Improving K-Mean Method by Finding Initial Centroid Points

Andleeb Aslam*, Usman Qamar*, Reda Ayesha Khan*, Pakizah Saqib*

* *Department of Computer and Software Engineering, National University of Sciences and Technology (NUST)*

Islamabad, Pakistan

andleebaslam0@gmail.com, usmanq@ceme.nust.edu.pk, reda.ayesha@gmail.com, pakizahfatima@gmail.com

***Abstract*— The paper is concerned with Improving k-Mean Algorithm in terms of accuracy by selecting the best initial seed points based on the provided k value. This paper presents two modified k-mean method for the selection of initial centroid points. In the first method based on the calculated k value with the help of elbow method, the original sorted data based on distances calculated using Euclidean distance method is divided into k equal partitions. And the mean of each partition is considered as initial centroid points. And in the second method the number of k is chosen randomly and the mean of each partition is considered as initial centroid points. We compared within cluster distance and number of iterations. Modified k-mean methods are better than original k-mean method as the distance within the clusters are less in modified k-mean than the original k-mean and the accuracy is also better.**

Keywords-component; k-mean, Centroid, Euclidean Distance, Clustering.

I. INTRODUCTION

Clustering is dividing a given dataset into partitions based on the k value i-e the required numbers of clusters to be formed[1]. Clustering helps in making clusters of similar elements having same attributes. K mean Algorithm is used for the clustering of data[2]. The number of clusters are equal to the value of k provided. Each cluster has element which are similar to one another than the elements present in other clusters. Clusters are made base on the near point distance to centroid. But In K-mean the Initial centroid points are selected randomly, which is also its limitation as selection of different initial centroids generates different results and clusters which make it less reliable[1]. Secondly in k-mean the number of k is provided by user.

There are many different methods for the selection of initial centroid point including K.A Abdul Nazir [6], KKZ method [7].

II. RELATED WORK

Wei Du, Hu Lin, Jianwei Sun, Bo yu, Haibo Yang proposed new solution for selection of initial centroid using distance computation and statistical information, for each dimension high density points are selected. Then for finding all possible center's density and distance are used. After this process work from high variance dimension to low variance ones, using k-nearest neighbours the final initial clusters centers are constructed [3].

Li Kangping, Wang Fei, Zhen Zhao, Mi Zengqiang, Sun Hongbin, Liu Chun proposed method of optimal selection of centroid point and thus they improved the k-mean algorithm using simulated annealing algorithm [4].Jie Yang, Yan Ma, Xiangfen Zhang, Shunbao Li, Yuping Zhang presented an optimal solution to select the initial centre points of clusters. They define a new distance measure having both density and Euclidean distance. Based on that, they proposed an efficient algorithm for selection of initial centre points of clusters that can dynamically adjust the weighting parameter [5].N. Nidheesh, K.A Abdul Nazir, P.M Ameer proposed an efficient and improved version of k mean based on density. Basically, the idea is to select the elements or data points which are adequately separated in feature space as initial centre point(centroid)and which belong to dense region [6].

KKZ method finds a point x preferably at the edge of the dataset to choose as first point.in the 2nd step the algorithm finds the point furthest from initial point x. Than the distance is calculated of all points to the first and second point. The next element (seed) is the point furthest from its nearest seed. This process of choosing seeds continue until k seeds are chosen [7].

Takashi Onoda, Miho Sakai, Seiji Yamada proposed a method for better initial centroid selection to form better clustering. For this a seeding method was proposed based on independent component analysis for clustering using k-mean [7].

K. Arai and Barakbah came up with the new method of calculating initial centroid. The algorithm basically works by taking the region of minimum distance which is the average of distances between data and then calculating the average distance of the near data points within it [8].

Rose Mawati, I Made Sumertajaya, Farit Mochamad Afendi gave a thought that cluster centroid is important in determining the effective cluster assignments. The proposed algorithm works by calculating the distances between the data points and the data points that have least distance difference are deleted from dataset and the process is repeated again until the number of datapoints that are being deleted and stored in another data set Am, reach to 0.75*(n/k). Then the mean of Am datasets are the initial centroid points [9].

III. METHODOLOGY

A. Original K-Mean Algorithm

In k-mean algorithm there exists basically two steps. In the first step the user randomly selects centroids from the given data and secondly based on the Euclidean distances the objects are assigned to clusters that are close. The algorithm's efficiency depends on the initial selection of centroids and on the value of k[10].

Input

K: number of user specified cluster

D= {d1, d2,…..dn}a dataset comprising of n objects.

Output

A set of resulting K clusters.

Method:

1. Select K data item arbitrarily from the given data set D.
2. Repeat

Allot each data object di from D to the nearest centroid on the basis of specified similarity measure.

Then, update current centroids of resulting clusters by computing new mean of all objects within ach cluster.

3. Repeat until no further changes occur

B. Modified K-Mean Algorithm (1st)

1. Calculate Euclidean distance among data points and sort them according to calculated distances.
2. Divide the dataset into half i.e. equal partitions for initial centroids. (For more accurate results determine k number using Elbow Method) [11].
3. Continue to divide each partition further into two until k partitions are made.
4. Calculate mean of each partition which will serve as initial centroids.

IV. CALCULATIONS AND RESULTS OF ORIGINAL K-MEAN AND PROPOSED K-MEAN METHOD

A. Original K-mean

Dataset

2,3,4,7,9,10,11,12,16,18,19,23,24,25,30

K=3

TABLE IV.1 K-MEAN EXTRACTION STEP I

Clusters	Randomly selected centroids	Clusters
2,3,4,7,9	c1=3	5
10,11,12,16,18,19	c2=16	14.33
23,24,25,30	c3=25	25.5

TABLE IV.2 K-MEAN EXTRACTION STEP II

New Clusters	New centroids from step I	Clusters
2,3,4,7,9	c1=5	5
10,11,12,16,18,19	c2=14.33	14.33
23,24,25,30	c3=25.5	25.5

Sum of distances within clusters=42

B. Modified K-mean

TABLE IV.3 K-MEAN EXTRACTION STEP I

Clusters	Calculated centroids using proposed methodology	Clusters
2,3,4	c1=5	3
7,9,10,11,12	c2=7.25	9.8
16,18,19,23,24,25,30	c3=20.8	22

TABLE IV.4 K-MEAN EXTRACTION STEP II

Clusters	Calculated centroids using proposed methodology	Clusters
2,3,4	c1=3	3
7,9,10,11,12	c2=7.25	7.25
16,18,19,23,24,25,30	c3=25.5	25.5

Sum of distances within clusters=39

V. RESULT ANALYSIS

Modified k-mean and the original k-mean are both applied on a same set of data to analyze the results and the difference they make. The proposed methodology results are better than the original k-mean. The distance within the clusters are less in modified k-mean than the original k-mean where the centroids are chosen randomly, which results in improved performance of it. The efficiency of the proposed method is also better because it depends on the distances within clusters, sum of distances which is reduced using modified methodology.

A. Elbow Method For Calculation of Number of K to Improve Accuracy And Performance

According to it, for the given dataset k = 5

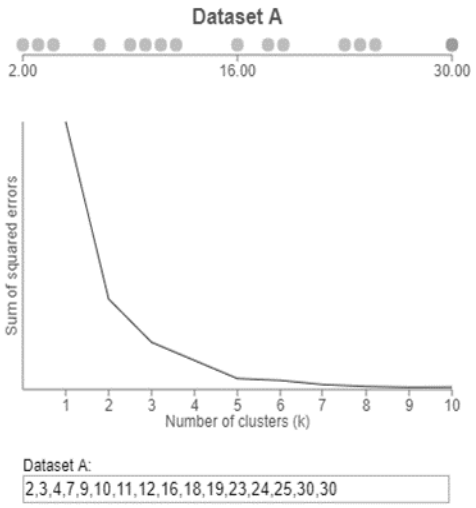

Figure 1. Elbow method graph showing no. of clusters

DATASET:
2,3,4,7,9,10,11,12,16,18,19,23,24,25,30
K=5
The dataset is already sorted according to distance. Dividing it into 5 equal partitions will result in:

TABLE V.1 CENTROID SELECTION USING ELBOW METHOD STEP II

Clusters	Calculated centroids using proposed methodology
2,3,4	c1=3
7,9,10	c2=16
11,12,16	c3=25
18,19,23	c3=20
24,25,30	c3=26

TABLE V.2 CENTROID SELECTION USING ELBOW METHOD STEP I

Clusters	Initial centroids	Distance within Clusters
2,3,4	c1=3	1,0,1=2
7,9,10	c2=16	2,0,1,2=5
11,12,16	c3=25	2,1,3=6
18,19,23	c4=20	2,1,3=6
24,25,30	c5=26	3,2,1,4=10

Sum of distances within clusters=39
Now the centroid points are not changing so the clusters formed are final clusters. From table V results it can be clearly seen that the best results for all possible scenarios or for any number of k the algorithm best works if initially the dataset is divided into two equal parts. Though the algorithm doesn't always need to start by dividing it into two parts as results remain the same, but we only need to keep in mind the required K value.

The proposed method is not so complex and its main advantage is simplicity as it the initial centroids are calculated easily. Furthermore, the modified algorithm shows the best value for k while in original k mean that value is selected by user while it requires no user involvement. It also increases the efficiency as the distances ae reduced within the clusters thus forming better clusters than original k-mean algorithm.

B. Modified K-Mean Algorithm (2nd)

1) Calculate Euclidean distance among data points and sort them according to calculated distances.

2) Based on provided K value by user, make k-equal partitions.

3) Take the mean of each partition and that mean will be the initial centroid point.

4) Make clusters according to distances calculated and the nearest distance to centroid.

5) Repeat III to IV unless and until centroids and clusters don't change.

TABLE V.3 COMPARISON BETWEEN ORIGINAL AND MODIFIED K-MEAN

Clusters	Initial centroids	Distance within Clusters
	k-mean	Modified k-mean
K=2	53	51
K=3	43	42
K=4	24	23
SUM	120	116

VI. CONCLUSION AND FUTURE WORK

Different researches have been conducted on the selection of initial centroid points for k-mean for making better clusters. This paper proposed modified k-mean methods for selection of initial cluster points. These methods surely increase the efficiency of k-mean by giving good initial clusters but on the other hand the computational cost is also
increased in calculation of initial centroid point. This algorithm also involves original k-mean calculation steps within it along

with modification method for selection of better centroid points. The results reflect that with the increase in k value, the within clusters distances also reduce and thus gave better clustering results. Furthermore, we are taking the mean for centroid calculation but not the midpoint, that also helps in better centroid selection. So, the modified k-mean is more accurate and efficient in making clusters than original k-mean that selects the initial centroid points randomly. The future work will focus on running these algorithms on UCI data sets.

REFERENCES

[1] Nie, Feiping, Cheng-Long Wang, and Xuelong Li. "K-Multiple-Means: A Multiple-Means Clustering Method with Specified K Clusters." Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. ACM, 2019.

[2] Rajeswari, K., et al. "Improvement in K-means clustering algorithm using data clustering." 2015 International Conference on Computing Communication Control and Automation. IEEE, 2015.

[3] Du, Wei, et al. "Combining Statistical Information and Distance Computation for K-Means Initialization." 2016 12th International Conference on Semantics, Knowledge and Grids (SKG). IEEE, 2016.Author, F., Author, S.: Title of a proceedings paper. In: Editor, F., Editor, S. (eds.) CONFERENCE 2016, LNCS, vol. 9999, pp. 1–13. Springer, Heidelberg (2016).

[4] Kangping, Li, et al. "Analysis on residential electricity consumption behavior using improved k-means based on simulated annealing algorithm." Power and Energy Conference at Illinois (PECI), 2016 IEEE. IEEE, 2016.

[5] Yang, Jie, et al. "An initialization method based on hybrid distance for k-means algorithm." Neural computation 29.11 (2017): 3094-3117.

[6] Nidheesh, N., KA Abdul Nazeer, and P. M. Ameer. "An enhanced deterministic K-Means clustering algorithm for cancer subtype prediction from gene expression data." Computers in biology and medicine 91 (2017): 213-221.

[7] Onoda, Takashi, Miho Sakai, and Seiji Yamada. "Careful seeding method based on independent components analysis for k-means clustering." Journal of Emerging Technologies in Web Intelligence 4.1 (2012): 51-59.

[8] Barakbah, Ali Ridho, and Kohei Arai. "Centronit: Initial Centroid Designation Algorithm for K-Means Clustering." EMITTER International Journal of Engineering Technology 2.1 (2014): 50-62.

[9] Mawati, Rose, I. Made Sumertajaya, and Farit Mochamad Afendi. "Modified Centroid Selection Method of K-Means Clustering."

[10] Liu, Zhe, Jianmin Bao, and Fei Ding. "An Improved K-Means Clustering Algorithm Based on Semantic Model." Proceedings of the International Conference on Information Technology and Electrical Engineering 2018. ACM, 2018.

[11] Bedi, Jatin, and Durga Toshniwal. "Empirical Mode Decomposition Based Deep Learning for Electricity Demand Forecasting." IEEE Access 6 (2018): 49144-49156.

Andleeb Aslam received the B.S. degree in Software Engineering from University of Engineering and Technology, Taxila, Pakistan. She is currently pursuing the M.S. degree in computer software engineering with the Computer and Software Engineering Department, College of Electrical and Mechanical Engineering, National University of Sciences and Technology (NUST), Islamabad, Pakistan. Her area of research is Natural Language Processing (NLP).

Usman Qamar has over 15 years of experience in data engineering and decision sciences both in academia and industry having spent nearly 10 years in the UK. He has a Masters in Computer Systems Design from University of Manchester Institute of Science and Technology (UMIST), UK. His MPhil in Computer Systems was a joint degree between UMIST and University of Manchester which focused on feature selection in big data. In 2008/09 he was awarded PhD from University of Manchester, UK. His PhD specialization is in Data Engineering, Knowledge Discovery and Decision Science. His Post PhD work at University of Manchester, involved various research projects including hybrid mechanisms for statistical disclosure (feature selection merged with outlier analysis) for Office of National Statistics (ONS), London, UK, churn prediction for Vodafone UK and customer profile analysis for shopping with the University of Ghent, Belgium. He has also done a post-graduation in Medical & Health Research, from University of Oxford, UK, where he worked on evidence-based health care, thematic qualitative data analysis and healthcare innovation and technology.

He is director of Knowledge and Data Science Research Centre, a Centre of Excellence at NUST, Pakistan and principal investigator of Digital Pakistan Lab, which is part of National Centre for Big Data and Cloud Computing. He has authored over 150 peer reviewed publications which includes 2 books published by Springer & Co, 21 Book Chapters, 38 Impact Factor Journal Publications with a combined impact factor of 104 (Clarivate Analytics Impact Factor) and over 100 Conference Publications. Many of his papers have been awarded best research paper awards by Higher Education Commission, Pakistan. Because of his extensive publications he is member of Elsevier Advisory Panel. He has successfully supervised 4 PhD students and over 70 master students.

Dr. Usman has been able to acquire nearly PKR 100 million in research grants. He has received multiple research awards, including Best Book Award 2017/18 by Higher Education Commission (HEC), Pakistan, Best Researcher of Pakistan 2015/16 by Higher Education Commission (HEC), Pakistan, Best Overall NUST University Researcher Award 2016 and Best College of E&ME researcher award 2016 as well as gold in Research & Development category by Pakistan Software Houses Association (P@SHA) ICT Awards 2013 & 2017 and Silver award in APICTA (Asia Pacific ICT Alliance Awards) 2013 in category of R&D hosted by Hong Kong. He is also recipient of the prestigious Charles Wallace Fellowship 2016/17 as well as British Council Fellowship 2018, visiting research fellow at Centre of Decision Research, University of Leeds, UK and scientific director of Data and Text Mining Lab, Manchester Metropolitan. He is also an expert committee member of engineering & technology for the evaluation/recognition of national research journals for Higher Education Commission (HEC), Pakistan. Finally, he has the honour of being the finalist of the British Council's Professional Achievement Award 2016/17

Reda A. Khan received the B.S. degree in Software Engineering from University of Engineering and Technology, Taxila, Pakistan. She is currently pursuing the M.S. degree in computer software engineering with the Computer and Software Engineering Department, College of Electrical and Mechanical Engineering, National University of Sciences and Technology (NUST), Islamabad, Pakistan. Her area of research is Ontology Engineering.

Pakizah Saqib received the B.S. degree in Software Engineering from PUCIT, University of the Punjab, Lahore, Pakistan. She is currently pursuing the M.S. degree in computer software engineering with the Computer and Software Engineering Department, College of Electrical and Mechanical Engineering, National University of Sciences and Technology (NUST), Islamabad, Pakistan. Her area of research is Gene expression dataset analysis using Machine Learning techniques.