. It also has been defined as one of the simplest data mining partitioning/clustering approaches that implements the Euclidean distance function. The main purpose of the K-means technique has been reducing the cluster performance index, the error sum of squares, and the error criterion that are the backbone of discovery of the optimal value of k divisions to meet a specific criterion are the functional objectives of this method. Some of the advantages of the K-means are brevity, efficiency, and swiftness [13]. Yet, this method relies mostly on initial data points and the variance in selecting initial samples that usually are directed to various outcomes [14]. What is more, the K-means method constantly utilizes a gradient technique based on the objective function to get a peak value. The trend seeking function in this gradient technique is primarily observed in the computation process in which the entire process will readily sink to the lowest point when the initial cluster focal point may not be appropriate that in turn, causes energy reductions [15]. The standard algorithm is described by the Hartigan-Wong algorithm [16] that defines the total within-cluster variation as the sum of the squared distances between items and the corresponding centroid.

$$W\left(C_k\right) = \sum_{x_i \in C_k} \left(x_i - \mu_k\right)^2 \tag{2}$$

This formulation is a specification in which $W\left(C_k\right)$ represents the within-cluster total, x_i is a data point for a cluster, C_k indicates a cluster for each data point, and μ_k defines the mean value of the points that is assigned to the cluster C_k. Therefore, the sum of total within-clusters of the sum of squares measures compactness (TW) as follows

$$\mathrm{TW} = \sum_{K=1}^{K} W\left(C_k\right) = \sum_{k=1}^{k} \sum_{x_i \in C_k} \left(x_i - \mu_k\right)^2 \tag{3}$$

2.3. Elbow Method

The Elbow approach is a technique which looks at the percentages of variance illustrated as a function of the optimal number of clusters in the K-means [17]. This approach exists based on the idea that a number of clusters have to be selected so that the means of one more cluster does not provide marginally better modeling of the data. The percentage of variance demonstrated by the clusters is plotted versus the optimal number of clusters. The first clusters will insert numerous amounts of information but at a certain point, the marginal number of clusters attained will fall significantly and provide an angle in the graph [18]. The proper value of "*k*", which is the number of clusters selected at this point, is called the Elbow criterion. The main concept could be described as beginning with $k = 2$, keep raising it by one point, computing the cluster and the cost that comes with the training [19]. At some value for *k*, the cost will substantially decline, and after that point, the cost rises when you raise the *k* point higher. At the point where the cost decline changes to a cost increase is the *k* value you look for to be the elbow. So, if the value of the cluster $k = 3$ to $k = 4$ and has a cost decline, then goes then from $k = 4$ to $k = 5$ and gives a sharp cost increase, there is an elbow at point $k = 4$ which means the perfect cluster *k* is $k = 4$ [20].

The Elbow method is described in Equation (4) as the within-groups sum of squares (WSS), where the squared average distance of all the data points (the means of each of the individual groups or group means) for a cluster is a distance that is statistically measured from the group means to the same cluster centroid [21].

$$\mathrm{WSS} = \sum_{i=1}^{m} \left(x_i - c_i \right)^2 \tag{4}$$

The combination of both the K-means and Elbow method can locate the value of *k* at the optimal cluster to determine *k* as the number of clusters formed. The Elbow method is used to choose the best number of *k* clusters for grouping data within the K-means technique. The Elbow method can be expressed by the sum of the squares error [21] as follows.

$$\mathrm{SSE} = \sum_{k-1}^{k} \sum_{x_i \in S_k} \left\| x_i - c_k \right\|_2^2 \tag{5}$$

This formulation is a specification with *k* that is equal to many clusters that formed *C*, which is the i^{th} cluster, with *X*, representing the data given at each cluster.

3. Results and Discussions

In this study, the previously described approach has been followed based upon a computer generated algorithm to group students in multiple clusters based on their performance.

From the proposed system, the preprocessing stage has been used to convert the data type from string to numeric value, as shown in **Table 1** and **Table 2**.

Table 1. Sample raw dataset profiling for students.

Course Number	*Grade*	*GPA*
CSI-1420	B+	1.54
CSI-2310	B–	1.54
CSI-2300	A	1.65
CSI-2999	A	1.65
CSI-3660	A–	1.87
CSI-2310	B–	1.87

Table 2. Converting dataset to numeric.

Course Number	*Grade*	*GPA*
CSI-1420	3.6	1.54
CSI-2310	3	1.54
CSI-2300	4	1.65
CSI-2999	4	1.65
CSI-3660	3.8	1.87
CSI-2310	3.2	1.87
CSI-2999	4	2
CSI-2999	4	2.1
CSI-2310	2.8	2.33
CSI-2999	2.9	2.36

Figure 2 displays the raw data set with no consideration given for clustering, as shown in Table 2.

In Figure 3, the data set is divided into three clusters. The clustering algorithms are affected by the data feature such that the grade/course feature is dominating the GPA.

The scaling technique is structured to standardize/normalize the data sets, which is a common procedure followed by many machine learning estimators that use Scikit-Learn. If one feature of the dataset has a value that is larger than others, it might dominate the objective function and misguide the estimator to learn from other features correctly as expected. So, if the data features do not look like normally distributed data, then the estimator may behave badly. For example, if it is assumed that there are two features of a person such as weight in pounds (lbs.) and height in feet (ft.), and there is a desire to predict whether a person needs an "S" or "L" size shirt based upon these two factors, the following formula can be used by taking the sum of weight and height to determine the best fit. To clarify, if it is assumed that there are two people, one in a cluster such that one person in cluster X = (175 lbs. + 5.9 ft.) in size "L", and another person in cluster Y = (115 lbs. + 5.2 ft.) in size "S"; if a third person in cluster Z = (140 lbs. + 6.1 ft.), then the previously described method will classify cluster Z in the

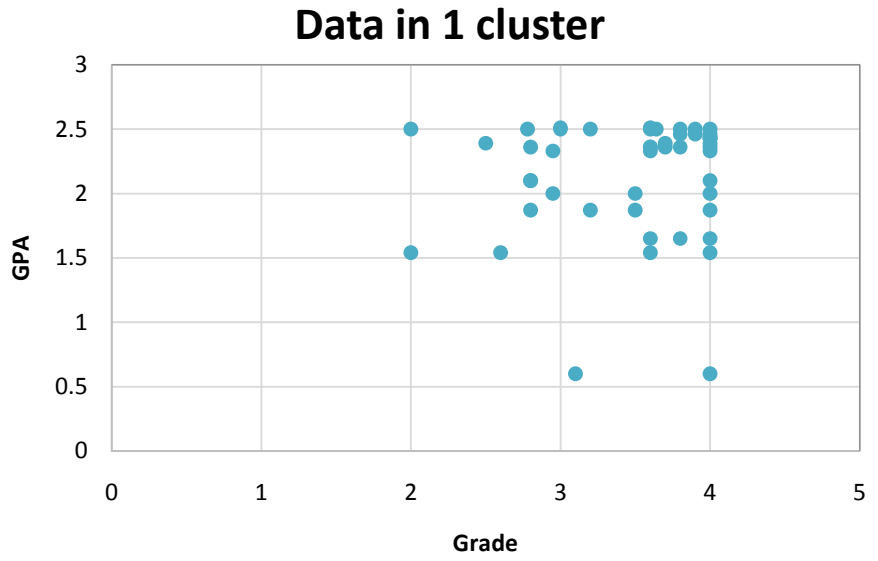

Figure 2. Raw dataset of one cluster.

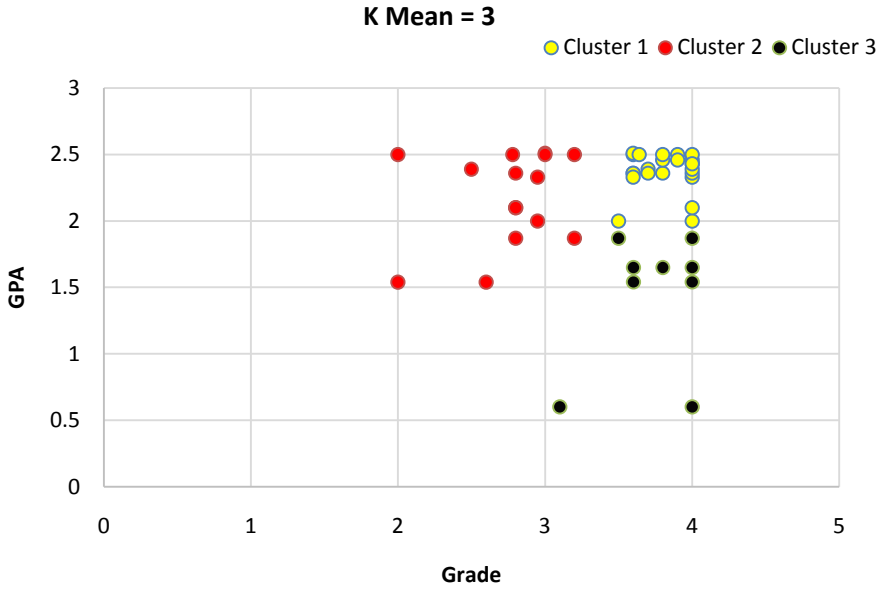

Figure 3. The K-mean applied at $k = 3$.

cluster nearest to cluster X or cluster Y. If the features are not scaled the height will not affect the clustering, and Z will be allotted in the cluster size "S". From Figure 4, WSS has been presented both before and after scaling the dataset.

From Figure 4, the increase of the score for WSS can be seen after scaling occurs at each iteration. In Figure 5, the elbow method has been implemented on the data set before scaling where the elbow curve starts roughly at WSS = 13 at $K = 2$.

In Figure 6 the elbow method has been implemented on the data set after scaling, where the elbow curve starts roughly at WSS = 35 at $K = 3$.

In Figure 6 the Elbow method is exactly located at $k = 3$ where k is the optimal number of clusters. In Figure 5 the Elbow method curve is not clear enough

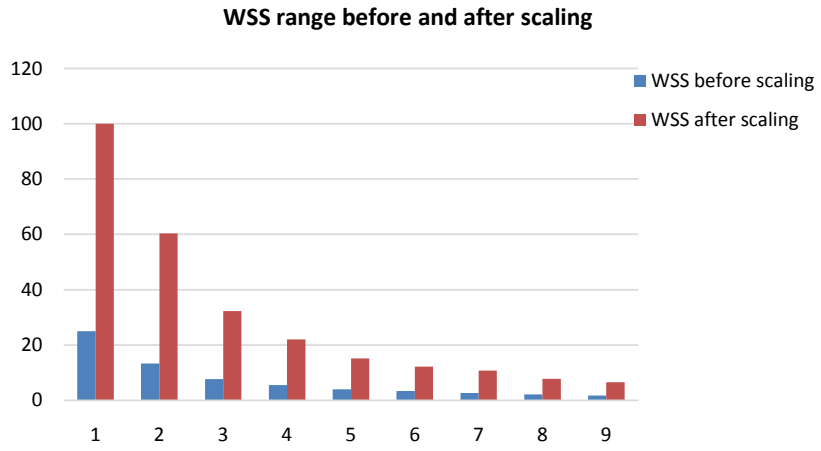

Figure 4. The WSS before and after scaling data.

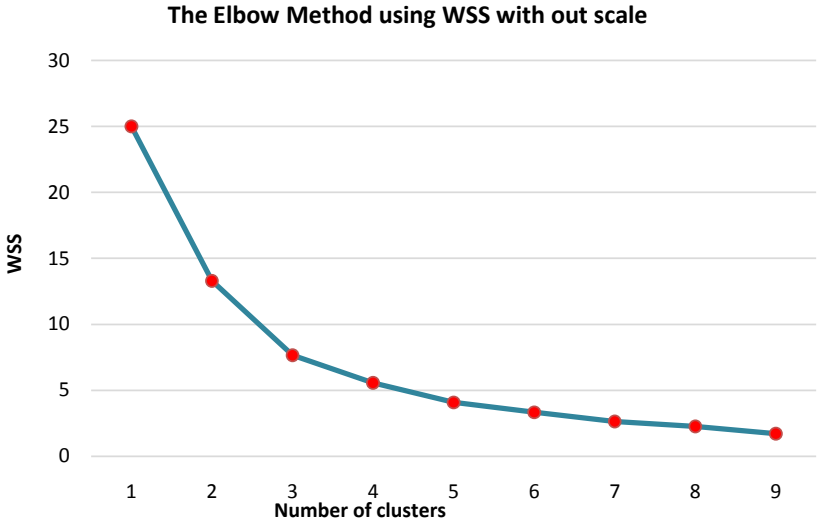

Figure 5. The Elbow method before scaling data.

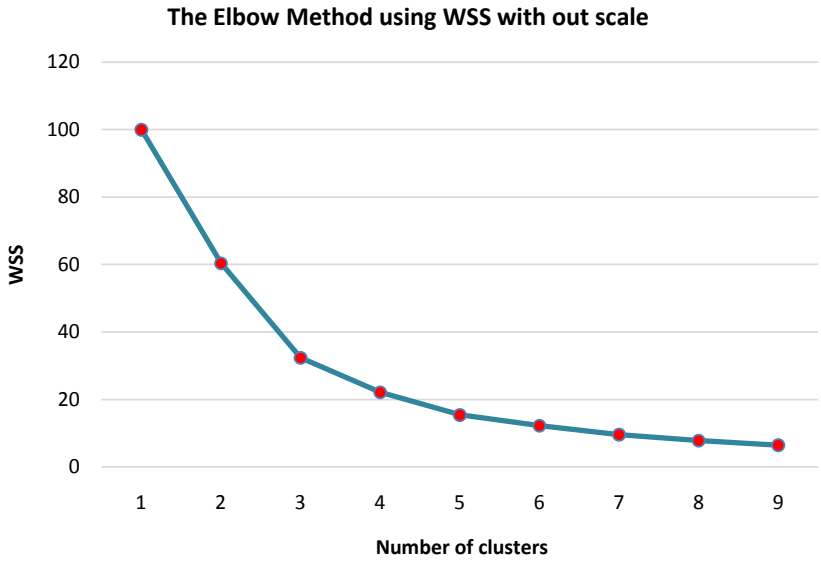

Figure 6. The Elbow method after scaling the data.

as expected at $k = 2$, because the data set features contain different ranges of values, so it is important to scale the values of the features to the same range to get more accurate results from the Elbow method. In this study, a small data set was considered; if a large scale of data set had been considered, then the differences could be noticed in the elbow between the scaled and unscaled data sets. Applying this model has demonstrated that the optimal number of clusters for any given dataset can be achieved.

4. Conclusion

This research paper provided a description of a simple and efficient method to help college students that are under-performing, are very close to falling under their university's minimum academic standards, or are on academic probation as indicated by their GPA. Many universities have 2.0 GPA minimum standards and will issue a warning to the students when a student's GPA falls below that standard, providing a probationary period of one semester to raise their GPA. Students on probation are usually not able to participate in college activities including working in school or receiving scholarships. Those students will find themselves out-of-school in one semester if they do not improve their GPA. Previous research combining the K-means clustering algorithm with different methodologies provided the inspiration in creating an efficient procedure for providing failing students with targeted suggestions for significantly raising their performance above the 2.0 minimum standards. This was a procedure that combined the K-means algorithm with four functionalities: the Elbow method, scaling, normalization, and standardization. First, a dataset was selected containing GPAs and grades per course for computer science students from Oakland University. The dataset was setup first so that students were clustered into groups based on performance similarity for GPA and Grades/course where the number of clusters k were provided to the K-means algorithm as input. Next, the Elbow method was applied to the same data set where the Elbow method defined the optimal number of clusters for the data set. Following that, the K-mean algorithm was combined with Elbow and scaled data where the data sets were passed through four stages of processing: 1) converting Grade/Course data type from string to a number; 2) scaling that applied both standardization and normalization to the dataset feature that had a different range to improve clustering accuracy; 3) applying the elbow method to the dataset to define the optimal number of clusters K to group students in clusters based on their performance while SSE (Sum Square Error) was calculated and recorded to determine the number of clusters; and 4) comparing the three scenarios, where the numerical result revealed that the third scenario approach where the data was scaled before combining k-means with Elbow method was accurate and efficient in optimally clustering students based on their academic results. This approach clearly demonstrated the advantage of scaling the data before combining the K-means algorithm with the Elbow method. After clustering students in groups, an improve-

ment plan was structured for each group of students based on their performance, focusing on the areas where the student was not performing well, suggesting chapters for review, homework to retake, and topics to dedicate more focus.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Raval, U.R. and Jani, C. (2016) Implementing & Improvisation of K-Means Clustering Algorithm. *International Journal of Computer Science and Mobile Computing*, **5**, 191-203.

[2] Yang, B., Fu, X., Sidiropoulos, N.D. and Hong, M. (2017) Towards K-Means-Friendly Spaces: Simultaneous Deep Learning and Clustering. *Proceedings of the* 34*th International Conference on Machine Learning-Volume* 70, Sydney, August 2017, 3861-3870.

[3] Marutho, D., Handaka, S.H. and Wijaya, E. (2018, September) The Determination of Cluster Number at K-Mean Using Elbow Method and Purity Evaluation on Headline News. 2018 *International Seminar on Application for Technology of Information and Communication*, Semarang, 21-22 September 2018, 533-538. https://doi.org/10.1109/ISEMANTIC.2018.8549751

[4] Jasser, J., Ming, H. and Zohdy, M.A. (2017) Situation-Awareness in Action: An Intelligent Online Learning Platform (IOLP). In: Kurosu, M., Ed., *Human-Computer Interaction*, *Interaction Contexts*, *HCI* 2017, *Lecture Notes in Computer Science*, Vol. 10272, Springer, Cham, 319-330. https://doi.org/10.1007/978-3-319-58077-7_25

[5] Qin, X., Ting, K.M., Zhu, Y. and Lee, V.C. (2019) Nearest-Neighbour-Induced Isolation Similarity and Its Impact on Density-Based Clustering. *Proceedings of the AAAI Conference on Artificial Intelligence*, **33**, 4755-4762. https://doi.org/10.1609/aaai.v33i01.33014755

[6] Gao, C., Sun, H., Wang, T., Tang, M., Bohnen, N.I., *et al.* (2018) Model-Based and Model-Free Machine Learning Techniques for Diagnostic Prediction and Classification of Clinical Outcomes in Parkinson's Disease. *Scientific Reports*, **8**, Article No. 7129. https://doi.org/10.1038/s41598-018-24783-4

[7] Sharma, M., Purohit, G.N. and Mukherjee, S. (2018) Information Retrieves from Brain MRI Images for Tumor Detection Using Hybrid Technique K-Means and Artificial Neural Network (KMANN). In: Perez, G., Mishra, K., Tiwari, S. and Trivedi, M., Eds., *Networking Communication and Data Knowledge Engineering*, *Lecture Notes on Data Engineering and Communications Technologies*, Springer, Singapore, 145-157. https://doi.org/10.1038/s41598-018-24783-4

[8] Kumar, J. and Vashistha, R. (2017) Estimation of Inter-Centroid Distance Quality in Data Clustering Problem Using Hybridized K-Means Algorithm. 2017 2*nd International Conference on Electrical*, *Computer and Communication Technologies* (*ICECCT*), Coimbatore, 22-24 February 2017, 1-7. https://doi.org/10.1109/ICECCT.2017.8117896

[9] Syakur, M.A., Khotimah, B.K., Rochman, E.M.S. and Satoto, B.D. (2018) Integration K-Means Clustering Method and Elbow Method for Identification of the Best Cus-

tomer Profile Cluster. *IOP Conference Series*: *Materials Science and Engineering*, **336**, Article ID: 012017. https://doi.org/10.1088/1757-899X/336/1/012017

[10] Bholowalia, P. and Kumar, A. (2014) EBK-Means: A Clustering Technique Based on Elbow Method and K-Means in WSN. *International Journal of Computer Applications*, **105**, 17-24.

[11] Huang, G.X. and Lin, D. (2019, October) Clustering Analysis and Visualization of Terrorist Attack Data. *Proceedings of the* 2019 *International Conference on Video, Signal and Image Processing*, Wuhan, October 2019, 136. https://doi.org/10.1145/3369318.3369343

[12] Tang, J., Wang, D., Zhang, Z., He, L., Xin, J. and Xu, Y. (2017) Weed Identification Based on K-Means Feature Learning Combined with Convolutional Neural Network. *Computers and Electronics in Agricultur*e, **135**, 63-70. https://doi.org/10.1016/j.compag.2017.01.001

[13] Macqueen, J.B. (1965) On the Asymptotic Behavior of K-Means. University of California, Los Angeles.

[14] Jamadi, N.A., Siraj, M.M., Din, M.M., Mammy, H.K. and Ithnin, N. (2018) Privacy Preserving Data Mining Based on Geometrical Data Transformation Method (GDTM) and K-Means Clustering Algorithm. *International Journal of Innovative Computing*, **8**, 1-7. https://doi.org/10.11113/ijic.v8n2.174

[15] Qiu, Q., Zhang, Q. and Guo, K. (2014) Grey Kmeans Algorithm and Its Application to the Analysis of Regional Competitive Ability. 2014 *IEEE 7th Joint International Information Technology and Artificial Intelligence Conference*, Chongqing, 20-21 December 2014, 249-253. https://doi.org/10.1109/ITAIC.2014.7065044

[16] Yedla, M., Rao, S.S., Pathakota and Srinivasa, T. (2010) Enhancing K-Means Clustering Algorithm with Improved Initial Center. *International Journal of Computer Science and Information Technologies*, **1**, 121-125.

[17] Amaral, G.J., Dore, L.H., Lessa, R.P. and Stosic, B. (2010) *K*-Means Algorithm in Statistical Shape Analysis. *Communications in Statistics-Simulation and Computation*, **39**, 1016-1026. https://doi.org/10.1080/03610911003765777

[18] Kodinariya, T.M. and Makwana, P.R. (2013) Review on Determining Number of Cluster in K-Means Clustering. *International Journal*, **1**, 90-95.

[19] Ghayekhloo, M., Ghofrani, M., Menhaj, M.B. and Azimi, R. (2015) A Novel Clustering Approach for Short-Term Solar Radiation Forecasting. *Solar Energy*, **122**, 1371-1383. https://doi.org/10.1016/j.solener.2015.10.053

[20] Sujatha, S. and Sona, A.S. (2013) New Fast K-Means Clustering Algorithm Using Modified Centroid Selection Method. *International Journal of Engineering*, **2**, 1-4.

[21] Selvida, D., Zarlis, M. and Situmorang, Z. (2020) Analysis of the Effect Early Cluster Centre Points on the Combination of K-Means Algorithms and Sum of Squared Error on K Centroid. *IOP Conference Series*: *Materials Science and Engineering*, **725**, Article ID: 012089. https://doi.org/10.1088/1757-899X/725/1/012089

Nomenclatures

- z: Standardized value.
- x: Raw value of the data point.
- μ: Population mean.
- σ: Population standard division for dataset.
- k:Number of clusters.
- $W\left(C_k\right)$ Total within-cluster.
- x_i : Data point for cluster.
- C_k : Cluster for each data points.
- WSS: Within Sum of Squares.
- SEE: Sum of Squared Error.
- μ_k : Mean value of the points that is assigned to the cluster C_k .
- TW: Sum of total within-clusters of the sum of squares measures compactness.