Educational Administration: Theory and Practice
2024,30(4), 1806-1813
ISSN:2148-2403
https://kuey.net/

Research Article

Study Of Existing Methods & Techniques Of K-Means Clustering

Sonia Yadav[1*,2], Dr. Sachin Sharma[3]

[1] *Research Scholar, School of Computer Applications, Manav Rachna International Institute of Research and Studies (MRIIRS), Faridabad, India.
[2] Department of Computer Science, Deshbandhu College, University of Delhi, New Delhi.
[3] Associate Professor, Faculty of Computer Applications, Manav Rachna International Institute of Research and Studies (MRIIRS), Faridabad, India.

Citation: Sonia Yadav et al., Study Of Existing Methods & Techniques Of K-Means Clustering, *Educational Administration: Theory And Practice*, *30*(4), 1806-1813 Doi: 10.53555/kuey.v30i4.1755

ARTICLE INFO

ABSTRACT

In the field of data mining, clustering is the technique of grouping millions of data points to form clusters. Data of the same class are grouped together. K-Means clustering is the most important and basic clustering technique for analyzing data points. K-means is the most widely used algorithm for clustering using a known set of medians. In the past, various efforts have been made to improve the performance of the k-means algorithm. Improvements in k-means significantly improve performance for small to medium-sized data. However, for big and very large amounts of data, k-means lags. This study explores and reviews existing techniques for adapting and developing data grouping methodologies for clustering k-devices.

Keywords: K-Means, K-Means algorithm, Nearest Neighbour, KDD, Clustering.

1. Introduction

Clustering is the process of reducing the amount of data by grouping data items so that the objects in the same cluster are similar to each other. Clustering is the most popular unsubstantiated and research data analysis. Clustering involves grouping data objects according to a certain degree of similarity. The primary purpose of grouping is to extract useful information and trends from raw data. Clustering is one of the possible solutions for data-driven decision making and to extract unknown models by grouping similar objects that ultimately reduces the time required to make real-time decisions. An effective clustering algorithm must focus on two issues: to increase the similarity between objects in a given database and to keep objects assigned to different clusters as different.

Clustering is the oldest data research technique in data analysis. Clustering is the process of examining a collection of objects or data points and grouping these points or data objects, based on a certain distance measure. The purpose of the grouping is to have a minimum distance from each other within the same cluster. Traditional clustering algorithms generally deal with low dimensional data. As the size of data for organizations such as climate, health care, and web documents has grown exponentially, the need for an effective grouping technique has arisen. The technique of clustering large data packets plays a key role in analytics. K-mean clustering is popular among the various clustering methods. The purpose of K-means is to find clusters "k" of a given database. These flocks are characterized by a similarity measure based on their meter distance. An efficient clustering method depends on how closely each object in the cluster is connected.

2. Related Studies

“K. A. Abdul Nazeer et al. proposes k-means algorithm, for different sets of values of initial centroids, produces different clusters. Final cluster quality in algorithm depends on the selection of initial centroids. Two phases include in original k means algorithm: first for determining initial centroids and second for assigning data points to the nearest clusters and then recalculating the clustering mean”.

“Soumi Ghosh et al. present a comparative discussion of two clustering algorithms namely centroid based K-Means and representative object based FCM (Fuzzy C-Means) clustering algorithms. This discussion is on the basis of performance evaluation of the efficiency of clustering output by applying these algorithms”.

“Junatao Wang et al. propose an improved k-means algorithm using noise data filter in this paper. The shortcomings of the traditional k-means clustering algorithm are overcome by this proposed algorithm. The algorithm develops density-based detection methods based on characteristics of noise data where the discovery and processing steps of the noise data are added to the original algorithm. By pre-processing the data to exclude these noise data before clustering data sets the cluster cohesion of the clustering results is improved significantly and the impact of noise data on k-means algorithm is decreased effectively and the clustering results are more accurate".

"Shi Na et al. present the analysis of shortcomings of the standard k-means algorithm. As k-means algorithm has to calculate the distance between each data object and all cluster centres in each iteration. This repetitive process effects the efficiency of clustering algorithm. An improved k-means algorithm is proposed in this paper. A simple data structure is required to store some information in every iteration which is to be used in the next iteration. Computation of distance in each iteration is avoided by the proposed method and saves the running time".

"Jancey proposed a variant which is a modification for the Forgy's k-means algorithm (cf. Anderberg, which is expected to accelerate convergence and inferior local minima. In this variant, the new cluster centre is not the mean of the old and added points, but the new centre is updated by reflecting the old center through the mean of the new cluster. In order to avoid poor local solutions, a number of genetic algorithm-based methods have been developed".

"Likas et al. developed the global k-means clustering algorithm which is a deterministic and incremental global optimization method. It is also independent on any initial parameters and employs k-means procedure as a local search procedure, since the exhaustive global k-means method is computationally expensive".

"Faber proposed a variant of the Lloyd's k-means algorithm called the continuous k-means algorithm. The reference points in the continuous k-means algorithm are chosen as a random sample from the whole population of the data point while in the standard k-means algorithm, the initial reference points are chosen more or less arbitrarily. During the update process, the continuous k-means algorithm examines only a random sample of the data points while the standard k-means algorithm examines all of the data set in sequence. If the data set is very large and the sample is a representative of the data set, then the continuous k-means algorithm should converge much faster than the algorithm that examines every point in sequence".

"Kanungo et al. presented a simple and efficient implementation of Lloyd's k-means clustering algorithm which they called the filtering algorithm. The filtering algorithm is easy to implement which requires a kd-tree as the only major data structure. A kd-tree is a binary tree, which represents a hierarchical sub-division of the point set's bounding box using axis aligned splitting hyperplanes. Each node of the kd-tree is associated with a closed box, called a cell. The root's cell is the bounding box of the point set. If the cell contains at most one point (or, more generally, fewer than some small constant), then it is declared to be a leaf. Otherwise, the root's cell is splitting into two hyper-rectangles by an axis orthogonal hyperplane. The points of the cell are then partitioned to one side or the other of this hyperplane. The resulting sub-cells are the children of the original cell, thus leading to a binary tree structure".

"Bagirov and Mardaneh proposed a new variant of the global k-means algorithm which is known as the Modified Global K-means (MGKM) algorithm because it is said to be effective for solving clustering problems in gene expression data sets. In their algorithm, a starting point for the cluster center is computed by minimizing the so-called auxiliary cluster function. The effectiveness of this algorithm highly depends on its starting point. The algorithm computes clusters incrementally and to compute k-partition of a data set, it uses cluster centers".

"Nazeer and Sebastian discussed in their paper about one major drawback of k-means algorithm, they proposed an enhanced method that deals with improving the accuracy and efficiency of k-means algorithm. Both the phases of the original k-means algorithm were modified. The initial centroids are determined systematically so as to produce clusters with better accuracy in the first phase". The second phase makes use of a variant of the clustering method discussed in Fahim et al. It starts by forming the initial clusters based on the relative distance of each data point from the initial centroids. These clusters are subsequently fine-tuned by using a heuristic approach there by improving the efficiency.

3. K-means Algorithms and its Application

Cluster analysis can be divided into hierarchical clustering methods and non-hierarchical clustering methods. Examples of stratification techniques include single link, common link, intermediate link, median, and word. Non-hierarchical methods include k-mean, adaptive k-mean, k-medoid, and fuzzy grouping. Which algorithm is appropriate depends on the type of data available and the specific purpose of the analysis. More objectively, cluster stability can be examined by simulation studies. The problem of choosing the "best" algorithm/setting parameters is difficult. A good clustering algorithm should ideally create groups with separate overlapping boundaries, but in practice perfect separation cannot be achieved. The idea of this validation approach is to reward the sequence his algorithm. In this work, we implemented the traditional k-means clustering algorithm and chose Euclidean similarity of distances to analyze the problem.

Data retrieval is the process of automatically searching or discovering useful knowledge. This process uses mathematically-based algorithms and statistical methodologies to extract data from databases to reveal unknown samples of data that may be of useful information. The information obtained during data extraction is very important knowledge that helps users to make appropriate business strategies. These processes are also known as “Knowledge Discovery in Databases (KDD)”, in the sense that much information and raw data can be used in databases to discover and analyze knowledge. Knowledge can be used in decision support systems, to predict customer behaviour, and to predict future product sales percentages.

This study explores various techniques for adapting and developing data grouping methodologies for clustering k-devices. The K-Means algorithm, which is based on partitioning, is a type of cluster algorithm and is recommended by J.B. MacQueen. Problems with grouping data with k-devices are the choice of start centres. Research focuses on the development of a method for clustering a k-device for centride selection. In this paper, we presented the basic idea of the technique for retrieving data by grouping data from raw data by selecting the appropriate starting centre. The techniques used for clustering in this document are the k-mean clustering method.

Data clustering is the processing of raw data to find clusters or groups of similar data. In each cluster, members have some similarities in data type. The principles of data clustering are to find the value of a similar result and to attribute each member to the same group of other members who have a similar or the same result. The technique for retrieving data when finding data clusters differs from data classification in that the user does not have to specify the target characteristic for assigning each data record to the appropriate cluster. Therefore, data clustering is a method of unattended training. The clustering method relies on the measurement of similarity automatically by groups of relevant or similar data members, as visually shown in Figure 1. After the clustering process, the user can apply some sorting algorithm to extract a data model in each cluster to better understand the cluster model.

K-means clustering algorithm is the most chosen data clustering technique. K-means is non-hierarchical clustering and the use of cycles to group data into K groups. K-clustering of K-means starts the iterative process by finding the starting centroid or the center point of each group, randomly selecting representative data from raw data to be the centroid in each group of K data. Then assign each information to the nearest group, calculating the Euclidean distance between each data record to each centroid to distribute the data record to the nearest group. Each cluster will then find a new centroid to replace the original one and repeat the steps of calculating the Euclidean distance to group data members and send each member in the group to the nearest centroid. The process will stop when each group has a stable centroid and the members do not change their groups.

The steps of k-means algorithm can be summarized as the following:
1) Specify group number and select initial centroid of each group.
2) Calculate Euclidean distance for each data member and centroid to assign members to the nearest centroid.
3) Calculate distance’s mean of every data member and own centroid to define new centroid in each group.
4) Repeat steps 2 and 3 until each group has stable centroid or same centroid.

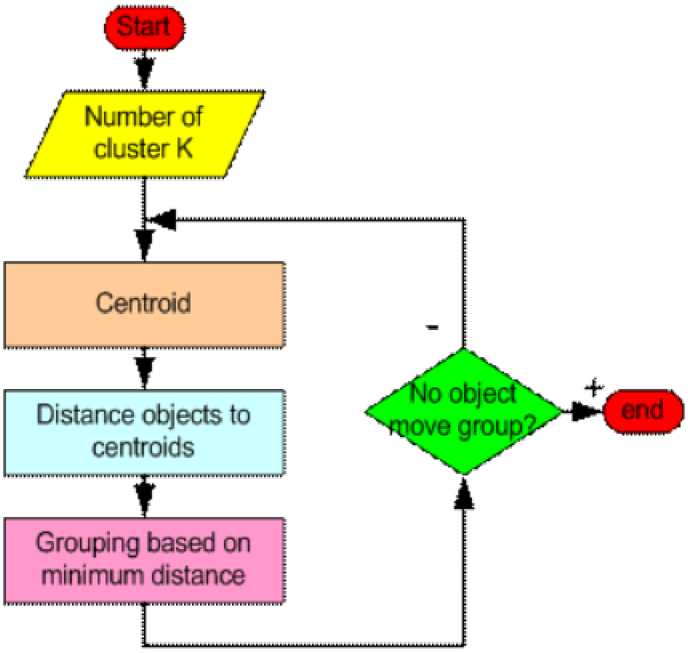

Start
Number of cluster K
Centroid
Distance objects to centroids
Grouping based on minimum distance
No object move group?
-
+
end

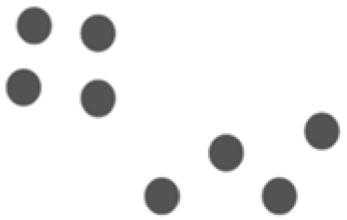
Fig.1. K-Means Generic Algorithm

K-means clustering intends to divide n objects into k clusters in which each object belongs to the closest cluster. This method produces exactly k different clusters with the greatest possible difference. The best number of clusters leading to the largest separation (distance) is not known as a priority and must be calculated from the data. The purpose of the K-Asset Clustering is to minimize the total variance within the cluster or the square error function:

number of clusters number of cases case *i* centroid for cluster *j*

$$\text{objective function} \leftarrow J = \sum_{j=1}^{k}\sum_{i=1}^{n} \underbrace{\left\| x_i^{(j)} - c_j \right\|^2}_{\text{Distance function}}$$

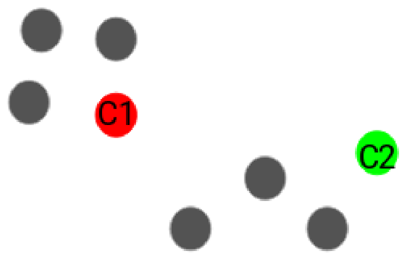
Let's take an example to understand how K-Means really works:



We have these 8 points and we want to apply k-means to create clusters for these points.

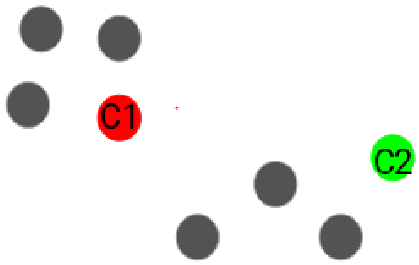
Step 1: Choose the number of clusters *k*
The first step in k-means is to pick the number of clusters, k.

Step 2: Select k random points from the data as centroids
Next, we randomly select the centroid for each cluster. Let's say we want to have 2 clusters, so k is equal to 2 here. We then randomly select the centroid:



Here, the red and green circles represent the centroid for these clusters.

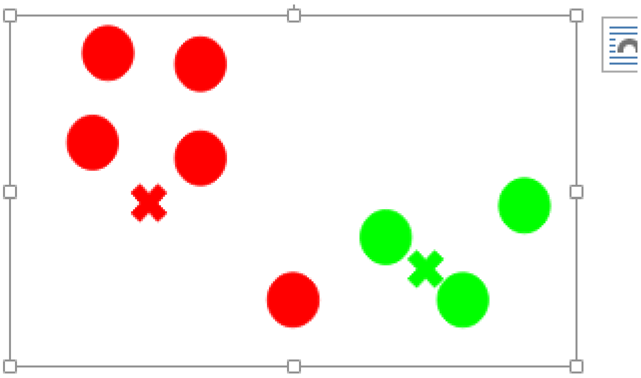
Step 3: Assign all the points to the closest cluster centroid
Once we have initialized the centroids, we assign each point to the closest cluster centroid:



Here you can see that the points which are closer to the red point are assigned to the red cluster whereas the points which are closer to the green point are assigned to the green cluster.

Step 4: Recomputed the centroids of newly formed clusters

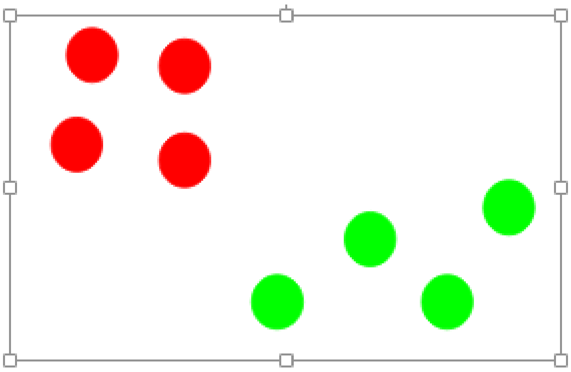
Now, once we have assigned all of the points to either cluster, the next step is to compute the centroids of newly formed clusters:



Here, the red and green crosses are the new centroids.

Step 5: Repeat steps 3 and 4

We then repeat steps 3 and 4:



"The step of computing the centroid and assigning all the points to the cluster based on their distance from the centroid is a single iteration. "But wait – when should we stop this process?

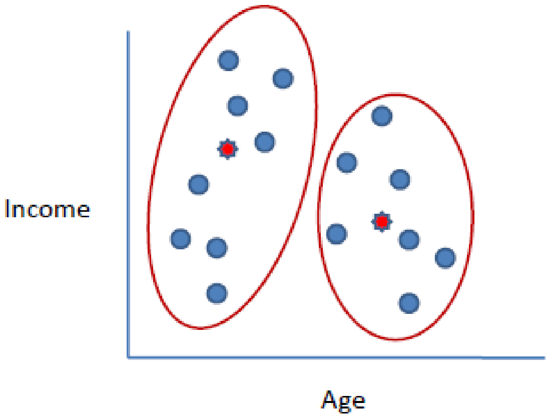
Stopping Criteria for K-Means Clustering

There are basically three stop criteria that can be adopted to stop the K-mean algorithm:
1. The centroids of the newly formed clusters do not change
2. The dots remain in the same cluster
3. The maximum number of repetitions is achieved

We can stop the algorithm if the centroids of the newly formed clusters do not change. Even after multiple repetitions, if we get the same centres for all clusters, we can say that the algorithm does not learn any new pattern and that is a sign to stop training.

Another clear sign that we need to stop the training process if the points remain in the same cluster even after training the multiple repetition algorithm.

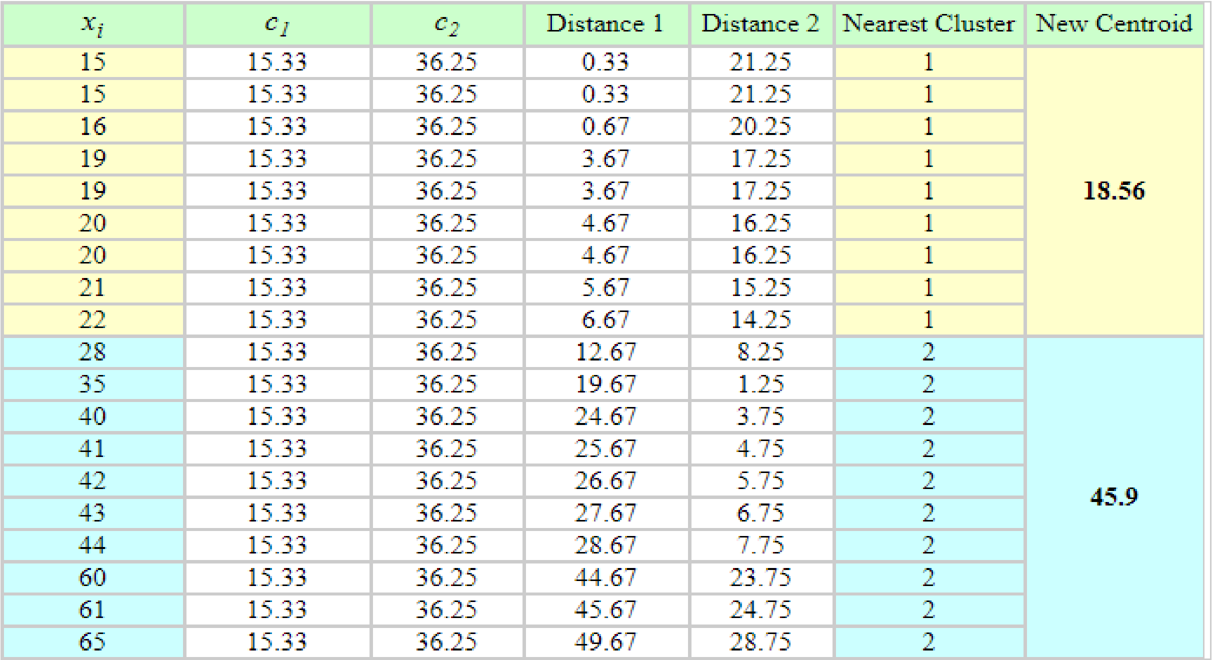
Finally, we can stop training if the maximum number of repetitions is achieved. Suppose we set the number of repetitions to 100. The process will be repeated for 100 repetitions before it stops.



Example:
Suppose we want to group the visitors to a website using just their age (one-dimensional space) as follows:
n = 19
15,15,16,19,19,20,20,21,22,28,35,40,41,42,43,44,60,61,65

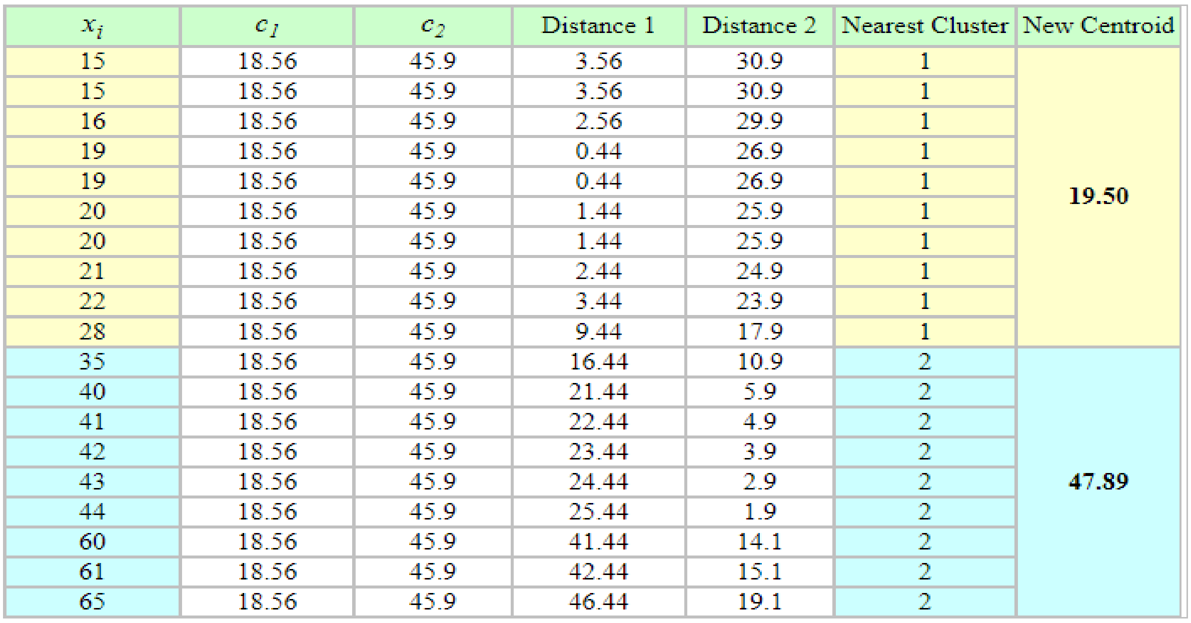
Iteration 1:
c1=15.33
c2=36.25

Iteration 2:
c1=18.56
c2=45.90



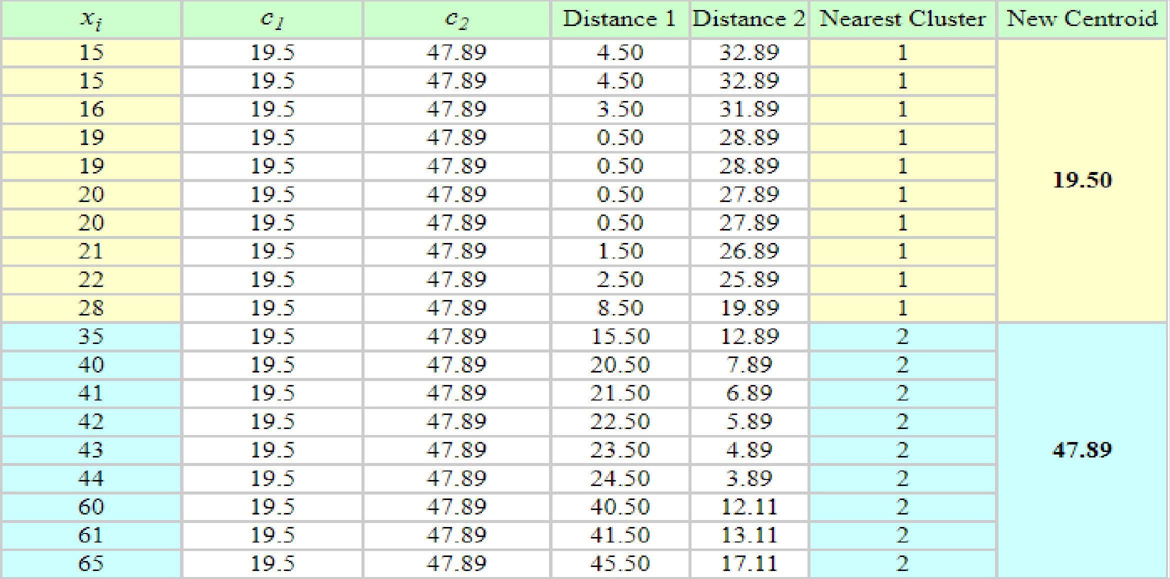
x_i	c_1	c_2	Distance 1	Distance 2	Nearest Cluster	New Centroid
15	15.33	36.25	0.33	21.25	1	**18.56**
15	15.33	36.25	0.33	21.25	1	
16	15.33	36.25	0.67	20.25	1	
19	15.33	36.25	3.67	17.25	1	
19	15.33	36.25	3.67	17.25	1	
20	15.33	36.25	4.67	16.25	1	
20	15.33	36.25	4.67	16.25	1	
21	15.33	36.25	5.67	15.25	1	
22	15.33	36.25	6.67	14.25	1	
28	15.33	36.25	12.67	8.25	2	**45.9**
35	15.33	36.25	19.67	1.25	2	
40	15.33	36.25	24.67	3.75	2	
41	15.33	36.25	25.67	4.75	2	
42	15.33	36.25	26.67	5.75	2	
43	15.33	36.25	27.67	6.75	2	
44	15.33	36.25	28.67	7.75	2	
60	15.33	36.25	44.67	23.75	2	
61	15.33	36.25	45.67	24.75	2	
65	15.33	36.25	49.67	28.75	2	

Iteration 3:
c1=19.50
c2=47.89



x_i	c_1	c_2	Distance 1	Distance 2	Nearest Cluster	New Centroid
15	18.56	45.9	3.56	30.9	1	**19.50**
15	18.56	45.9	3.56	30.9	1	
16	18.56	45.9	2.56	29.9	1	
19	18.56	45.9	0.44	26.9	1	
19	18.56	45.9	0.44	26.9	1	
20	18.56	45.9	1.44	25.9	1	
20	18.56	45.9	1.44	25.9	1	
21	18.56	45.9	2.44	24.9	1	
22	18.56	45.9	3.44	23.9	1	
28	18.56	45.9	9.44	17.9	1	
35	18.56	45.9	16.44	10.9	2	**47.89**
40	18.56	45.9	21.44	5.9	2	
41	18.56	45.9	22.44	4.9	2	
42	18.56	45.9	23.44	3.9	2	
43	18.56	45.9	24.44	2.9	2	
44	18.56	45.9	25.44	1.9	2	
60	18.56	45.9	41.44	14.1	2	
61	18.56	45.9	42.44	15.1	2	
65	18.56	45.9	46.44	19.1	2	

Iteration 4:
c1=19.50
c2=47.89



x_i	c_1	c_2	Distance 1	Distance 2	Nearest Cluster	New Centroid
15	19.5	47.89	4.50	32.89	1	**19.50**
15	19.5	47.89	4.50	32.89	1	
16	19.5	47.89	3.50	31.89	1	
19	19.5	47.89	0.50	28.89	1	
19	19.5	47.89	0.50	28.89	1	
20	19.5	47.89	0.50	27.89	1	
20	19.5	47.89	0.50	27.89	1	
21	19.5	47.89	1.50	26.89	1	
22	19.5	47.89	2.50	25.89	1	
28	19.5	47.89	8.50	19.89	1	
35	19.5	47.89	15.50	12.89	2	**47.89**
40	19.5	47.89	20.50	7.89	2	
41	19.5	47.89	21.50	6.89	2	
42	19.5	47.89	22.50	5.89	2	
43	19.5	47.89	23.50	4.89	2	
44	19.5	47.89	24.50	3.89	2	
60	19.5	47.89	40.50	12.11	2	
61	19.5	47.89	41.50	13.11	2	
65	19.5	47.89	45.50	17.11	2	

No changes were observed between repetitions 3 and 4. Using grouping, 2 groups 15-28 and 35-65 were identified. The initial selection of centroids can affect the output herds, so the algorithm is often performed multiple times with different initial conditions in order to obtain a fair view of what clusters should be.

4. Conclusion

The K-Means algorithm, which is not controlled, is commonly used in data retrieval and pattern recognition. The goal of minimizing the cluster performance index, the quadratic error, and the error criterion are the foundations of this algorithm. To look for the optimizing result, this algorithm tries to find K divisions to satisfy a certain criterion. First, select some points to represent the initial focal points of the cluster (usually we choose the first K sample points of income to represent the initial focal point of the cluster); second, we add the remaining sample points to their focal points according to the minimum distance criterion, then we get the initial classification, and if the classification, if unreasonable, we modify it (recalculate all focal points of the cluster), we repeat until we get reasonable classification. The K-Means algorithm, based on separation, is a kind of cluster algorithm and has the advantages of brevity, efficiency and purity. However, this algorithm depends a lot on the initial points and the difference in the choice of initial samples, which always leads to different results. Moreover, this algorithm, based on the objective function, always uses a gradient method to obtain extremism. The direction of search in the gradient method is always in the direction in which the energy decreases, which will lead to the fact that when the initial focal point of the cluster is not correct and then the whole algorithm will easily sink to the local minimum point.

In this paper k-means clustering techniques and method are reviewed. K-means being most famous among data scientist need further improvement in various section of algorithm. The outliers, empty clusters and selecting centroid for datasets are still a challenging task. Hence various further research needed to focus on these mentioned issues. Table I. presents various techniques and its limitation are present in proposed kmeans algorithm. They need further enhancement due to increase of size of data as of now. This paper has made an attempt to review a significant number of papers to deal with the present algorithm of k-means. Present study illustrates that k-means algorithm can be enhanced by selecting centroid point appropriately.

References

1. Amber Abernathy, M. Emre Celebi,The incremental online k-means clustering algorithm and its application to color quantization,Expert Systems with Applications,Volume 207,2022,117927,ISSN 0957-4174,https://doi.org/10.1016/j.eswa.2022.117927. (https://www.sciencedirect.com/science/article/pii/S0957417422011708)
2. Laith Mohammad Abualigah, Ahamad Tajudin Khader, Essam Said Hanandeh, A new feature selection method to improve the document clustering using particle swarm optimization algorithm, Journal of Computational Science,Volume 25,2018,Pages 456-466, ISSN 1877-7503, https://doi.org/10.1016/j.jocs.2017.07.018. (https://www.sciencedirect.com/science/article/pii/S1877750316305002)
3. Rasim M. Alguliyev, Ramiz M. Aliguliyev, Lyudmila V. Sukhostat,Parallel batch k-means for Big data clustering,Computers & Industrial Engineering,Volume 152,2021,107023,ISSN 0360-8352,https://doi.org/10.1016/j.cie.2020.107023. (https://www.sciencedirect.com/science/article/pii/S0360835220306938)

4. Liang Bai, Jiye Liang, Fuyuan Cao, A multiple k-means clustering ensemble algorithm to find nonlinearly separable clusters, Information Fusion,Volume 61,2020,Pages 36-47,ISSN 1566-2535,https://doi.org/10.1016/j.inffus.2020.03.009. (https://www.sciencedirect.com/science/article/pii/S1566253518305153)
5. Asma Belhadi, Youcef Djenouri, Kjetil Nørvåg, Heri Ramampiaro, Florent Masseglia, Jerry Chun-Wei Lin, Space–time series clustering: Algorithms, taxonomy, and case study on urban smart cities, Engineering Applications of Artificial Intelligence,Volume 95,2020,103857, ISSN 0952-1976, https://doi.org/10.1016/j.engappai.2020.103857. (https://www.sciencedirect.com/science/article/pii/S0952197620302141)
6. Li Chen, Shuisheng Zhou, Jiajun Ma, Mingliang Xu,Fast kernel k-means clustering using incomplete Cholesky factorization, Applied Mathematics and Computation,Volume 402, 2021,126037,ISSN 0096-3003,https://doi.org/10.1016/j.amc.2021.126037. (https://www.sciencedirect.com/science/article/pii/S0096300321000850)
7. Fatéma Zahra Benchara, Mohamed Youssfi, A new scalable distributed k-means algorithm based on Cloud micro-services for High-performance computing, Parallel Computing,Volume 101,2021,102736,ISSN 0167-8191,https://doi.org/10.1016/j.parco.2020.102736. (https://www.sciencedirect.com/science/article/pii/S0167819120301186)
8. Ioan-Daniel Borlea, Radu-Emil Precup, Alexandra-Bianca Borlea, Daniel Iercan, A Unified Form of Fuzzy C-Means and K-Means algorithms and its Partitional Implementation, Knowledge-Based Systems, Volume 214, 2021, 106731,ISSN 0950-7051, https://doi.org/10.1016/j.knosys.2020.106731. (https://www.sciencedirect.com/science/article/pii/S0950705120308601)
9. Rasim M. Alguliyev, Ramiz M. Aliguliyev, Lyudmila V. Sukhostat, Parallel batch k-means for Big data clustering, Computers & Industrial Engineering, Volume 152, 2021, 107023, ISSN 0360-8352, https://doi.org/10.1016/j.cie.2020.107023. (https://www.sciencedirect.com/science/article/pii/S0360835220306938)
10. Xiaohui Chen, Yun Yang, Diffusion K-means clustering on manifolds: Provable exact recovery via semidefinite relaxations, Applied and Computational Harmonic Analysis, Volume 52, 2021, Pages 303-347, ISSN 1063-5203, https://doi.org/10.1016/j.acha.2020.03.002. (https://www.sciencedirect.com/science/article/pii/S106352032030021X)