# A Proposed Modified K-means Algorithm for Effective Clustering based on Element Properties

**Shraddha N. Umarkar** **Ankita S. Kadu** **Prof. N. J. Padole**

***Abstract*—Clustering is technique of grouping objects in single group having nearly equal properties. There are number of different clustering algorithms but k-means algorithm is a most popular clustering algorithm. K-means algorithm is most effective and simple clustering algorithm; it is one of the most widely used algorithms for clustering. But it is very difficult to predict number of cluster and also lead to underflow condition if specify number of cluster initially, so authors tried to introduce modified k-means algorithm which include properties of data elements and also overcome problem of predicting number of cluster in k-means algorithm. In this paper analyse basic k-means algorithm and tried to introduce modified k-means algorithm.**

***Keywords*: K-means, clustering.**

## I. INTRODUCTION

Data clustering is process of arranging an object in a single cluster having some similar attribute, objects having different attribute belongs to different cluster. Cluster analysis is widely used methods to analysing data for many practical applications such as engineering, bioinformatics. The k-means algorithm [1,2,4,3,4,5,6] is effective in producing clusters for many practical applications. But the computational complexity of the original k-means algorithm is very high, especially for large data sets. Moreover, this algorithm results in different types of clusters depending on the random choice of initial number of cluster. K-means algorithm has applications in many areas like data mining, clustering. It is very difficult to predict initially number of cluster for data set; it may lead to empty clusters and if data in more quantity then it may also lead to underflow of cluster number which has initially passed and data element having different properties place in improper cluster.

## II. RELATED WORK

The researchers introduced an enhanced K-means algorithm to improve the time complexity using uniform data. They make clusters in two phases. In phase one, they find initial clusters on similarity basis, while in second phase they finalize clusters [7]. Similarly all of the requirements, advantages and disadvantages of basic K-mean clustering are discussed [8]. Mary, C.I. & et al, proposed technique of Ant Colony Optimization (ACO) is proposed to improve K-means clustering. The researchers have contributed to improve the cluster quality after grouping. Their proposed method has two phases. In the first phase, on the basis of statistical modes, initial centroids for K-mean clustering are selected. In the second phase, they improve the cluster quality by using ant refinement algorithm [9]. Visalaksh and et al, an alternate distance measure namely Max-min measure is proposed. By using the Max-min normalization, entire data is adjusted in limits [0, 1] and after this normalization clustering is done [10]. Similarly Singh R.V. and et al, A data clustering technique by using K-means algorithm is presented, which is based on the initial mean of the cluster, According to this algorithm, whole data space is divided into segments (k*k) and the frequency of data points in each segment is calculated. The segment having the highest frequency will have maximum probability of having centroid. If more than one consecutive segment having the same frequency then that segments are merged. After this, distances of data points and centroids are calculated. In same manner the process is continued [11].

## III. BASIC K-MEAN ALGORITHM

In basic k-means algorithms clusters are total depend on initial cluster value which we have passed. Distance formula is used to calculate the distance between data elements and moved to appropriate cluster. The process is continued until no more changes occur in clusters.

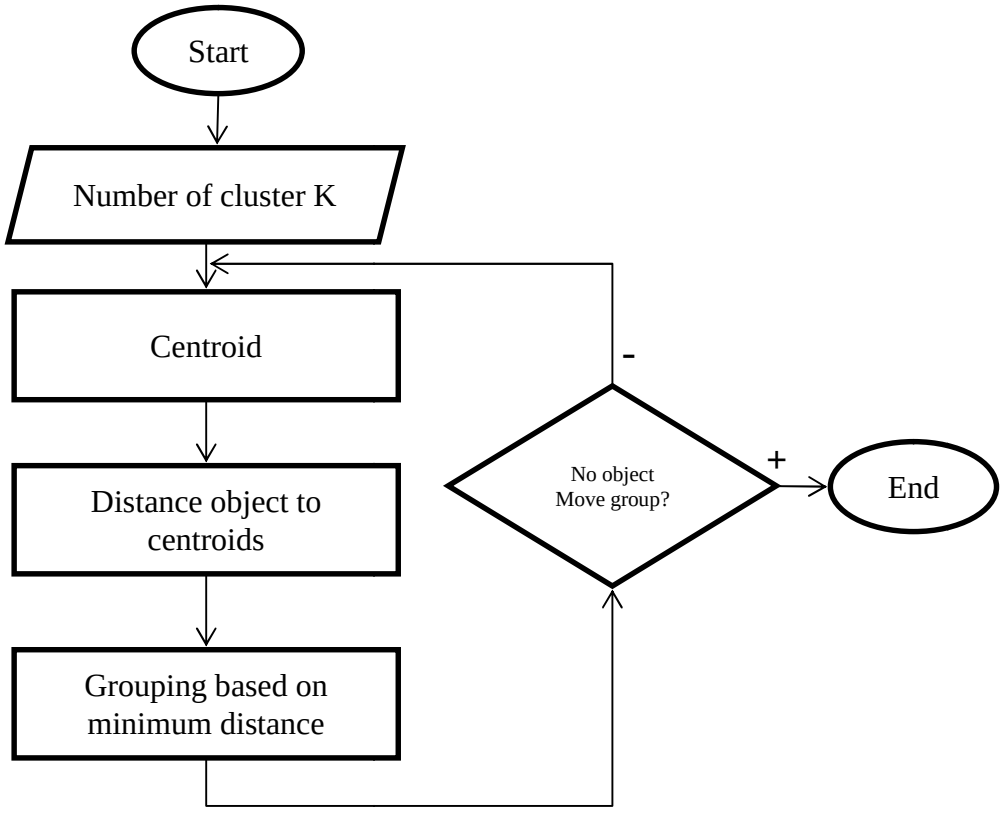


Fig 1: flowchart of basic k-means algorithm

## IV. PROBLEM DEFINITION

By analysing whole work on k-means algorithm there is calculate distance of each data element from centroid and moved in appropriate cluster, according to this data element having different properties are leads to improper clustering. In this paper authors tries to introduce modified k-means algorithm which based on different properties of data elements and dynamically creates a clusters.

## V. MODIFIED K-MEANS ALGORITHM

*Steps:-*

Step 1: input D of element data set and integer k
Step 2: randomly select data element form first cluster (first time only)
Step 3: select next data element
Step 4: calculate distance from centroid by using distance formula
Step 5: match some properties of data element to previous clustered data element
If some properties are match then put this data element in cluster else form new cluster
Step 6: repeat steps 2 to step 5 until no change occur
Step 7: output k – clusters and exit

From the above algorithm of K-means clustering, it is clear that the proposed system architecture adopts flexible, effective in clustering data set according to their properties. The basic steps of K-mean clustering is simple, firstly we determined the input D of element data set and integer k. when we randomly select the data set from first cluster, only first cluster is created. We cannot predefined the cluster it created by their own properties. Now calculate the distance from centroid formula that is Euclidean distance formula. The cluster are totally depend on initial cluster value which we passed. The distance formula is used to calculate the distance between data elements then matching the properties of the data set and move the data to appropriate cluster. Whenever matching step is completed then put this data element in previous cluster otherwise we create the new cluster and the output k cluster is to be created.

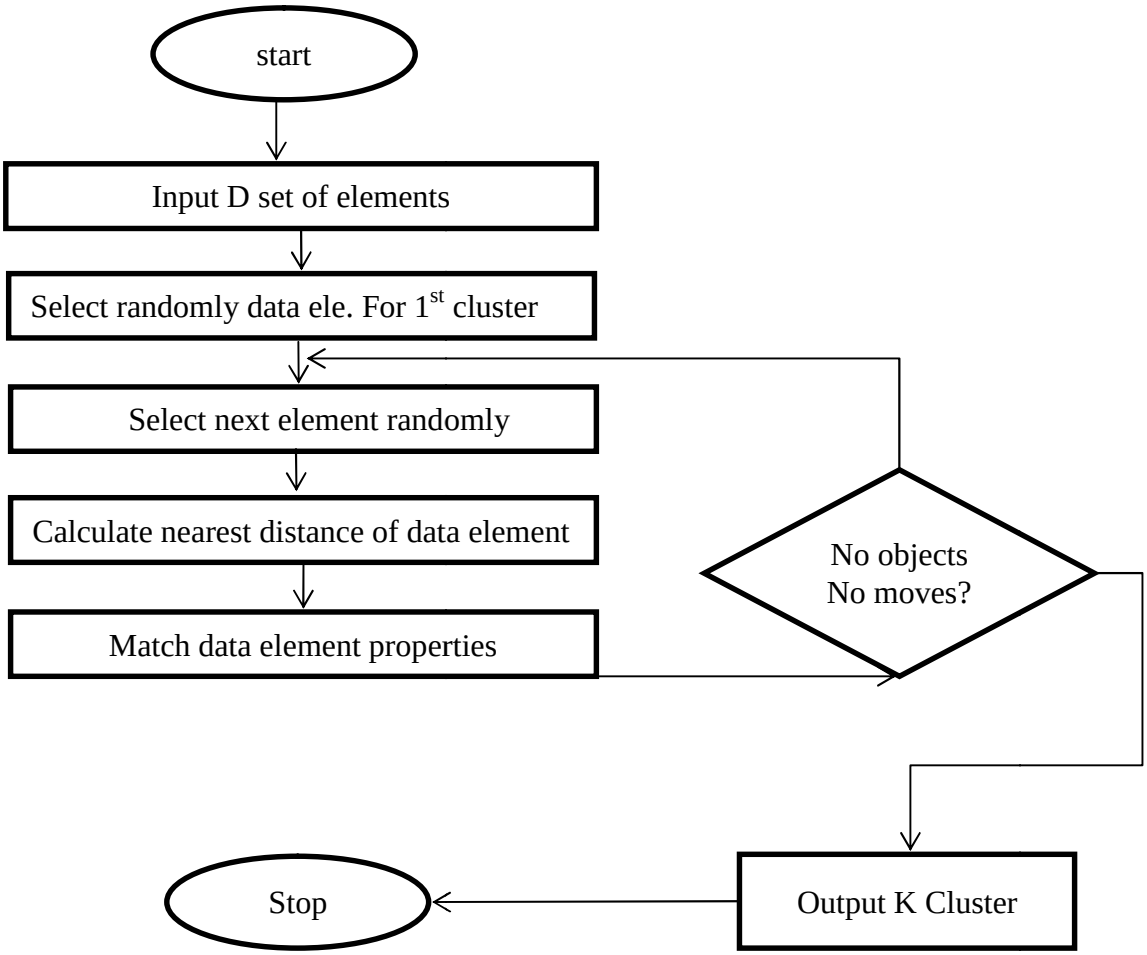


*Fig 2: flowchart of modified k-means algorithm*

## ADVANTAGES

1. If variables are huge then k-mean most of times computationally faster than hierarchical clustering.
2. K-means produce clusters based on distance and data set properties.
3. It prevents empty clustering.

## DISADVANTAGES

1. Different initial partitions can result in different final cluster.
2. If doesn't work well with cluster of different size and different density.

## CONCLUSION

In this paper, K-means algorithm for effective cluster technique is used. The k-means cluster technique and they all have slightly different goals from other technique. The domain topics K-means algorithm and clustering are highlight in the paper. As author's proposed K-mean cluster algorithm based on distance and properties of data set. K-means algorithm is widely used for clustering. In these proposed modified K-mean algorithm introduced a new concept of pattern matching which is very beneficial to cluster the data element according to their data property.

## REFERENCES

[1] Jiawei Han M. K, Data Mining Concepts and Techniques, Morgan Kaufmann Publishers, An Imprint of Elsevier, 2006.
[2] Margaret H. Dunham, Data Mining- Introductory and Advanced Concepts, Pearson Education, 2006.
Proceedings of the World Congress on Engineering 2009 Vol I WCE 2009, July 1 - 3, 2009, London, U.K.
[3] McQueen J, "Some methods for classification and analysis of multivariate observations," Proc. 5th Berkeley Symp. Math.Statist. Prob., (1):281–297, 1967.
[4] Merz C and Murphy P, UCI Repository of Machine Learning Databases, Available: ftp://ftp.ics.uci.edu/pub/machine-learning-databases
[5] Pang-Ning Tan, MichaelSteinback and Vipin Kumar, Introduction to Data Mining, Pearson Education, 2007.
[6] Stuart P. Lloyd, "Least squares quantization in pcm," IEEE Transactions on Information Theory, 28(2): 129-136.
[7] Napoleon,D. and P.G. Lakshmi, 2010. "An Efficient K-means Clustering Algorithm for Reducing Time Complexity using Uniform Distribution Data Points," in Trendz in Information Sciences and Computing (TISC), Chennai
[8] Master, C.P. and X.G. Professor, 2011. "A Brief Study on Clustering Methods Based on the K-means algorithm," in 2011 International Conference on E-Business and E-Government (ICEE), Shanghai, China.
[9]Mary, C.I. and S.V.K. Raja, 2009. "Refinement of clusters from k-means with ant colony optimization," Journal of Theoretical and Applied Information Technology, 9(2): 28-32
[10] Visalakshi, N.K. and J. Suguna, 2009. "K-means Clustering using Max-min Distance Measure," in Annual Meeting of the North American Fuzzy Information Processing Society, NAFIPS, Cincinnati,OH
[11] Singh,R.V. and M.P. Bhatia, 2011. "Data Clustering with Modified K-means Algorithm," in International Conference on Recent Trends in Information Technology (ICRTIT), Chennai, Tamil Nadu.

## AUTHOR'S PROFILE

**Shraddha N. Umarkar** received her BSc in Computer Science from SGB Amravati University. Presently student of MSc in Computer Science from HVPM an autonomous college.

**Ankita S. Kadu** received her BSc in Computer science from SGB Amravati University. Presently student of MSc in Computer Science from HVPM an autonomous college.

**Prof. N. J. Padole**
Assistant prof in PG Department of computer science and technology in HVPM auto autonomous college.