A-4 Paper size, 2,5 cm up, down, right, left
All text should be written in Times New Roman and 12pt

send your paper WORD format, NOT pdf

Title must be in capital letters, Times New Roman, 12pt, BOLD

STANDARD DEVIATION-BASED CENTROID INITIALIZATION FOR K-MEANS

title, name, surname

Dr. Ali SENOL

. Engineering Faculty, Tarsus University FACULTY AND UNICERSITY NAME

alisenol@tarsus.edu.tr, Whatsapp: +903456543222 email and whatsapp number

ORCID NO: 0000-0003-0364-2837 orcid

ABSTRACT

One of the most important problems related to k-means clustering algorithm is the random selection of initial centroids which affects the quality of resultant clusters directly. If the initial centroids are not optimal, the quality of produced clusters reduces. Although various clustering approaches that try to overcome the mentioned issue have been proposed, there is no global solution to this problem. In this study, we propose a standard deviation-based initial centroid determination approach for k-means that is named SSDK-means (Standard Deviation and Silhouette Coefficient-based k-means). In our algorithm, we select k randomly selected points from the dataset as initial centroids in the first step. Then, we assign each data to the closest cluster. In the third step, we calculate the average cost of all clusters by using standard deviation and silhouette coefficient. After the three steps are completed, we come back to the first step and repeat the processes for i times. The initial centroids which minimize the cost of all clusters are selected as the best centroids for k-means. In the experimental study, we compared our algorithm with Lloyd's k-means, k-means++, and k-medoids (PAM) on various datasets. According to obtained results, our algorithm outperformed other algorithms in the aspect of clustering quality.

Keywords: k-means, centroid initialization, standard deviation.

at least 3 keywords should be added

INTRODUCTION

Clustering algorithms are commonly used in many areas like machine learning [1, 2], bioinformatics [3, 4], data mining [5, 6], pattern recognition [7, 8], and web mining [9, 10] because they do not need the actual class labels. The main purpose of clustering algorithms is to define clusters according to the similarity among data. In the literature, there are mainly 5 types of clustering algorithms partitioning-based, hierarchical-based, model-based, density-based, and grid-based clustering algorithms. k-means [11], DBSCAN [12], OPTICS [13], Affinity Propagation [14], BIRCH [15], and WBFC [16] are some examples of clustering algorithms.

k-means is the most known clustering algorithm because of its usage simplicity, easy implementation, and effectiveness on datasets that have spherical clusters. However, it has disadvantageous. Its disadvantages can be listed as follows:

- The number of clusters must be defined,
- It can not handle outliers/noisy data,
- Its clustering quality depends on randomly selected initial centroids,
- It can not define clusters in arbitrary shapes.

In this study, we propose a new approach to overcome the problem of the randomness of selecting initial centroids. The aim is to select *k* centroids which makes the cost of clusters minimum. The details about the cost function are provided in the third section. For this purpose, we select the best initial centroids among randomly selected centroids.

The rest of the paper is organized as follows: in the next section related works are presented. The 3th section provides details about the material and method while the results and discussions are given in the 4th section. The study is concluded in the last section.

RELATED WORKS

In the literature, many approaches have been proposed to overcome the problem of selecting the initial centroids randomly. One of these approaches was proposed by Arthur and Vassilvitskii [17]. They select the first centroid randomly. For following centroids, they select data points that are farthest to determined centroids. This process continues until k centroids are determined.

Another study to find the best initial centroids was proposed by Yuan et al. [18]. They select the first initial center by finding the nearest two points in the dataset and delete these two points with their n-neighbors. Then, the next centroid is selected in the same way and the process continues until k sub-sets are created. The mean of each sub-set is calculated and these centroids are defined as the initial centroids.

Zalik [19] proposed a novel method that uses a cost function to find the best initial centroids. In his approach, the aim is to minimize the cost function. The cost function is an extended version of the mean-square-error cost-function of k-means. The most important advantage of the algorithm is that it does not need to define the number of clusters.

Thangavel and Kumar [20] proposed a novel clustering algorithm that does not use distance-based clustering. They used Combined Standard Deviation (CSD) as measure of similarity. The aim is to minimize the error. In each step, the minimum combined standard deviation is used which leads to optimal clusters. Their algorithm does not depend on centroids. This was the most important advantage of their algorithm.

Rahman et. al [21] proposed a method for centroid initialization based on Convex Hull algorithm. They select two initial centroids to make clusters well separated from each other as much as possible. Therefore, all selected centroids are selected according to to maximize the separation.

MATERIAL AND METHODS

In this section, we explain the details of the proposed approach.

Standard Deviation and Silhouette Coefficient-based Centroid Initialization

The main idea underlying the proposed method is that we select k centroids which minimize the average standard deviation of clusters and maximize the silhouette coefficient. For this purpose, we first select k points as initial centroids and assign the data to clusters according to the distance to these centroids. Then, we calculate both the average standard deviation and silhouette coefficient values of constructed clusters. The process is repeated i times and the centroids that minimize the overall standard deviation and maximize the silhouette coefficient are selected as the best centroids for k-means.

Let x be n-element vector and $\bar{x}$ be the mean, standard deviation (s) is calculated by Equation (1).

$$s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad (1)$$

In our approach, we find the average standard deviation of clusters using Equation (1) for each feature. Then, the average standard deviation of any cluster is calculated by calculating the average standard deviation of features. Finally, the overall standard deviation is calculated by calculating the average standard deviation of all clusters. So, let n be the number of data the cluster has, d be the number of features each record has, and k be the number of clusters; the overall average standard deviation (STD) of clusters is calculated by Equation (2).

$$STD = \frac{1}{k}\sum_{l=1}^{k}\frac{1}{d}\sum_{j=1}^{d}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad (2)$$

On the other hand, the silhouette coefficient (SC) is calculated by the formula given in Equation (3). Here, a is the average distance from x (the selected data) to all other data of the cluster to which x belongs and b is the average distance from x to all the data of the closest cluster to that x does not belong.

$$SC = \frac{1}{n}\sum_{i=1}^{k}\sum_{x \in C_i}\frac{b(x) - a(x)}{\max(a(x), b(x))} \qquad (3)$$

To make the produced clusters compact and separated, we use the formula given below to calculate the cost function ($SSDK$). This is the overall cot function of our method. We try to minimize this value as much as possible.

$$SSDK = \frac{STD}{SC}$$

Experimental Study

The datasets given in Table 1 were used in the experimental study. To make parameter selection easier Min-Max normalization is applied to all datasets. For algorithms, k value is selected from interval of [2, 20] while the maximum iteration is selected as 50.

In the experimental study, our algorithm is compared with Lloyd’s k-means, k-means++, and k-medoids (PAM). To select the best initial centers, each algorithm was run 100 times for each k-value on each dataset to ensure that enough trials were performed. Obtained results are compared in the aspects of ARI (Adjusted Rand Index), Purity, and Silhouette Index (SI). The results are presented in figures.

Table 1. Used datasets.

Datasets	**Type**	**# of records**	**# of attributes**	**# of classes**
2d-4c	Synthetic	863	2	4
Aggregation	Synthetic	788	2	7
Outliers	Synthetic	700	2	4

RESULTS AND DISCUSSIONS

The comparisons of obtained results for algorithms are provided in Figures 1, 2, 3, 4, 5, and 6. It is clear that our algorithm was the best among the others in the aspect of clustering quality. In Aggregation dataset, it was the most successful one in the ARI index as seen in Figure 1. Its ARI value was 0.8155 while the second-best one’s value was 0.8057. The curves of ARI, Purity, and SI against the number of clusters are provided in Figure 2. These curves also support our belief. It can be said that the clustering quality of our algorithm was better than the others in this dataset.

In Outliers dataset, the success of our algorithm is clear as seen in Figures 3 and 4. The ARI value of our algorithm was 0.9284 while the successes of the other algorithms were 0.8463. As illustrated in the curves that are given in Figure 4, our algorithm was better than the others in the ARI, Purity, and SI against the number of clusters curves.

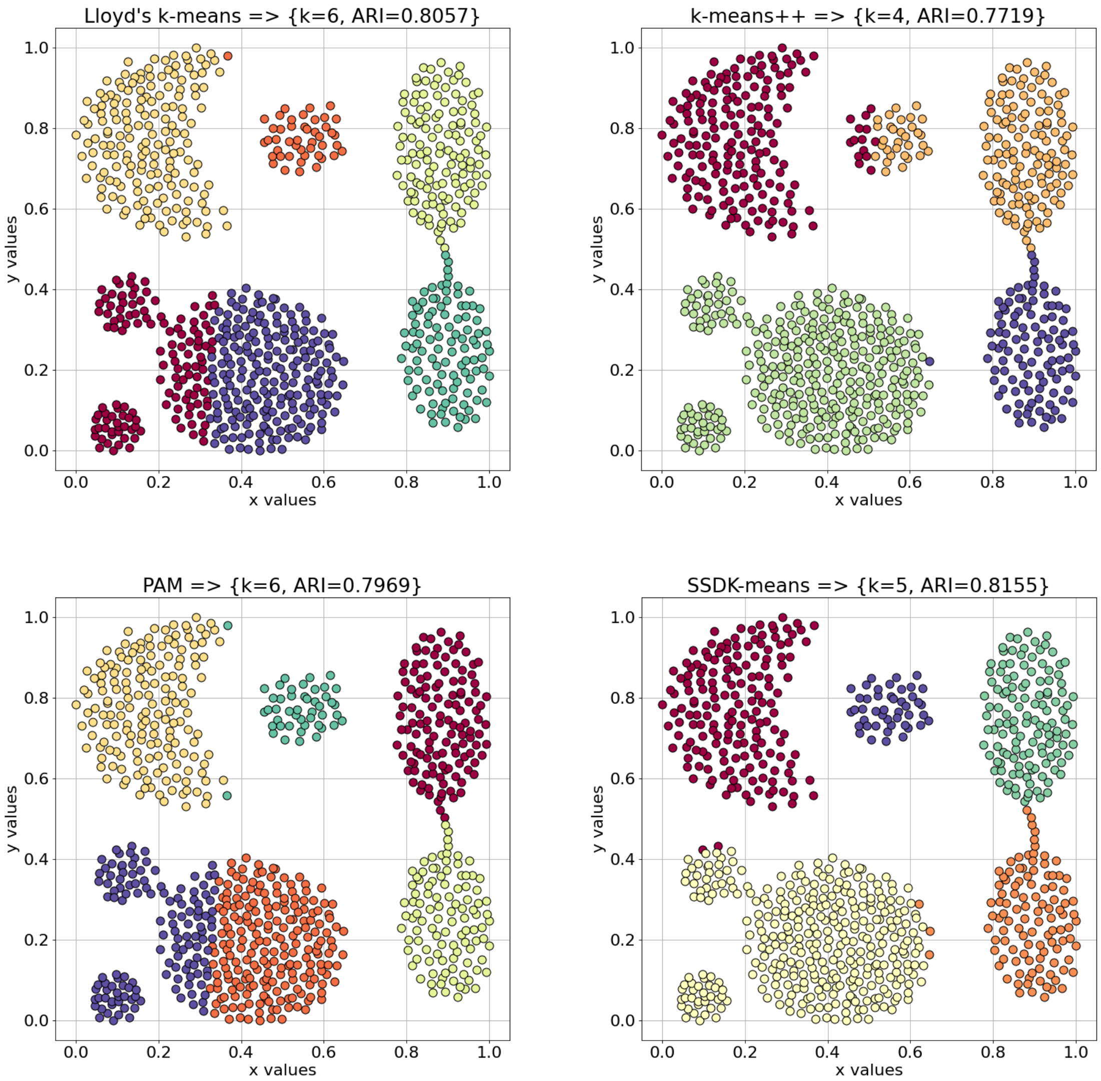

Figure 1. Visual results obtained on Aggregation dataset.

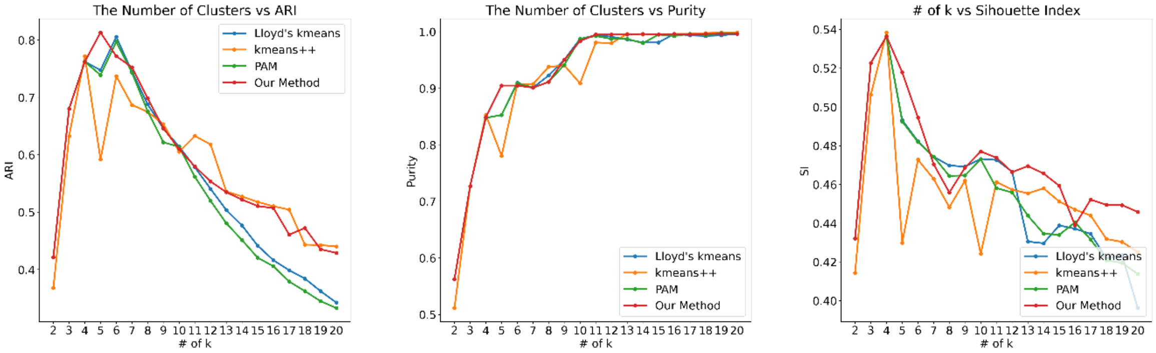

Figure 2. Comparison of ARI, Purity, and SI of algorithms on Aggregation dataset.

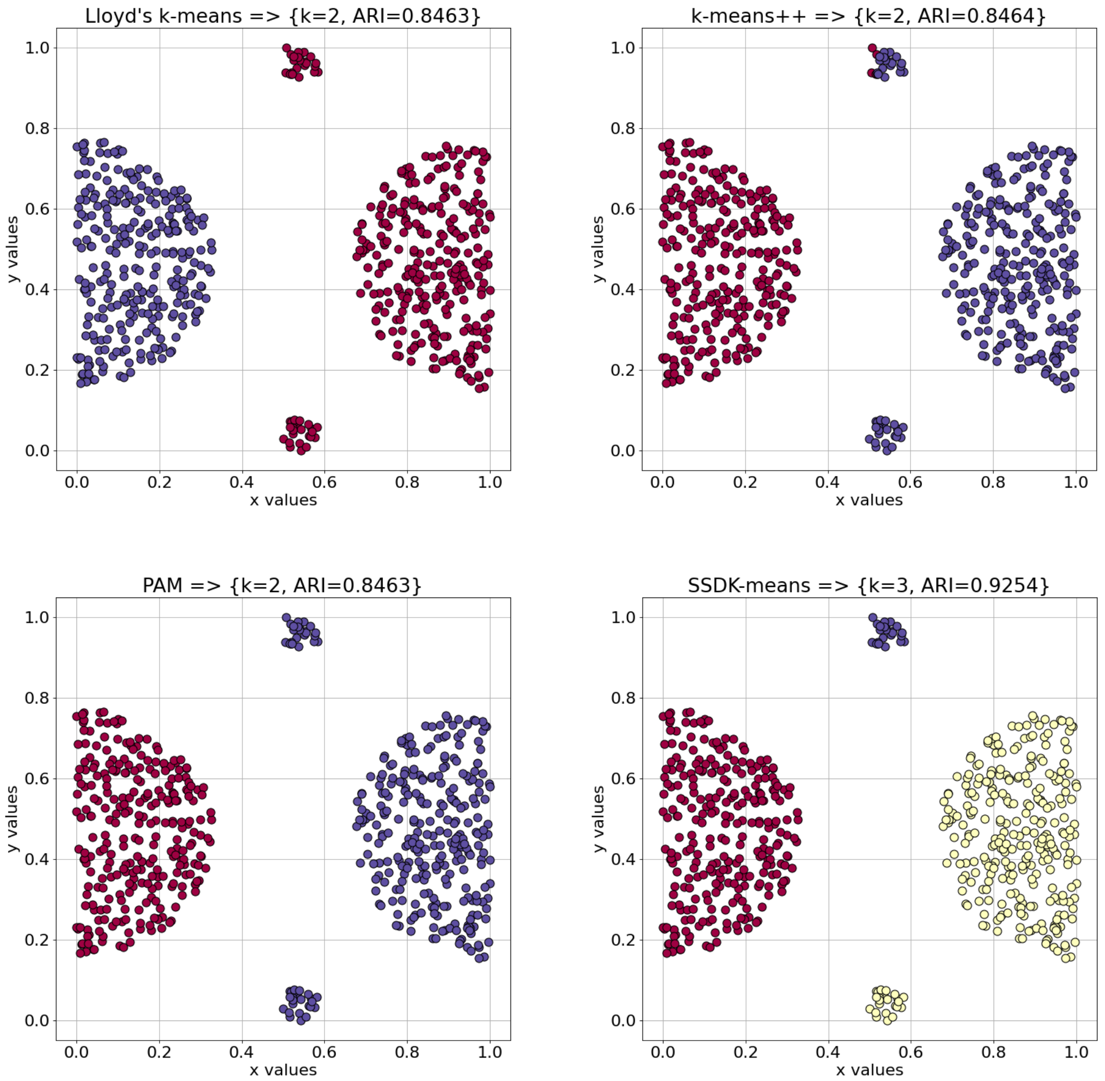

Figure 3. Visual results obtained on Outliers dataset.

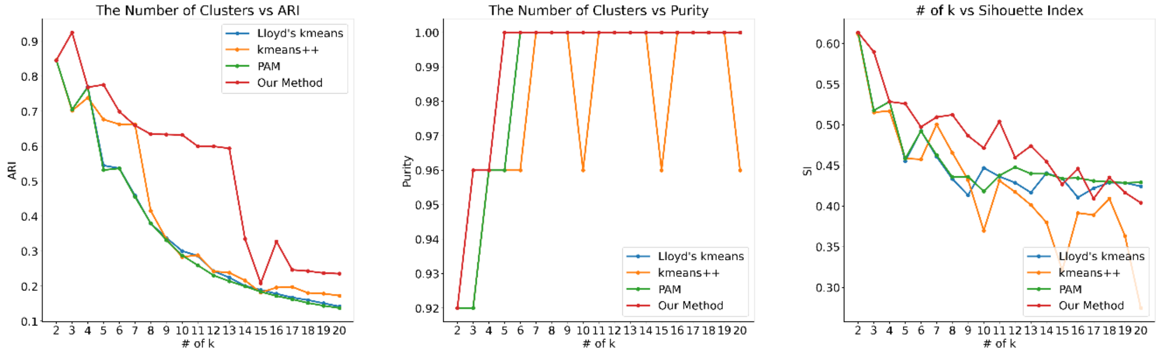

Figure 4. Comparison of ARI, Purity, and SI of algorithms on Outliers dataset.

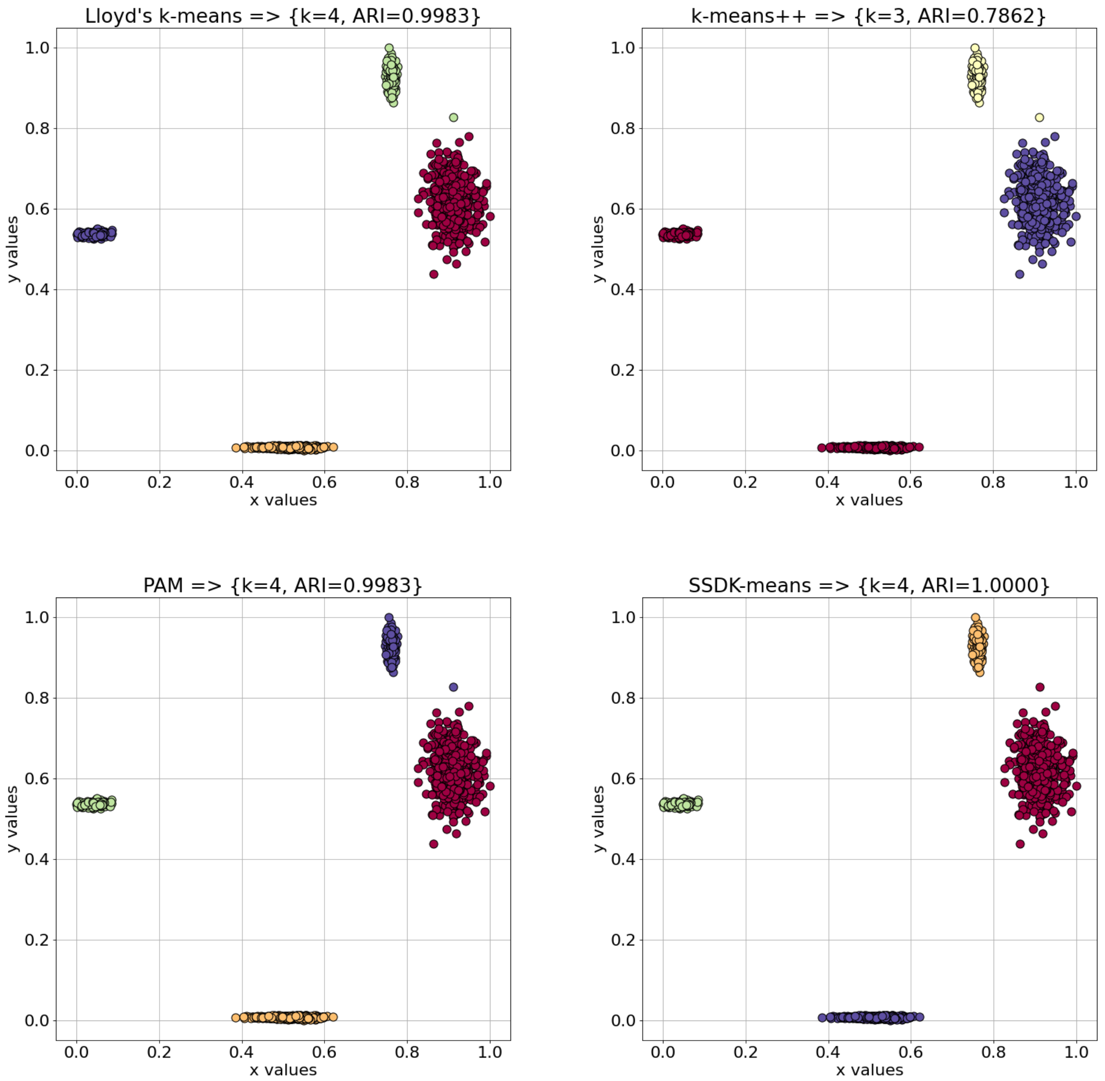

Figure 5. Visual results obtained on 2d-4c dataset.

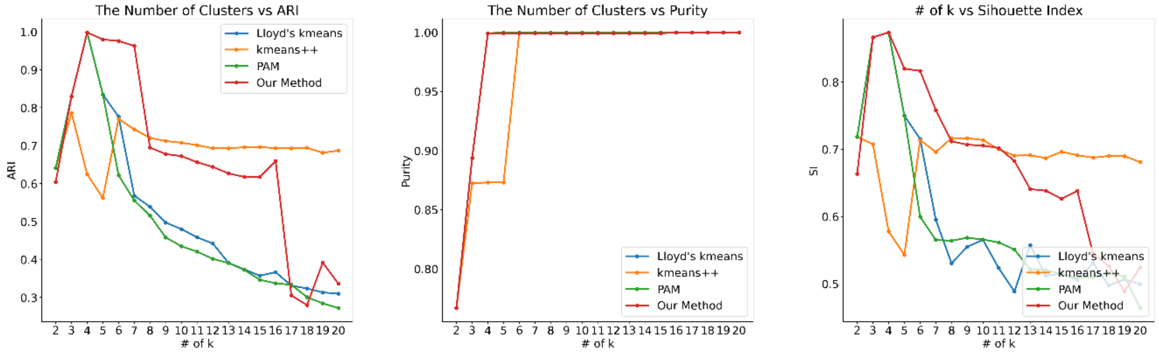

Figure 6. Comparison of ARI, Purity, and SI of algorithms on 2d-4c dataset.

As for 2d-4c dataset, our algorithm was the best one among the others. the ARI value of our algorithm was 1.0000 while Lloyd's k-means' and PAM's ARI values were 0.9983 as shown in Figure 5. Besides, ARI, Purity, and SI against the number of clusters curves also support this idea as can be seen in Figure 6.

CONCLUSION

In this study, we proposed a standard deviation-based centroid initialization method for k-means clustering algorithm. It selects the centroids which minimize overall standard deviations of clusters among possible centroids as initial centroids. In the experimental study, we compared our method with Lloyd's k-means, k-means++, and PAM algorithms. According to the experiments, our algorithm outperformed the other algorithms in the mean of clustering quality. It can find the best initial centroids in terms of various numbers of clusters when compared with the other algorithms.

REFERENCES

1. Şenol, A. and H. Karacan, *A Survey on Data Stream Clustering Techniques.* European Journal of Science and Technology, 2018(13): p. 17-30.
2. Kumar, V., M.S. Chauhan, and S. Khan, *Application of Machine Learning Techniques for Clustering of Rainfall Time Series Over Ganges River Basin*, in *The Ganga River Basin: A Hydrometeorological Approach*. 2021, Springer. p. 211-218.
3. Yu, Z., H.-S. Wong, and H. Wang, *Graph-based consensus clustering for class discovery from gene expression data.* Bioinformatics, 2007. **23**(21): p. 2888-2896.
4. Zou, Q., et al., *Sequence clustering in bioinformatics: an empirical study.* Briefings in bioinformatics, 2020. **21**(1): p. 1-10.
5. Han, J., M. Kamber, and J. Pei, *Data mining concepts and techniques third edition.* The Morgan Kaufmann Series in Data Management Systems, 2011. **5**(4): p. 83-124.
6. Sabor, K., et al., *A data mining approach for improved interpretation of ERT inverted sections using the DBSCAN clustering algorithm.* Geophysical Journal International, 2021.
7. Sathya, B. and R. Manavalan, *Image Segmentation by Clustering Methods: Performance Analysis.* International Journal of Computer Applications, 2011. **29**: p. 27-32.
8. Li, C., et al., *A Review of Clustering Methods in Microorganism Image Analysis*, in *Information Technology in Biomedicine*, E. Pietka, et al., Editors. 2021, Springer International Publishing: Cham. p. 13-25.
9. Aggarwal, C.C. and C.R.D. Clustering, *Algorithms and Applications*. 2014, CRC Press Taylor and Francis Group.
10. Rambabu, M., S. Gupta, and R.S. Singh, *Data Mining in Cloud Computing: Survey*, in *Innovations in Computational Intelligence and Computer Vision*. 2021, Springer. p. 48-56.
11. Lloyd, S.P., *Least squares quantization in PCM.* IEEE Trans. Inf. Theory, 1982. **28**: p. 129-136.
12. Ester, M., et al., *A density-based algorithm for discovering clusters in large spatial databases with noise*, in *Proceedings of the Second International Conference on*

Knowledge Discovery and Data Mining. 1996, AAAI Press: Portland, Oregon. p. 226-231.

13. Ankerst, M., et al., *OPTICS: ordering points to identify the clustering structure.* SIGMOD Rec., 1999. **28**(2): p. 49-60.
14. Frey, B.J. and D. Dueck, *Clustering by Passing Messages Between Data Points.* 2007. **315**(5814): p. 972-976.
15. Zhang, T., R. Ramakrishnan, and M. Livny, *BIRCH: an efficient data clustering method for very large databases.* SIGMOD Rec., 1996. **25**(2): p. 103-114.
16. Zhang, C., J. Xue, and X. Gu, *An Online Weighted Bayesian Fuzzy Clustering Method for Large Medical Data Sets.* Computational Intelligence and Neuroscience, 2022. **2022**: p. 6168785.
17. Arthur, D. and S. Vassilvitskii, *k-means++: The advantages of careful seeding*. 2006, Stanford.
18. Fang, Y., et al. *A new algorithm to get the initial centroids*. in *Proceedings of 2004 International Conference on Machine Learning and Cybernetics (IEEE Cat. No.04EX826)*. 2004.
19. Žalik, K.R., *An efficient k'-means clustering algorithm.* Pattern Recognition Letters, 2008. **29**(9): p. 1385-1391.
20. D, A.K. and D.T. Kuttiyannan, *A Combined Standard Deviation Based Data Clustering Algorithm.* Journal of modern applied statistical methods: JMASM, 2006. **5**: p. 258-265.
21. Rahman, Z., et al. *An enhanced method of initial cluster center selection for K-means algorithm*. in *2021 Innovations in Intelligent Systems and Applications Conference (ASYU)*. 2021.