IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 10, Issue 6, June 2022

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.165

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 6, June 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1006109 |

A Survey on K-Means Clustering Algorithm

Raghavendra D, Ajay Y P, Mr.Srinivasulu M*

Department of Master of Computer Application, University B.D.T College of Engineering. Davanagere, Karnataka, India

ABSTRACT: It's called "clustering" when you're analysing data this way. K-means is the most used and oldest method for clustering data. There are both benefits and downsides to using the K-means approach. This collection of articles also includes research on enhanced K-means. The accuracy and efficiency of traditional K-means may also be enhanced. Two adjustments may be made to K-means to make it more accurate: It is important to choose a few beginning locations. After determining the average and distance between two data points, each data point is placed in a cluster that is closest to it using the formulae. In terms of speed and efficiency, the suggested K-means technique is superior to the present method. It'll also be more precise. Every day, the globe generates enormous amounts of data in several sectors. Data mining and data analytics are the techniques used to sift through large amounts of data to uncover hidden information. Data clustering is one of the most common techniques used in data mining. When data is grouped together, it is simpler to extract certain pieces of information from each group. Clustering may be done using a variety of methods. It's a well-known technique that's been used for more than 50 years to group items together. k-means K-means remains the most widely used clustering technique, despite several improvements and modifications. It's still in use today in a variety of different professions and industries. An SLR (Structural Literature Review) is being used in this work to gather primary papers on k-means clustering algorithms and organise, analyse, and summarise the findings.

I. INTRODUCTION

As computer technology has been more widely used, large data sets with many dimensions have emerged [2]. Because the data is kept digitally in an electronic media, automated data analysis, classification, and retrieval may be feasible. Using clustering in data analysis, it is possible to group together data points that are related but yet different. It is possible to extract useful information from enormous amounts of digital data using clustering analysis. You may use it if you're into bioinformatics or pattern recognition. Marketers and economists alike may gain from its use.

Since its inception, K-means has been one of the most extensively used techniques for categorising data. The parameters of this approach are used to pick a predetermined number of clusters (k) before doing the analysis. After then, data points are randomly assigned to their nearest centroids in a second round. This method will continue forever after the convergence requirements are met. Although K-means has its flaws, new methods are needed to improve the results of the study. The clustering technique's accuracy and efficiency have been significantly improved by the research considered in this article [3, 6, 7, 8, and 9]. The major goal of this study is to get initial centroids and assign data points to neighbouring clusters. This study's clustering algorithm has space for improvement. It's possible that subsequent ways to selecting values for the initial clusters will be improved. The most widely used clustering method in research and industry is K-means clustering. When N items are divided into k groups, each object is placed in the cluster that is closest to its mean. [3]

The Traditional K-Means clustering algorithm

- Select the value of K i.e. Initial centroids.
- Calculate the new global centroid for each cluster
- Form K cluster by assigning each point to its closest centroid
- Find the nearest point from that centroids in the Dataset. 4
- Repeat step 3 and 4 for all data points in dataset.

Properties of k-meansclustering algorithm

- Efficient while processing large data set.
- It works only on numeric values.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 6, June 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1006109 |

- The shapes of clusters are convex.

Because of its simplicity and effectiveness, K-means is the most often used technique for breaking up data in cluster analysis. If you have a huge dataset, you can't utilise this method since it's time-consuming, sensitive to outliers, and its findings rely on the initial centroids, which are selected at random. Many suggestions have been made on how to improve K-Means. None of the solutions offered are worldwide. Some of the offered methods are quick, however they don't maintain the quality of the clusters. However, they may be highly costly in terms of the difficulty with which they can be used. The quality of clusters will be harmed by outliers. Algorithms that only operate with sets of numbers are called set-based algorithms.

1.1 Issues and limitations ofK-means clustering algorithm

The researchers initially sought to find out what K-means couldn't achieve in order to come up with a solution. More than half of the 114 publications we looked at for this study dealt with the K-means method's shortcomings. The majority of research have indicated that despite minor variances in their attribute lists, they share many of the same characteristics. Locating k, finding the first cluster centres, dealing with noise, and dealing with huge datasets are the most common issues with Kmeans, according to some. Figure 4 displays our findings on the frequency with which certain articles discussed the same issues.

- **Figuring out how many clusters there are**: The number of clusters (the value of k) must be specified as a fixed input before starting the K-means algorithm. The number of clusters (the value of k) is one of the most problematic aspects of K-Means, according to 12 of the 17.65 percent of studies discussing the issue. For example, [8] said that determining the number of clusters is one of the most difficult aspects of cluster analysis. Like [9] said, when it comes to K-means, the ultimate outcome of clustering relies largely on various factors, such as the cluster centre and cluster count. It's also not always feasible for all sorts of applications to figure out the proper amount of clusters in advance. In contexts where datasets fluctuate over time, the value of k cannot be predetermined. Because of the lack of an accurate technique to determine the optimal value of k, K-means is of little practical utility. [11].

- **Initial cluster centres:**K-means is a clustering technique that uses a point known as a centroid to identify groups. The first cluster centres must be selected before performing the algorithm. So, k cluster centres must be selected at random to represent k clusters. K-means uses an initial cluster as a starting point, and then requires more clustering processing [12] to produce a more accurate representation of the data. Algorithm performance might be negatively affected by the selection of the cluster's centre of gravity. More than half (51 of 75) of the 75 studies discussing the constraints of the K-means method for clustering said that the initialization of the centroid is critical. According to [13], [14], [15], and [16], you should consider how sensitive K-means is to the initialization of the cluster's centroid if you want the most accurate and efficient clustering. According to [17], the clustering effect is made even more random by the random selection of the initial clustering centres.

The real quantity of data is large and sophisticated as a result of the growth of various data structures. As a result of this, it's possible to overcomplicate things and so make clustering difficult. In an important work, [18] shown that the K-means clustering algorithm may avoid the local optimum condition by selecting the best cluster centres. It decreases the number of clustering iterations, which is d, and improves the accuracy and stability of the clustering process. The following criteria should be used to choose the initial cluster centres: 16 Clusters with good separation and density may be produced using the fewest iterations feasible, according to the Systematic Review of K-Means Clustering Algorithm ICNCC 2020, Tokyo, Japan, December 18–20, 2020.

- **Noise and anomaly:** Basically, it is hard to cluster an object with noise (values that are either very high or very low). Outliers and noisy data can throw off the k-means method [19]. Noise can also change the results of clustering. A good algorithm should be able to handle data that is noisy or has outliers. The K-means algorithm is vulnerable to noise points and isolated points, and the noise points affect the results of the K-means clustering process [20]. Seven papers (10.29%) that talked about the problems with K-Means said that it can't handle noise and outlier data points well, and that even a small amount of this kindof data can have a big effect on the mean value[21].

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 6, June 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1006109 |

- **Size of dataset:** When K-means is used on a large amount of data, it often runs into problems [22]. 8.76% of the papers that talked about this subject were about this. [23] said that the current clustering algorithms need scalable ways to deal with big datasets. Also, if the random algorithm is run several times under the same conditions, it might produce results that don't match up, and it doesn't work well with large datasets [24]. In the K-means method, you have to keep changing the location of the new cluster centre until the location stops changing. When the K-means algorithm is used on big datasets, it is hard to make it work well in terms of both performance and efficiency. [13]. Also, from the steps of the K-means algorithm, it's clear that K-means requires frequent sample grouping and debugging, which could make the cost of the algorithm high. If there are a lot of data, [12]. In short, the number of iterations is affected by the initial centroid of a cluster and the size of the dataset [25]. The longer algorithm takes longer to run because the bigger dataset makes the cluster size and number of iterations bigger.

1. Literature Review on K-means clustering Algorithm

Making observations, compiling data, and forming groups are all part of clustering. The term "cluster" refers to a collection of records that are related to one another. Clustering is distinct from classification since there is no goal variable in clustering. Clustering is very essential when the mean is close.

The traditional K-Means method [3] is really simple: The value of K, which stands for "starting centres," should be selected in step one. 2. Repeat steps 3 and 4 for each dataset point. 3. Locate the nearest point in the dataset to the centre of interest. Give each point to the centroid closest to it in a K cluster 5 The new global centroid must be determined for each cluster separately. These are the properties of the k-means algorithm: One of the best ways to handle a vast amount of data. 2. It can only be used with numbers. There are three characteristics of clusters: Because of its simplicity and effectiveness, K-means is the most often used technique for breaking up data in cluster analysis. However, it cannot be used on big datasets since it is difficult to calculate, it is susceptible to outliers, and its findings are dependent on the starting centroids, which are picked at random. There have been several suggestions for improving K-Means. No one has come up with a universal answer.

Although some of the suggested techniques are quick, they don't maintain the high quality of the clusters. It is possible to build excellent clusters with a high price tag, though. Clusters will suffer if there are outliers in them. A collection of integers is all that certain algorithms need. Text mining, geographical database applications, and online analysis are just a few examples of data mining applications. The hierarchical approach and the non-hierarchical method of clustering are two options. Using the hierarchical technique is the best option if you don't yet know how many groups you need. The non-hierarchical strategy is used if you already know how many groups you want. As part of the hierarchical clustering process, the Fuzzy K-Means algorithm is termed [6] Fuzzy K-means has become a popular cluster technique since it is simple to implement. However, this approach may also be used to create a database cluster using qualities from other kinds by converting these attributes into an index of how similar or distinct they are. By default, the K-Means method creates a predetermined number of clusters, K, in which items with similar qualities are grouped together. Members of a cluster are selected depending on their distance from the cluster centre (centroid) [1]. Partitioning and clustering data may be accomplished using the K-Means technique [11]. Clustering using K-means may be done as follows. a. Count the number of clusters K. b. Finding the cluster's centre (centroid) may be accomplished in a variety of methods. But picking a random number is the most usual method. Random selection is used to determine the locations of the nodes that make up a cluster. K-Means with Map Reduce, suggested by Amira Boukhdhir, Oussama Lachiheb, and Mohamed Salah Gouider [1], was shown to be more efficient when used to huge datasets. Traditional K-Means, PK-Means, and Fast K-Means take longer to execute than the algorithm does. Take the highest extreme number in a collection of numbers. also There is a technique called Map Reduce that is used to choose the first clusters of interest and to build clusters of interest.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 6, June 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1006109 |

Nevertheless, there are some limitations, such as having to input the number of centroids. Number sets are the only ones that are compatible with this. Furthermore, the number of clusters isn't predetermined. Kmeans can now run more quickly because to an algorithm developed by Duong Van Hien and Phayung Meesad. They were able to do this by omitting a couple of the process's final phases. This experiment approach reduces the number of repeats by 30%, resulting in a 30% time savings while maintaining good accuracy. Randomly selected centres, on the other hand, produce unstable clusters. The accuracy of clustering findings might be tainted by noise points. Traditional k-means clusters might be improved, thanks to work by Li Ma and colleagues [3, 4]. They employed a methodical approach to choosing the number of clusters and the first cluster centres. The outlier issue was overcome by reducing the amount of noise points. This method generates high-quality clusters, but it takes a long time to complete. Algorithm named an enhanced k-means by Xiaoli Cui and colleagues [4].

Instead of analysing the whole dataset, this algorithm uses a sampling strategy to focus on a subset of important data points. Both the I/O and the network costs were reduced as a result of Parallel Kmeans. According to the trials, although the algorithm works well and is superior than k-means in certain ways, it lacks precision in others. G. Sahoo [5] and Yugal Kumar [4] focused on issues with how K-Means begins. Both the question of how many clusters are required for clustering and the question of setting up the initial centres for the clusters are valid methods of stating the K-Means algorithmic issue. In this study, we'll look at a method for establishing the first cluster centres. The first cluster points, also known as the first centroid, are put up by the K-Means algorithm using a binary search initialization approach. We used data from the UCI repository to evaluate the performance of an algorithm. It's difficult to analyse patterns of user behaviour since the numbers aren't sensitive enough and the data isn't dispersed uniformly over location and time, as Huang Xiuchang and SU Wei [6] found out the hard way.

Using the previous clustering approach doesn't provide accurate results anymore. In this study, current techniques for clustering, trajectory analysis, and behaviour pattern analysis are examined, and the clustering algorithm is applied to the trajectory analysis. Changes were made to the classic K-MEANS clustering algorithm in order to provide a better technique of analysing user behaviour patterns. We tested the algorithms using simulated and actual data, with positive findings for the new approach in dealing with issues related to user behaviour patterns. Singh, Nidhi and Singh, Divak [7] K-means is often used in clustering methods. In the iris dataset, k-means outperforms hierarchical clustering, but in the diabetes dataset, hierarchical clustering outperforms k-means. Using k-means, data sets may be grouped together in a fraction of the time. When putting things into groups, it's important to ensure that items in the same group are more similar to each other than those in separate groups. The K-Means method performs well on huge datasets in this study. Kedar B. Sawan (author) We now have a lot of issues with the K-means clustering technique. How many clusters will develop and where they will be located depends on where you begin. The number of clusters, K, and a novel method for selecting the initial centroid locations for the K-means algorithm are both discussed in this study. Clustering algorithms may be improved by using a modified version of the K-means technique. There are similarities between the new approach and K-means clustering since it takes into consideration data regarding how effectively an algorithm performs. The new algorithm makes advantage of a

Finding the first centroid spots in a methodical manner. This reduces the amount of dataset scans and improves the K's accuracy by iterating less often. [6] According to A. Abdul Nazeer et al., distinct clusters are formed by the kmeans algorithm depending on the original centroids. The final quality of the clusters is directly related to the starting quality of the centres. First, the initial centroids are found, and then the data points are placed in clusters that are closest to them, and the clustering mean is recalculated. Both the centroid-based K-Means clustering technique [7] and the representative object-based FCM clustering algorithm [7] are examined by Soumi Ghosh et al. in this paper. On the basis of this assessment, we're going to address this topic.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 6, June 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1006109 |

It is possible to increase the quality of the clusters and discover the optimal number of clusters using a modified K-means method, as shown by Shafeeq et al. The number of groups is provided by the user as input to the K-means algorithm (K). However, in the actual world, it is impossible to predict in advance the number of clusters that will exist. The strategy proposed in this study may be used whether or not the number of clusters is known in advance. You may either provide an exact number of clusters or enter the minimum number of clusters that are required. Each iteration of the method adds a new cluster centre to the cluster counter until the quality of the cluster is verified. This method will discover the optimal number of clusters on the fly to tackle this issue. According to Junatao Wang et al. [9], the K-means method may be improved with a noisy data filter.

Traditional k-means clustering suffers from a number of drawbacks, which this new approach addresses. A density-based detection technique is generated by the algorithm based on the properties of noisy data. The original algorithm is used to find and analyse the noisy data. The accuracy of the k-means method is considerably enhanced by pre-processing the data to remove the noise data before clustering the data sets, which improves the cohesiveness of the clustering results and reduces the influence of noise data on the algorithm. Shi Na and others Take a look at the shortcomings of the typical k-means method. It is necessary for the K-means algorithm to calculate the distance between each data point and the centres of all the clusters at least once throughout each iteration. The clustering algorithm's performance suffers as a result of this iterative procedure. K-means can be improved by using this paper's recommendations. When you go through an iteration, you'll need to keep track of certain data that will be useful in the following one. Time is saved by not having to calculate the distance at each step.

KA The Kimply method proposed by Abdul Nazeer et al. [6] generates groups that are unique to each of the original centroid estimations. The starting centroids used in the computation determine the quality of the final group. The unique k-involves computation has two stages: first, the initial centroids are determined, and then the information centroids are assigned to the closest clusters and the cluster mean is recalculated.. K-Means based on centroid and FCM (Fuzzy C=Means) clustering computations based on agent queries are presented in a relevant lecture by Soumi Ghosh et al. [7]. This exchange is based on the use of these computations to evaluate the efficiency of clustering performance during the run.

It is possible to increase cluster quality and discover the optimal number of clusters using a modified Kimplies computation, as shown by Shafeeq et al[8].

Customer-supplied number of clusters (K) used to compute Kimplies. In any event, it is impossible to predict the number of groups in advance in practise. It is possible to use this approach for both known and unknown groups, as shown in this text. It is possible for a client to customise the number of clusters or the information about the minimum number of clusters necessary. Every cycle, the cluster counter is increased until it reaches cluster quality validity, at which time new cluster foci are calculated. By determining the optimal amount of bursts in the run, this calculation solves the issue.

Here, Junatao Wang et al. [9] provide a new k-means algorithm that makes use of commotion data. This new Kimplies clustering calculation addresses the shortcomings of the standard Kimplies clustering computation. Companies' disclosure and production of requirements information are added to a first calculation to build a thickness-based identification approach for the quality of the information. Cluster results and cluster connectedness are much enhanced, and the influence of motion information on k-means calculations is suitably minimised, as a consequence of pre-treating prohibition information from this shell data before clustering the data sets.

II. RESEARCH METHODOLOGY ON K-MEANS CLUSTERING ALGORITHM

It is our intention to discover new applications for the K-means method. Furthermore, this study's purpose is to demonstrate the issues that academics have identified and the solutions they have proposed in the context of the k-means clustering method. A Systematic Literature Review (SLR) was conducted to gather, sort by primary papers that

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 6, June 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1006109 |

employed various versions of the K-means clustering method, and to analyse the results of this analysis. SLR enables you to identify, assess, and comprehend all papers related to a certain study field. Many research have attempted to refine and optimise the K-means algorithm, which is one of the most common techniques to learn without being seen.

In the ACM Digital Library, IEEE Xplore Digital Library, ScienceDirect journal, SpringerLink journal, and the Web of 13 ICNCC 2020 December 18–20, Tokyo (Japan), Japan the Ashabi et al. In scientific databases, the years 2015 to 2020 were searched. As part of the search, K-means and the phrases "modified" and "optimised" were placed together. Results were limited to those that were authored in English alone. An Elearning application at SMK Negeri 2 Bengkulu Tengah has been used to analyse the implementation of the K-Means Clustering Algorithm for electronic-based learning. Figure 2 depicts the overall structure of the study.

Afterwards, a literature review was conducted by comparing the findings of prior studies. [1] The purpose of data partitioning is to organise data into groups with similar qualities and groups with differing features. After establishing the extent of the issue, it was analysed to determine what the intended outcomes were. Study of K-Means Clustering Algorithm literature, as well as computations and grouping findings, led to a grouping of student activity in the E-Learning application's learning activities. We gathered our information from the Elearning application's database, which contains information on what students and teachers have done while using the app[12]. Following is a flowchart of the k-means clustering method [1].

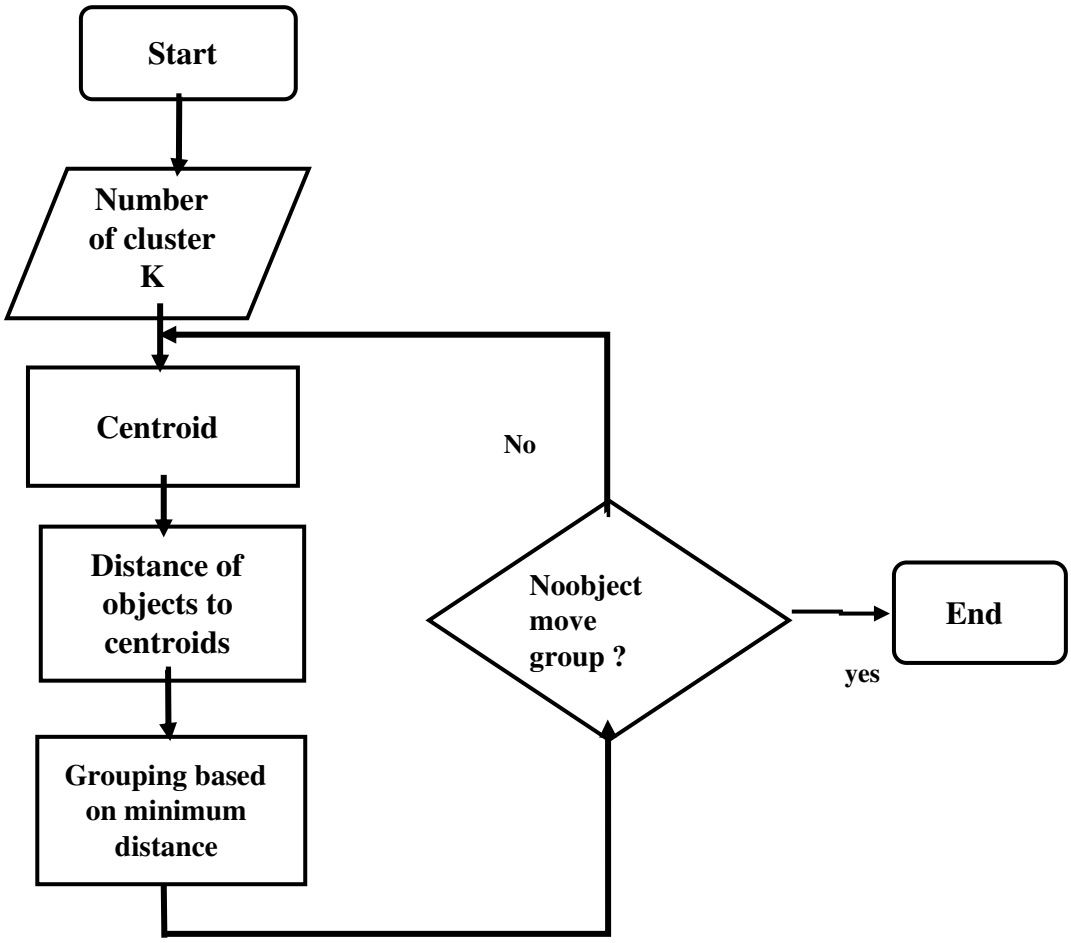

Figure 1: Flowchart of K-Means Clustering Algorithm

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 6, June 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1006109 |

Sl No.	Author	Methodology Used	Objectives	Limitations
1.	K. A. Abdul Nazeer et al	K-Means Algorithm	Presented an improved clustering method that calculates the initial centroid values and also effectively allocated the data points to the clusters. Enhanced the correctness of K-Means Algorithm.	The limitation of this algorithm lies on the fact that despite the distribution of the various data points, it is still required to give count of clusters as an input
2.	Soumi Ghosh et al.	K-Means Algorithm, Fuzzy C- Means Algorithm	Performs relative investigation of Fuzzy C Means and K means algorithm based on the criteria of time complexity. K- Means algorithm seems far better than Fuzzy C-Means.	The calculation time taken is more because of the fuzzy measurements.
3.	Shafeeq et al.	Modified K-Means Algorithm	The exact number of devised on the basis of run method of clustering. It works good for both familiar and non-familiar number of clusters.	The technique devised takes more time for calculation than k means in case of big data sets.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 6, June 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1006109 |

4.	Junatao Wang et al.	K-Means Algorithm	The updated algorithm produces less nose data as compared to the earlier researches	The noise impact is more in cluster forming.
5.	Shi Na, Liu Xumin	K-Means Algorithm	Enhances the speed and decreases the calculative complexity.	The algorithm used for the selection of the centroid is not very effective.

Table 1: Comparison among various existing Approaches and its Limitations

III. CONCLUSION

When clustering data, this is the most common method and relies on assigning data points to clusters that are closest to a starting point. There are more benefits than downsides to K-means clustering, although it still needs some improvement. An improved method for determining where the initial centroids are and how to allocate data points more accurately is presented in this study. They also speed up the procedure in comparison to the conventional k-means method. This remains a source of worry since it has the potential to enhance clustering accuracy, paving the door for a more advanced approach in the future.

REFERENCES

1. H. Jiawei,M. Kamber, and J. Pei, Data Mining: Concepts and Techniques, San Francisco California, Morgan Kaufmann Publishers, 2012.
2. Amira Boukhdhir Oussama Lachiheb, Mohamed Sala Gouider. “An improved Map Reduce Design of Kmeans for clustering very large datasets”, IEEE transaction
3. V. Duon, M. Phayung. ”Fast K-Means Clustering for very large datasets based on Map Reduce Combined with New Cutting Method (FMR KMeans)”, Springer International Publishing Switzerland, 2015.
4. M. Li and al. “An improved k-means algorithm based on Map reduce and Grid”, International Journal of Grid Distribution Computing, (2015)
5. C. Xiaoli and al. “Optimized big data K-means clustering using Map Reduce”, Springer Science + Business Media New York (2014)
6. Yugal Kumar and G. Sahoo, “A New Initialization Method to Originate Initial Cluster Centers for K-Means Algorithm”, International Journal of Advanced Science and Technology Vol.62, (2014).
7. Huang Xiuchang , SU Wei ,”An Improved K-means Clustering Algorithm“ ,JOURNAL OF NETWORKS, VOL. 9, NO. 1, JANUARY 2014
8. Nidhi Singh, Divakar Singh,” Performance Evaluation of K-Means and Hierarchal Clustering in Terms of Accuracy and Running Time”, (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 3 (3) , 2012.
9. Kedar B. Sawant, “Efficient Determination of Clusters in K-Mean Algorithm Using Neighborhood Distance “International Journal of Emerging Engineering Research and Technology Volume 3, Issue 1, January 2015.
10. Bapusaheb B. Bhusare, S. M. Bansode, ”Centroids Initialization for K-Means Clustering using Improved Pillar Algorithm” , International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 3 Issue 4, April 2014 .

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 6, June 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1006109 |

11. Kamaljit Kaur, Dr. Dalvinder Singh Dhaliwal, Dr. Ravinder Kumar Vohra ,"Statistically Refining the Initial Points for K-Means Clustering Algorithm ",International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 2, Issue 11, November 2013.
12. Abhijit Kane," Determining the number of clusters for a Kmeans clustering algorithm", Indian Journal of Computer Science and Engineering (IJCSE) Vol. 3 No.5 Oct-Nov 2012
13. Omar Kettani, Faical Ramdani, Benaissa Tadili, "AKmeans: An Automatic Clustering Algorithm based on Kmeans ", Journal of Advanced Computer Science & Technology, 4 (2) (2015) .
14. Avni Godara, Varun Sharma," Improvement of Initial Centroids in Kmeans clustering Algorithm", Vol-2 Issue-2 2016 IJARIIE
15. D. Sharmila Rani, V.T. Shenbagamuthu,"Modified KMeans Algorithm for Initial Centroid Detection", International Journal of Innovative Research in Computer and Communication Engineering (An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014
16. Effat Naaz, Divya Sharma, D Sirisha, Venkatesan M," Enhanced Kmeans clustering approach for healthcare analysis using clinical documents", International Journal of Pharmaceutical and Clinical Research 2016
17. A.K. Jain, Data clustering: "50 years beyond K-means, Pattern Recognition Letters", Elsevier, vol.31, pp.651-666, 2010
18. M. Yedla, S.R. Pathakota, and T.M. Srinivasa, "Enhancing K-means Clustering algorithm with Improved Initial Center", International Journal of Computer Science and Information Technologies, vol.1 (2), pp.121-125, 2010.
19.] S. S. Yu, S. W. Chu, C. M. Wang, Y. K. Chan, and T. C. Chang, "Two improved k-means algorithms," Appl. Soft Comput. J., vol. 68, pp. 747–755, 2018.
20. T. Wang and J. Gao, "An Improved K-Means Algorithm Based on Kurtosis Test," J. Phys. Conf. Ser., vol. 1267, no. 1, 2019.
21. X. Wang and Y. Bai, "The global Minmax k-means algorithm," Springerplus, vol. 5, no. 1, 2016
22. C. Lutz, S. Breb, T. Rabl, S. Zeuch, and V. Mark, "Efficient and Scalable k-Means on GPUs," Datenbank Spektrum, pp. 157–169, 2018.
23. C. Sreedhar, N. Kasiviswanath, and P. Chenna Reddy, "Clustering large datasets using K-means modified inter and intra clustering (KM-I2C) in Hadoop," J. Big Data, vol. 4, no. 1, 2017.
24. R. M. Esteves, T. Hacker, and C. Rong, "Competitive K-means: A new accurate and distributed K-means algorithm for large datasets," Proc. Int. Conf. Cloud Comput. Technol. Sci. CloudCom, vol. 1, pp. 17–24, 2013.
25. B. Xiao, Z. Wang, Q. Liu, and X. Liu, "SMK-means: An improved mini batch k-means algorithm based on mapreduce with big data," Comput. Mater. Contin., vol. 56, no. 3, pp. 365–379, 2018.
26. G. Zhang, C. Zhang, and H. Zhang, "Improved K-means algorithm based on density Canopy," Knowledge-Based Syst., vol. 145, pp. 289–297, 2018.
27. M. Ashkartizabi and M. Aminghafari, "Functional data clustering using K-means and random projection with applications to climatological data," Stoch. Environ. Res. Risk Assess., vol. 32, no. 1, pp. 83–104, 2018.
28. S. Y. Huang and B. Zhang, "Research on improved k-means clustering algorithm based on hadoop platform," Proc. - 2019 Int. Conf. Mach. Learn. Big Data Bus. Intell. MLBDBI 2019, pp. 301–303, 2019.
29. R. A. Haraty, M. Dimishkieh, and M. Masud, "An Enhanced k-Means Clustering Algorithm for Pattern Discovery in Healthcare Data," Int. J. Distrib. Sens. Networks, vol. 11, no. 6, p. 615740, 2015. [
30. X. Hou, "An Improved K-means Clustering Algorithm Based on Hadoop Platform," Advances in Intelligent Systems and Computing, vol. 928. pp. 1101–1109, 2020.
31. S. Dhanasekaran, R. Sundarrajan, B. S. Murugan, S. Kalaivani, and V. Vasudevan, "Enhanced Map Reduce Techniques for Big Data Analytics based on K-Means Clustering," IEEE Int. Conf. Intell. Tech. Control. Optim. Signal Process. INCOS 2019, pp. 0–4, 2019.
32. X. Wei and Y. Li, "Research on improved k-means algorithm based on hadoop," Proc. - 2017 4th Int. Conf. Inf. Sci. Control Eng. ICISCE 2017, pp. 593–598, 2017.
33. K. Wu, W. Zeng, T. Wu, and Y. An, "Research and improve on K-means algorithm based on hadoop," Proc. IEEE Int. Conf. Softw. Eng. Serv. Sci. ICSESS, vol. 2015- Novem, pp. 334–337, 2015.P. Rai, and S. Sing, "A survey of clustering techniques", International Journal of computer Applications, vol. 7(12), pp.1-5, 2010.
34. T. Soni Madhulatha, "An overview on clustering methods", IOSR Journal of engineering, vol. 2(4), pp.719-725, 2012.
35. K. A. Abdul Nazeer, M. P. Sebastian, Improving the Accuracy and Efficiency of the kmeans Clustering Algorithm, Proceedings of the World Congress on Engineering 2009 Vol I WCE 2009, July 1 - 3, 2009, London, U.K

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 6, June 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1006109 |

36. Soumi Ghosh, Sanjay Kumar Dubey, Comparative Analysis of K-Means and Fuzzy CMeans Algorithms, International Journal of Advanced Computer Science and Applications, Vol. 4, No.4, 2013
37. Shafeeq, A., Hareesha ,K., Dynamic Clustering of Data with Modified K-Means Algorithm, International Conference on Information and Computer Networks, vol. 27 ,2012
38. Junatao Wang, XiaolongSu, An Improved Kmeans Clustering Algorithm, Communication Software and Networks (ICCSN), 2011 IEEE 3rd International Conference on 27 may,2011 (pp. 44- 46)

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor: 8.165

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details