Study of Centroid initialization in K means clustering algorithm

A DISSERTATION

SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS

FOR THE AWARD OF DEGREE

OF

MASTER OF TECHNOLOGY

IN

SOFTWARE ENGINEERING

Submitted by:

ASHUTOSH

2K20/SWE/06

Under the supervision of

Dr. Sonika Dahiya

(Assistant Professor)

DEPARTMENT OF SOFTWARE ENGINEERING

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Bawana Road, Delhi-110042

MAY, 2022

DEPARTMENT OF SOFTWARE ENGINEERING

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Bawana Road, Delhi - 110042

CANDIDATE'S DECLARATION

I, Ashutosh, Roll No. 2K20/SWE/06 student of M. Tech (Software Engineering), hereby declare that the project Dissertation titled "Study of centroid initialization in K means clustering algorithm" which is submitted by me to the Department of Software Engineering, Delhi Technological University, Delhi in partial fulfillment of the requirement for the award of the degree of Master of Technology, is original and not copied from any source without proper citation. This work has not previously formed the basis for the award of and Degree, Diploma Associateship, Fellowship or other similar title or recognition.

Ashutosh.

Place: Delhi

Date:

Ashutosh

2K20/SWE/06

DEPARTMENT OF SOFTWARE ENGINEERING

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Bawana Road, Delhi - 110042

CERTIFICATE

I hereby certify that the Project Dissertation titled **"Study of centroid initialization in K means clustering algorithm"** which is submitted by Ashutosh, 2K20/SWE/06 Department of Software Engineering, Delhi Technological University, Delhi in partial fulfillment of the requirement for the award of the degree of Master of Technology, is a record of the project work carriedout by the student under my supervision. To the best of my knowledge, this workhas not been submitted in part or full for any Degree or Diploma to this Universityor elsewhere.

Sonika 30/05/2022

Place: Delhi

Date:

Dr. Sonika Dahiya

Assistant Professor

Department of SWE

ACKNOWLEDGMENT

The success of this project requires the assistance and input of numerous people and the organization. I am grateful to everyone who helped in shaping the result of the project.

I express my sincere thanks to **Dr. Sonika Dahiya,** my project guide, for providing me with the opportunity to undertake this project under her guidance. Her constant support and encouragement have made me realize that it is the process of learning which weighs more than the end result. I am highly indebted to the panel faculties during all the progress evaluations for their guidance, constant supervision and for motivating me to complete my work. They helped me throughout with new ideas, provided information necessary and pushed me to complete the work.

I also thank all my fellow students and my family for their continued support.

ASHUTOSH

2K20/SWE/06

ABSTRACT

K-mean Clustering is a feature-based feature detection and similarity grouping approach. Massive datasets are no problem for this approach. The K-means clustering outcome is influenced by the initial points in the optimization process. Cluster centres should be chosen at random for each cluster. These centres should be as widely apart as feasible. The clustering process and outcomes are influenced by the starting points used. Effective cluster assignment is made possible in large part by the Centroid initialization. As a result of the initial centroid values assigned, the clustering convergence behavior is also reliant on these values. In order to improve the clustering performance of the K-Means clustering method, this work focuses on the assignment of cluster centroid selection. An initial cluster centroid is assigned by using centres derived from partitioning of data along the data axis with the highest variance, as described in this study. The experimental findings show that the suggested method is superior to the standard method in terms of clustering outcomes.

CONTENTS

LIST OF FIGURES

LIST OF ABBREVIATIONS

1. ML: Machine Learning
2. EF: Euclidean Formula
3. k-NN: k- Nearest Neighbor
4. KN: Clustering algorithm
5. KM++: K-means++
6. CI: Centroid index
7. RKM: Repeated k-means

CHAPTER 1

INTRODUCTION

1.1 Machine Learning:

ML is discipline that allows machines to learn and develop without having to be explicitly taught to do so. Unsupervised learning occurs when a machine learns to label data automatically without knowing the pattern of the data beforehand. Since it is complex to predict the pattern of data in advance, unsupervised learning is critical.

Clustering is an unsupervised learning strategy that separates a data set into groups with the goal of maximizing both the similarity of data points within the same group and the un-similarity of data points among groups. It makes the machine read, understand, and generate meaning for human language.

This discipline is growing as fast and much study is conducted in this area. Because it is skilled in doing tasks that are complex for a human to accomplish directly, machine learning is needed

There are a few different ways to define Machine Learning Algorithms, however they may generally be categorized into groups based on their purpose, with the following being the major:

- **Supervised learning**
- **Unsupervised Learning**
- **Semi-supervised Learning**
- **Reinforcement Learning**

1.2 Classification of machine learning algorithm

1.2.1 Supervised Learning

Here models are trained using a labelled dataset, in which the model learns about every type of information. Output is generated after the model is evaluated using test data.
It is done after the training phase.

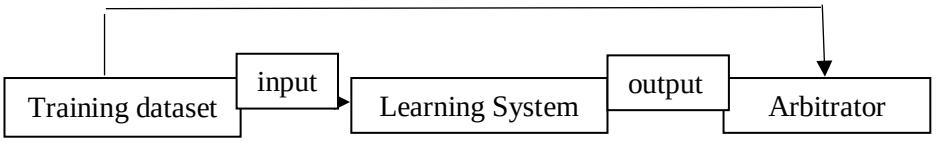

Fig. 1 Supervised learning block diagram

1.2.2 Unsupervised learning

It is a technique that does not use a training dataset to supervise models. whereas models employ data to identify previously unknown patterns and insights. It's similar to the learning that takes place in the human brain as we learn new stuff.

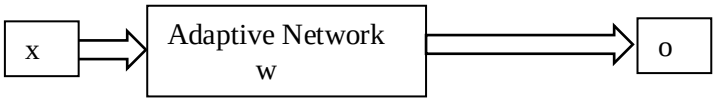

Fig. 2 Unsupervised learning block diagram

1.2.3 Semisupervised learning

The system is taught using both labelled and unlabeled data in this sort of learning. This blend will often have a small bit of labelled data and a significant amount of unlabeled data. The basic technique entails the programmer clustering similar data using an ULA before labelling the remaining unlabelled data with the current labelled data. The primary application cases for this type of algorithm all have one thing in common: acquiring unlabeled data is reasonably cheap, whereas labelling it is quite expensive

1.2.3 Reinforcement learning

It differs from SL in that latter includes the solution key, so the model to be trained with the genuine result, whereas RL does not. Instead, the RL selects what to do to complete the task.

1.3 Clustering

Clustering is a type of unsupervised learning technique. In the literature clustering is of following four types:

- Density based method clustering
- Hierarchical based method clustering
- Partitioning based method clustering
- Grid-based method clustering

The density-based method considers the clusters as the dense region having some similarities and differences from the lower dense region of the space. Example is DBSCAN,OPTICS etc. In Hierarchical based method the clusters formed in this method form a tree-type structure based on the hierarchy, for example CURE and BIRCH. The partition based method partition the objects into k clusters and each partition forms one cluster. This method is used to optimize an objective criterion similarity function such as when the distance is a major parameter. Example include K-means, CLARANS. In grid based method all the data space is formulated into a finite number of cells that form a

grid-like structure. Example include STING ,CLIQUE etc.

In the thesis, various methods for the initialization of centroid in k means proposed in literature is studied and presented.

1.4 K-means clustering algorithm

The K-means algorithm divides N data points into k clusters by reducing the sum of squared distances between each point and its nearest centroid. Sum-of-squared errors is the objective function (SSE) for k-means. K-means is famous for a reason. To begin with, it is straightforward to execute. Next, individuals generally favor to employ a well-studied algorithm with well-known restrictions over a potentially improved but less premidated algorithm with unfamiliar or unseen constraints.

But with the advantages, k-means has disadvantages. The iterations of k-means can lead to inferior local minimum if the initiation is poor.

Several initialization approaches have been suggested to solve this problem.

The three approaches to improve k-means are:

- Better Initialization

Approaches for selection of centroid in initial phase is used. Example: Random centroid, Maxmin, sorting heuristic etc.

- Repeated k-means(RKM)

k-means is repeated some specified number of times to reduce the error. Example: Steinley's algorithm etc.

- Replace k-means by another better algorithm

Some standalone better approaches are used to replace k-means.

In this paper various algorithms proposed for better initialization of centroid in k-means are studied.

The k-means clustering algorithm primarily accomplishes two goals:

- Iteratively determines the optimal value for K centre points or centroids.

- Assigns each data point to its closest k-center. A cluster is formed by data points that are close to a specific k-center.

  As a result, every cluster has points with some similarities and is isolated from the others.

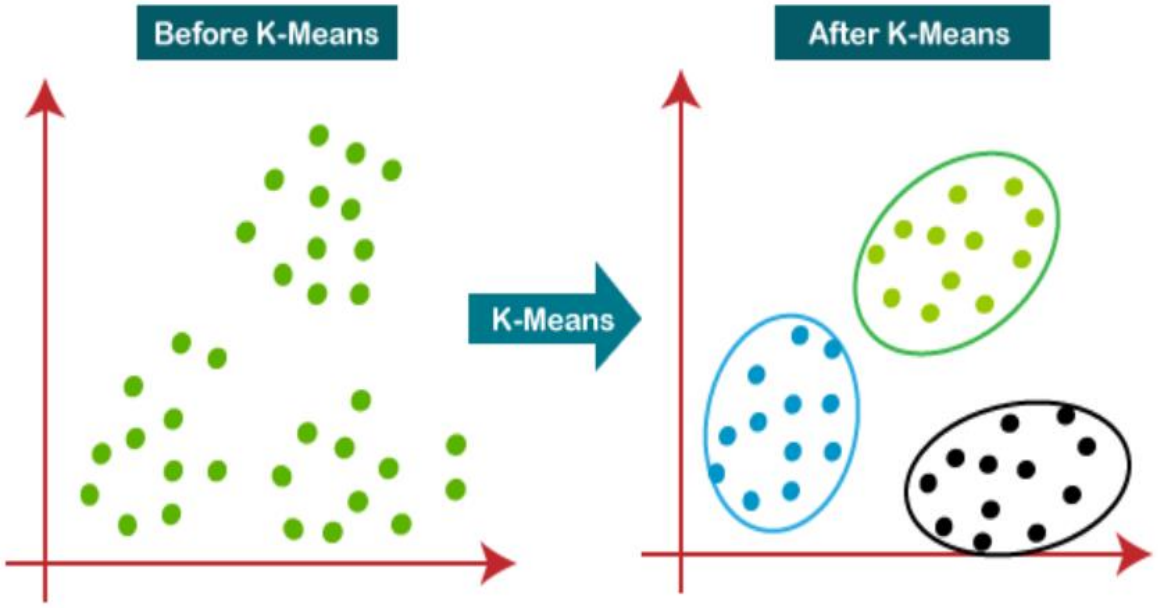

Fig. 3 Working of K-means clustering algorithm

The K-means algorithm divides N data points into k clusters by reducing the sum of squared distances between each point and its nearest centroid. Sum-of-squared errors is the objective function (SSE) for k-means. K-means is famous for a reason. To begin with, it is straightforward to execute. Next, individuals generally favor to employ a well-studied algorithm with well-known restrictions over a potentially improved but less premidated algorithm with unfamiliar or unseen constraints.

Some of the concerns with the K-means clustering algorithm have been addressed in previous studies. However, they did not address the K-means clustering algorithm's drawbacks in a unified approach.

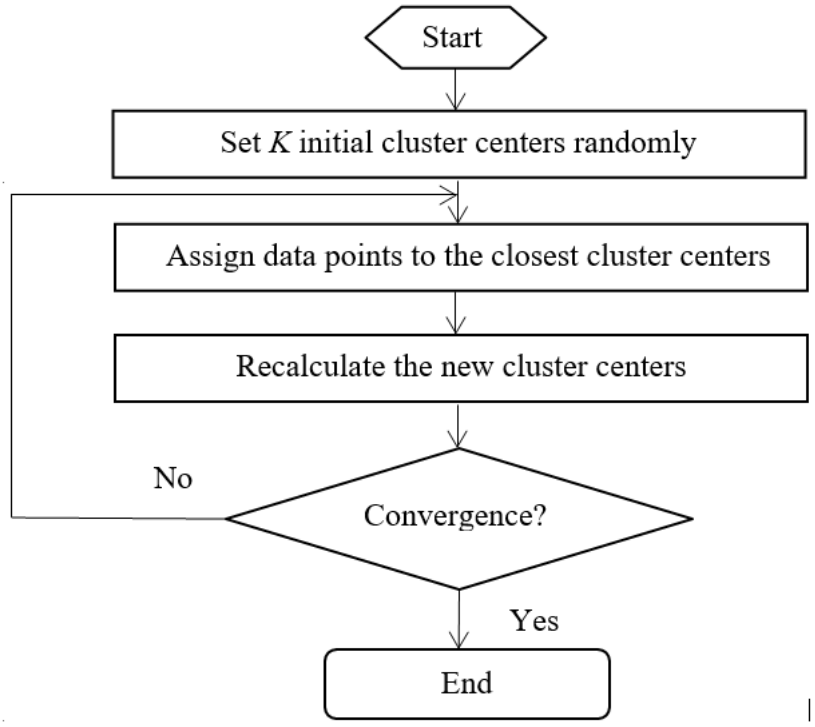

Fig. 4 K means clustering flow

1.5 Initialization of centroid in k-means

Even though k-means is a simple clustering approach, it is usually quite effective. Using k random data points from a particular dataset as centroids, all training instances are shown and allocated to the nearest cluster. When all of the clusters are brought together, the recalculated centroids indicate the average value for each cluster, and these recalculated centroids become the new centres of their respective clusters.

When a training set is re-plotted and added to a new cluster, all cluster memberships are reset. If and only if the centroids and cluster members remain constant, we can call the procedure iterative. When the recalculated centroids are within a predetermined margin of the preceding iteration's centroids, convergence has occurred. Convergence Distances are frequently calculated using the Euclidean k-means distance between two points in the form (x, y).

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (xi - yi)^2} \quad (1)$$

In the context of k-means clustering, it stands to reason that the more optimal the position of these initial centroids is, the lesser repetitions of the k-means CL will be necessary for junction. This tells that considerations during the planning stage could help with the initialization of these initial centroid.

1.6 Advantages of k-means:

- Relatively simple to implement
- Scales to large dataset
- Guarantees convergence
- Can warm-start the positions of centroids
- Easily adapts to new examples

- Generalizes to clusters of different shapes and sizes, such as elliptical clusters.

1.7 Disadvantages of k-means:

Along with the merits KM has demerits too.The iterations of k-means can lead to inferior local minimum if the initiation is poor.
The three approaches to improve k-means are:

- Better Initialization

Approaches for selection of centroid in initial phase is used. Example: Random centroid, Maxmin, sorting heuristic etc.

- Repeated k-means(RKM)

k-means is repeated some specified number of times to reduce the error. Example: Steinley's algorithm etc.

- Replace k-means by another better algorithm

Some standalone better approaches are used to replace k-means.

In the research paper various algorithms proposed for better initialization of centroid in k-means are studied

CHAPTER 2

LITERATURE SURVEY

A significant amount of work has been done in the field of initialization of centroid in k means clustering. Clustering with k-means is a basic but effective method. Typically, k points are chosen at random as cluster centres, or centroids, from a dataset, and all learning algorithms are charted and added to the cluster with the least training examples.. After all instances have been added to clusters, the centroids, which reflect the mean of each cluster's instances, are recalculated, and these recalculated centroids become the new cluster centres.

2.1 Centroid initialization method:

The three approaches to improve k-means are:

- Better Initialization

Approaches for selection of centroid in initial phase is used. Example: Random centroid, Maxmin, sorting heuristic etc.

- Repeated k-means(RKM)

k-means is repeated some specified number of times to reduce the error. Example: Steinley's algorithm etc.

- Replace k-means by another better algorithm

Some standalone better approaches are used to replace k-means.

In this paper various algorithms proposed for better initialization of centroid in k-means are studied.

Centroid initialization presented in this thesis are:

2.1.1 Random partitioning:

In this method each a point is placed within a randomly selected cluster. Centroid calculation is done after that. It leads to avoid outliers selection from border areas. However, the disadvantage is that the generated centroids are accumulated in the mean of data .

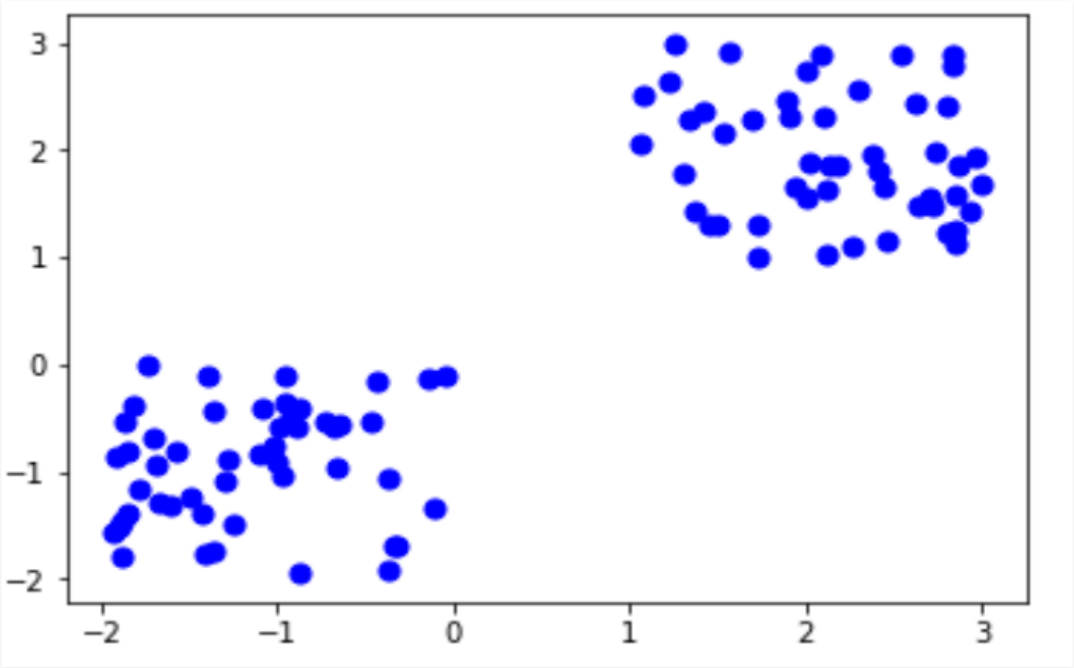

Fig. 5 Random Partitioning

Flowchart:

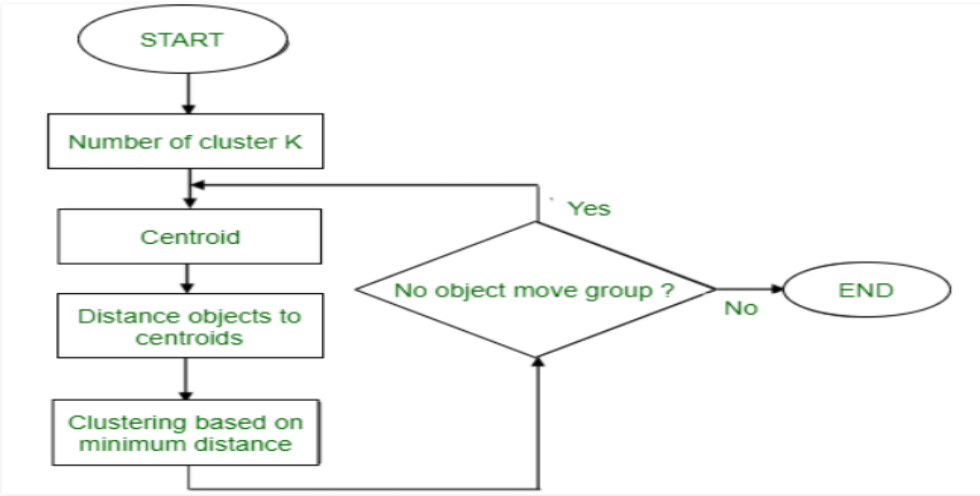

Fig. 6 Centroid initialization using random data points

2.1.2 Random centroid:

In this method k random points are chosen as started set of centroids. It ensures that each cluster has at least one point. Then position of each point is swapped with another randomly selected point.

It then takes first k points from array to avoid selection of k points twice.

For generating random number, random number generator is used. Hence this approach is called random centroids.

The time complexity of this approach is O(N).

2.1.3 Farthest point heuristic:

In this method a random point is selected as initial centroid. And then new centroids are added one after another. Next, at each stage centroid is point farthest(max) from its closest (min) current centroid. This has time complexity of

O(N). Some ways to select initial centroid are arbitrary, furthest pair of points, maximum distance

from origin, maximum density). This method is also known as Maxmin.

Maxmin prevents the worst-case scenario of random centroids, where worst case occurs when cluster size has serious unbalance.

2.1.4 K-means++:

This method is modified version of maxmin. In this method first centroid is chosen randomly and then next centroid is chosen using weighted probability

$$pi= cost_i/SUM(cost_i), \quad (2)$$

here $cost_i$ is squared distance of points x_i to its closest centroids. It has time complexity of O(logk).

2.1.5 Sorting heuristic:

This technique sorts points according to a few criteria. The criteria is based on the following points:

- Distance to center point
- Density
- Centrality
- Attribute with greatest variance

Here sorting takes time complexity of O(NlogN), Now k points are chosen from sorted list based on assumptions such as: 1^{st} k points, each $(N/k)^{th}$ point etc.
This method is good for well separated clusters.

2.1.6 Density based heuristic:

In this method density is utilized. Main problem associated to this algorithm is how to calculate density. For that following are the approaches:

- Buckets.
- Ɛ-radius circle.
- K-nearest neighbour.

In 1st technique , divide space by regular grid then count frequency of points in each bucket. Density of points is now taken from bucket.

For the other two methods density is evaluated for each point separately.

Once determined, the density can be combined with a nearest point heuristic, a sorting heuristic, or both.

2.1.7 Splitting Algorithm:

Every point is arranged into a single cluster by the Split algorithm .After then, one cluster at a time is divided unless k clusters are attained.. The selection of how to split and which cluster to split is to be taken into account.

Some methods used to split are mentioned. In the context of vector quantization, one way uses binary split to initialise their LBG algorithm.

There is also a split-kmeans variant that applies k-means iteration after every split in. In this paper simple variant is implemented where biggest cluster is selected to split. Two random points are chosen from cluster to split it. K-means clustering is used then within cluster

2.1.8 Repeated k-means:

Repeated k-means runs k-means in several instances, each time with a different

initialization, and keeps the outcome with the smallest SSE-value. Multi-start k-means is another term for this. The repeating done to increase the chances of success. The repetition number(R) to find accurate clustering is:

$$R=1/p \quad (3)$$

If the initialization procedure is deterministic (i.e., there is no variability), it either succeeds (p = 100%) or fails (p = 0%).

The time complexity of this method is $O(N^2)$.

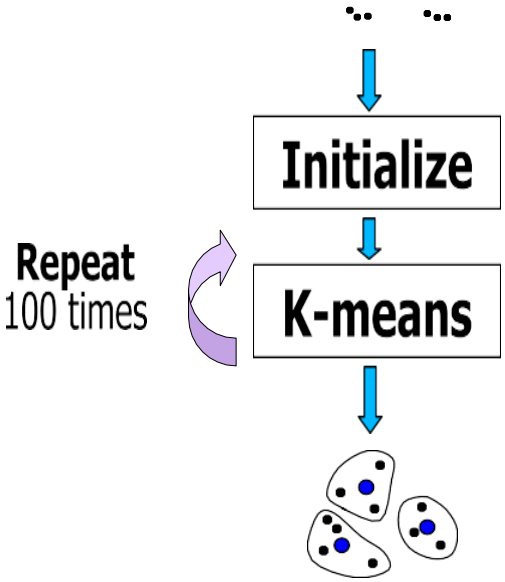

Fig. 7 Repeated k-means

2.1.9 Stenly's method:

This method repeats random centroid method 5000 times and then having smallest SSE value is chosen.

2.2 Data Set used in literature:

The analysis of initialization techniques was done by Pasi Franti and Sami Sieranoja for which

the data set was taken from: http://cs.uef.fi/sipu/datasets/.

Because SSE can appropriately group all of these datasets, a basic clustering benchmark was used. A brief description of dataset is given in table 1.

Table 1
Basic clustering benchmark

Dataset	Varying	Size	Dimensions	Clusters	Per cluster
A	Number of clusters	3000-7500	2	20-50	150
S	Overlap	5000	2	32-1024	333
Dim	Dimensions	1024	32-1024	16	64
G2	Dimensions + overlap	2048	2-1024	2	1024
Birch	Structure	1000,000	2	100	1000
Unbalance	Balance	6500	2	2	100-1024

Table 2
Time Complexity of techniques

Technique	Complexity
Random Partitions	O(N)
Random Centroids	O(N)
Maxmin	O(kN)
Kmeans++	O(kN)
Sorting heuristic	O(NlogN)
Repeated k-means	$O(N^2)$
Density based	$O(N^{1.5})$

CHAPTER 3

PERFORMANCE METRIC

The result in the literature is based on following performance metrics:

3.1 Centroid Index:

The centroid index (CI) is the fundamental indicator of success. It counts number of actual clusters lacking centroid, or the clusters having excessive number of centroids.

If CI is zero, it is deduced that the clustering is accurate. The algorithm then solves the problem.

Then success rate is defined as the proportion of times (in percent) that an algorithm achieves the desired grouping (CI=0).

It counts the number of times the best potential clustering result is discovered.

3.2 SSE:

It is the sum of the squared Euclidean distances of each point to its nearest centroid. The objective function's value is measure of the algorithm's success. SSE is used in existing literature assessments of k-means. It's calculated as follows:

$$\text{SSE} = \sum_{i=1}^{N} ||xi - cj|| \qquad (4)$$

Here, x_i is a point and c_j is its closest centroid.

3.3 Purity:

The purity of each cluster is given by.

$$P = \sum(i = 1 \rightarrow k)\ (n_k/n)\ P_{ck} \qquad (5)$$

Here k is number of cluster.

The purity index is confined between 0 and 1; higher purity values indicate correct performance efficiency, and a purity of 1 indicates 100 percent efficiency.

3.4 Silhouette index:

It is a number that indicates how similar a point is to other points in the similar cluster, as well as how unlike it is to points in unsimilar clusters.

It can be calculated as:

$$S = 1/n \sum_{i=1}^{n} Sxi \qquad (6)$$

The Silhouette index is a popular method of estimating the number of clusters in a data set, but it may also be used to evaluate the quality of clustering.

The value is restricted between -1 and 1, with 1 indicating greatest cluster separation and maximum density inside clusters.

CHAPTER 4

HEURISTIC ANALYSIS

Based on the study the overall performance of various initialization strategies is discussed in this section.
The effect of the following factors on result is also discussed.

- Overlap of cluster
- Number of clusters
- Dimensions
- Cluster size imbalance

CI values: The random partition performs much worse than the random centroids. The best approach is maxmin heuristics (Maxmin and kmeans++).

SI and Purity: Maximin is the best in terms of total performance.

Cluster overlap: If overlapping is less, k means accuracy is decreased. Maxmin(more overlap) approach overcomes this weakness.

Number of clusters: If the number of clusters is more , CI value will be high and hence the success rate will reduce.

Dimensionality: From the study it is found that the performance of k- means is independent of dimensionality.

Cluster size imbalance: K-means++ has the highest rate of success considering this factor.

CHAPTER 5

CONCLUSION AND FUTURE SCOPE

One of the most widely used clustering methods is KM and it is mostly favoured when tackling the problem about how to cluster data.

One of the most critical tasks in the K-means algorithm is to start the centroids in a cluster. Choosing an initial centroid at random does not always result in the best solution. The algorithm is run numerous times with different random initializations to overcome this difficulty.

The KM++ algorithm is more calculatively intensive than the standard KM algorithm., the run-time in KM++ is remarkably lesser. The reason for this is that the centroids chosen at the start are likely to already be in different clusters.

Overall, we discovered that kmeans++ was the most effective way for initialising k-Means. Even though kmeans++ provided excellent initial results, it is still advised that k-Means be performed from several beginning points.

Given the fact that KM is a reasonably basic algorithm with a number of possibilities, it is not without flaws, its speed, scalability, and ease of interpretation make it particularly useful in a variety of scenarios.

The future work will include the performance of KM on more factors such as varying density, varying shape etc. Also, some standalone algorithms will be studied for improving the accuracy of k-means.

REFERENCES

[1] Khan, S. S., Ahmad, A., Cluster center initialization algorithm for k-means clustering, Pattern Recognition Letters 25 (11), 2004, pp. 1293-1302.

[2] S. Deelers, and S. Auwatanamongkol, Enhancing K-Means Algorithm with Initial Cluster Centers Derived from Data Partitioning along the Data Axis with the Highest Variance, Proceedings of world academy of science ,engineering and technology, Vol 26,December 2007 pp. 323-328.

[3] M.Sakthi and Dr. Antony Selvadoss Thanamani, An Effective Determination of Initial Centroids in K-Means ClusteringUsing Kernel PCA., International Journal of computer and information technologies, Vol. 2 (3). 2011, pp. 955-959.

[4] Neha Aggarwal and Kirti Aggarwal (2012a), A mid-point based k-mean clustering algorithm for data mining, InternationalJournal on Computer Science and Engineering, Vol. 4, No. 06.

[5] Raed T. Aldahdooh and Wesam Ashour (2013), DIMK-means ―Distance-based Initialization Method for K-meansClustering Algorithm, I.J. Intelligent Systems and Applications, 02, pp. 41-51.

[6] S.Z. Selim and M.A. Ismail, "K-means type a lgorithms: a generalized convergence theorem and characterization of local optimality," *IEEE Trans. Pattern Anal*, vol. 6,pp. 81–87, Mach 1984.

[7] R. Batra, S. Sharma, A. Shrivastav, P. Goyal, “Efficiently de-noising sms text for faq retrieval”, IEEE International Conference on Data Mining and Intelligent Computing, Delhi, Sep. 2014.

[8] D. Arthur and S. Vassilvitskii, "k-means++: The advantages of careful seeding," *ACM-SIAM Symposium on Discrete Algorithms (SODA 2007) Astor Crowne Plaza, New Orleans, Louisiana*, pp. 1–11, 2007.

[9] J. Waghmare, and M. A. Potey., “Survey of sms based faq retrieval systems”, International Journal Of Engineering And Computer Science 4, 2 (2015), 10259–10263.

[10] Chun Sheng Li, "Cluster Center Initialization Method for K -means Algorithm Over Data Sets with Two Clusters," *2011 International Conference on Advances in Engineering*, vol. 24, pp. 324 – 328, 2011.

[11] Kohei Arai and Ali Ridho Barakbah, "Hierarchical K-means: an algorithm for centroids initialization for K-means ," *Rep. Fac. Sci. Engrg, Saga Univ.* , vol. 36, 2007.

[12] M.C. Naldi, R.J.G.B. Campello, E.R. Hruschka, and A.C.P.L.F. Carvalho, "Efficiency issues of evolutionaryk-means," *Applied Soft Computing*, vol. 11 , pp. 1938–1952, (2011).

[13] Madhu Yedla, S.R. Pathakota, T.M. Srinivasa, Enhancing K-means Clustering Algorithm with Improved Initial Centre, International Journal of Computer Science and Information Technologies, 1 (2) , 2010, pp. 121-125.

[14] Koheri Arai and Ali Ridho Barakbah, Hierarchical k-means: an algorithm for centroids initialization for k-means, Reports of The Faculty of Science and Engineering Saga University, vol. 36, No.1, 2007.

[15] Pasi Fränti, Sami Sieranoja How much can k-means be improved by using better initializationand repeats, Pattern Recognition 93 (2019) 95–112.

[16] Joonas Hämäläinen, Tommi Kärkkäinen and Tuomo Rossi Improving Scalable K-Means++, MDPI(2020).

[17] Avgoustinos Vouros· Stephen Langdell · Mike Croucher2· Eleni Vasilaki An empirical comparison between stochastic and deterministic centroid initialisation for K-means variations, Machine Learning (2021) 110:1975–2003.

[18] E. Forgy, Cluster analysis of multivariate data: efficiency vs. interpretability of classification, Biometrics 21 (1965) 768–780.

[19] P. Fränti, J. Kivijärvi, Randomized local search algorithm for the clustering problem, Pattern Anal. Appl. 3 (4) (2000) 358–369.

[20] J.M Peña, J.A. Lozano, P. Larrañaga, An empirical comparison of four initializa- tion methods for the k-means algorithm, Pattern Recognit. Lett. 20 (10, Octo- ber) (1999) 1027–1040.

[21] J. He, M. Lan, C-L Tan, S-Y Sung, H-B Low, Initialization of Cluster Refinement Algorithms: a review and comparative study, IEEE Int. Joint Conf. Neural Netw. (2004).

[22] D. Steinley, M.J. Brusco, Initializing k-means batch clustering: a critical evalua- tion of several techniques, J. Classification 24 (2007) 99–121.

[23] M.E. Celebi, H.A. Kingravi, P.A. Vela, A comparative study of efficient initial- ization methods for the k-means clustering algorithm, Expert Syst. Appl. 40 (2013) 200–210.

[24] D. Steinley, Local optima in k-means clustering: what you don't know may hurt you, Psychol. Methods 8 (2003) 294–304.

[25] P. Bradley, U. Fayyad, Refining initial points for k-means clustering, in: Inter- national Conference on Machine Learning, San Francisco, 1998, pp. 91–99.

[26] P. Fränti, S. Sieranoja, K-means properties on six clustering benchmark datasets, Appl. Intel. 48 (12) (2018) 4743–4759.

[27] L. Morissette, S. Chartier, The k-means clustering technique: general consider- ations and implementation in Mathematica, Tutor. Quant. Methods Psychol. 9 (1) (2013) 15–24.

[28] J. Liang, L. Bai, C. Dang F. Cao, The k-means-type algorithms versus imbal- anced data distributions, IEEE Trans. Fuzzy Syst. 20 (4, August) (2012) 728– 745.

[29] I. Melnykov, V. Melnykov, On k-means algorithm with the use of Mahalanobis distances, Stat. Probab. Lett. 84 (January) (2014) 88–95.

[30] V. Melnykov, S. Michael, I. Melnykov, Recent developments in model-based clustering with applications, in: M. Celebi (Ed.), Partitional Clustering Algo- rithms, Springer, Cham, 2015.

[31] M. Rezaei, P. Fränti, Set-matching methods for external cluster validity, IEEE Trans. Knowl. Data Eng. 28 (8, August) (2016) 2173–2186.

[32] P. Fränti, M. Rezaei, Q. Zhao, Centroid index: cluster level similarity measure, Pattern Recognit. 47 (9) (2014) 3034–3045.

[33] S.J. Redmond, C. Heneghan, A method for initialising the K-means clustering algorithm using kd-trees, Pattern Recognit. Lett. 28 (8) (2007) 965–973.

[34] M.J. Norušis, IBM SPSS Statistics 19 Guide to Data Analysis, Prentice Hall, Upper Saddle River, New Jersey, 2011.

[35] T. Gonzalez, Clustering to minimize the maximum intercluster distance, Theor. Comput. Sci. 38 (2–3) (1985) 293–306.

[36] M.M.-T. Chiang, B. Mirkin, Intelligent choice of the number of clusters in k-means clustering: an experimental study with different cluster spreads, J. Classification 27 (2010) 3–40.

[37] J. Hämäläinen, T. Kärkkäinen, Initialization of big data clustering using distri- butionally balanced folding, Proceedings of the European Symposium on Arti- ficial Neural Networks, Comput. Intel. Mach. Learn.-ESANN (2016).

[38] I. Katsavounidis, C.C.J. Kuo, Z. Zhang, A new initialization technique for gener- alized Lloyd iteration, IEEE Signal Process Lett. 1 (10) (1994) 144–146.

[39] F. Cao, J. Liang, L. Bai, A new initialization method for categorical data cluster- ing, Expert Syst. Appl. 36 (7) (2009) 10223–10228.

[40] D. Arthur, S. Vassilvitskii, K-means++: the advantages of careful seeding,ACM-SIAM Symp. on Discrete Algorithms (SODA'07), January 2007.

[41] M. Erisoglu, N. Calis, S. Sakallioglu, A new algorithm for initial cluster centers in k-means algorithm, Pattern Recognit. Lett. 32 (14) (2011) 1701–1705.

[42] C. Gingles, M. Celebi, Histogram-based method for effective initialization of the k-means clustering algorithm, Florida Artificial Intelligence Research So- ciety Conference, May 2014.

[43] J.A. Hartigan, M.A. Wong, Algorithm AS 136: a k-means clustering algorithm, J. R. Stat. Soc. C 28 (1) (1979) 100–108.

[44] M.M. Astrahan, Speech Analysis by Clustering, Or the Hyperphome Method, Stanford Artificial Intelligence Project Memorandum AIM-124, Stanford Univer- sity, Stanford, CA, 1970.