# Improved K-Means Algorithm for Categorical Dataset

Mr. Yuvraj Sase
Department of Computer Engineering,
Vishwabharti Academy's College Of Engineering,
Ahmednagar, India

Prof. Jaypal P. C.
Department of Computer Engineering,
Vishwabharti Academy's College Of Engineering,
Ahmednagar, India

***Abstract*— K-Means algorithm is most popular clustering algorithm. K-Means algorithm classifies the objects based attribute / features into K number of groups where user input K is number of cluster. But there is no any thumb rule to calculate K value. User has to calculate K value by trial and error method. It is very hard and inefficient for user to check and analyze every possible input and best among them. There is need to find an algorithm independent of K. So that improved K-Means algorithm was proposed which does not require K as input. But improved K-Means algorithm doesn't work for categorical datasets. The real world generates lot off categorical data. There is need K-Means which work for categorical dataset. This paper proposed "an improved K-Means algorithm" which works for categorical datasets.**

***Keywords*— K-Means algorithm, Clustering, Categorical , Numerical, Distance, Datapoints, Grouping, dissimilarity, Centroid, Mean***

## I. INTRODUCTION

In today's world data is in every sector like banking, hospital, trade marketing, social media etc. They produce both types of data like categorical data, numerical data. There are actually two types of datasets first numerical and second categorical. Numerical contains numeric data like 5, 10 ,15 etc. and categorical dataset contains data which can be divided into categories for example color attribute having categorical data as red, green, yellow etc.

There is always need categorize such data for example selecting area wise customer, grouping patient according their disease etc. For this purpose many clustering algorithms are proposed. Clustering is process which group all objects depending on their dissimilarity i.e distance between objects. Among all of them K-Means algorithm is most popular clustering algorithm. K-means clustering algorithm divide data into K number of cluster where K is user input, but user does not have any thumb rule to calculate this K-value. User has calculate this value by trial and error method where user has to check every possible value for K. User analyze result generated from every value and find best K value. This is very inefficient and time consuming process. So there was need to propose algorithm which is independent of K. For this purpose an improved K-means algorithm was proposed that independent of K. This algorithm finds clusters without K value. But improved K-Means algorithm does not work for categorical type of data. This paper mainly focuses on improved K-means for categorical type of data. The main idea is convert this categorical data set into numerical dataset, so that categorical data get some value which can be used by improved K-Means algorithm and form clusters. So any type of data can be grouped by proposed algorithm.

## II. BACKROUND AND RELATED WORK

Dynamic clustering is achieved by improving K-Means algorithm. The main purpose of dynamic clustering is to improve quality of cluster. This algorithm also fixes the optimal number of cluster. For this algorithm both fixed and dynamic work well. This algorithm use intra distance and inter distance to calculate cluster where intra distance is distance between cluster centroid and cluster data points , inter distance is distance between cluster centroids of each cluster.

Variance and median also can be used to initialize cluster center where variance means how far a set of numbers is spread out i.e. distance between each point and mean value , median means middle value. Diagonal also can be used to calculate initial cluster center. Firstly data points are divided into K number of rows and K number of columns. Then width and height is calculated as in,

$$X_w = \frac{X_{max} - X_{min}}{K}$$

Where $X_{max}$ = biggest x value.

$X_{min}$= Lowest x value.

$$Y_w = \frac{Y_{max} - Y_{min}}{K}$$

Where $Y_{max}$ = biggest Y value.

$Y_{min}$= Lowest Y value.

Now Upper left point of cell is selected as base points. Area for initial centroid is chosen by moving this base points left up to $\frac{x_w}{2}$ in X-axis and $\frac{y_w}{2}$ in Y-axis and right up to $X_w$ in X-axis and $Y_w$ in Y-axis. Now centroids are selected randomly in area form by diagonal points. When centroids are found then K-Means algorithm is applied on it. Y-Means algorithm is also used to find out initial centroids. Y-means uses sequence of splitting, deleting, and merging the clusters.

## III. Existing System

K-Means algorithm takes K (no. of clusters) input from user and there is no any thumb rule to calculate K value. User has to calculate by trial and error method and this is very inefficient. So an algorithm was proposed called as Improved K-Means algorithm [1]. Improved K-Means algorithm works independent on K value for numerical type of data. Number of clusters (K) does not require as input in modified K-Means algorithm. This algorithm use outliers to calculate value of number of clusters i.e. K. Categorical data is one of problem for this algorithm.

```
Input:
    D: The set of n tuples with attributes Al, A2 . . . Am
       where m = no. of attributes. All attributes are numeric
Output:
    Suitable number of clusters with n tuples distributed properly.
Method:
  1) Compute sum of the attribute values of each tuple (to find the points in the data set which are farthest apart).
  2) Take tuples with minimum and maximum values of the sum as initial centroids.
  3) Create initial partitions (clusters) using Euclidean distance between every tuple and the initial centroids.
  4) Find distance of every tuple from the centroid in both the initial partitions. Take d=minimum of all distances (other than zero).
  5) Compute new means (centroids) for the partitions created in step 3.
  6) Compute Euclidean distance of every tuple from the new means (cluster centers) and find the outliers depending on the following objective function: If Distance of the tuple from the cluster mean < d then not an Outlier.
  7) Compute new centroids of the clusters.
  8) Calculate Euclidean distance of every outlier from the new cluster centroids and find the outliers not satisfying the objective function in step 6.
  9) Let B= {Yl, Y2 ...Y p) be the set of outliers obtained in Step 8 (value of k depends on number of outliers).
  10) Repeat until (B== φ)
      a) Create a new cluster for the set B, by taking mean value of its members as centroid.
      b) Find the outliers of this cluster, depending on the objective function in step 6.
      c) If no. of outliers = p then
            i.  Create a new cluster with one of the outliers as its member and test every other outlier for the objective function as in step 6.
            ii. Find the outliers if any
      d) Calculate the distance of every outlier from the centroid of the existing clusters and adjust the outliers the existing which satisfy the objective function in step 6.
      e) B = {ZI, Z2 .... Zq} be the new set of outliers (value of q depends on number of outliers).
```

## IV. Proposed Work

The proposed algorithm converts categorical into numerical data set. Some equations are formed using distance between data points. These equations are used to calculate value of categorical data. The calculated values are used as input dataset for improved K-Means to divide data into groups without knowing value of K. In this way grouping is done for categorical dataset. Steps of proposed algorithm for converting categorical data into numerical data are given below. template is designed so that author affiliations are not repeated each time for multiple authors of the same affiliation. Please keep your affiliations as succinct as possible (for example, do not differentiate among departments of the same organization).

1) Take input dataset D contains categorical as well as numerical dataset.

2) Calculate number of categories as cate [].

3) Until (cate[] == φ)

4) Select two categories as C1 and C2.

5) Search two data points as P1 and P2 having C1 category.

6) Search for one data point as P3 having C2 category.

7) Calculate distance between data point P1 and P2 as D12.

$$D_{12}^{\ 2} = N_{12}^{\ 2} + C_{12}^{\ 2} \ldots\ldots\ldots\ldots\ldots\ldots \text{(1).}$$

Where,

$N_{12}^{\ 2}$ = distance due to numerical data which can calculate from numerical data.

$C_{12}^{\ 2}$ = distance due to categorical data.

8) Since both P1 and P2 have same value for categorical data. Therefore,

$$C_{12}^{\ 2} = 0 \ldots\ldots\ldots\ldots\ldots \text{(2).}$$

9) From equation (2),

$$D_{12}^{\ 2} = N_{12}^{\ 2} \ldots\ldots\ldots\ldots \text{(3)}$$

10) Now calculate distance between data point P2 and P3 having categories C1 and C2 respectively as D23.

,

$$D_{23}^{\ 2} = N_{23}^{\ 2} + C_{23}^{\ 2} \ldots\ldots\ldots\ldots\ldots\ldots \text{(4).}$$

11) Add equation (3) and equation (4),

$$D_{12}^{\ 2} + D_{23}^{\ 2} = (N_{12}^{\ 2} + N_{23}^{\ 2}) + C_{23}^{\ 2} \ldots\ldots \text{(5).}$$

12) Subtract equation (3) and equation (4).

$$D_{12}^{\ 2} - D_{23}^{\ 2} = (N_{23}^{\ 2} - N_{12}^{\ 2}) + C_{23}^{\ 2} \ldots\ldots \text{(6).}$$

13) Add equation 5 and equation 6

$$2D_{12}^{\ 2} - N^2 = 2C_{23}^{\ 2} \ldots\ldots\ldots\ldots \text{(7).}$$

From above equation value of categorical data can be calculate. Repeat above steps till all categorical data values are not foundThese values are used as input for improved K-Means algorithm. Proposed improved algorithm is independent of K and also work for categorical data set.

## V. CONCLSION

In real world, almost all dataset contains categorical dataset. So clustering is hard in real time datasets. Improved K-Means algorithm was proposed to remove dependency on K, but it does not work for categorical data. In this paper algorithm to convert categorical data to numerical data is proposed which is used to remove problem for categorical data in improved K-Means algorithm. Proposed algorithm uses distance to calculate value of categories. After calculation of all categories this values can be used as input to improved K-Means algorithm. In this paper proposed algorithm remove dependency on K and work for categorical dataset. For future work efficiency of this algorithm can be improved further by checking different methods to calculate initial centroid

## ACKNOWLEDGMENT

I would like to thank Prof. Kshirsagar M. (M.E coordinator) and Prof. Jaypal P. C. (Project Guide) for their support and their efforts in making sure that sufficient facility were always available. I am indebted to the authors whose works have been referred in completing this paper. I would also like to thank Dr. Kureshi (Principal, Vishwabharti Academy's College Of Engineering) for his guidance and support. I would also like to thank all my M.E classmates for their help to understand topic.

## REFERENCES

[1] Anupma Chadha, Suresh Kumar, “An Improved K-Means Clustering Algorithm : A Step Forward For Removal Of Dependency on K”, 2014 International Conference on Reliability, Optimization and Information Technology - ICROIT 2014, India, Feb 6-8 2014.

[2] B M Ahamed Shafeeq, K S Hareesha, " Dynamic Clustering of Data with Modified K-Means Algorithm", International Conference on Information and Computer Networks (ICICN 2012), IPCSIT, vol. 27,pages 221-225,2012

[3] M. AI Dauod, "A New Algorithm for Cluster Initialization", World Academy of Science, Engineering and Technology, issue 4 ,2007

[4] Mohammed EI Agha, Wesam M. Ashour, " Efficient and Fast Initialization Algorithm for K-means Clustering" ,1.1. Intelligent Systems and Applications, vol. 4, issue 1, pages 21-31, 2012

[5] Kohei Arai, Ali Ridho Barakha, "Hierarchical K-means: An algorithm for centroids initialization for K-means", Reports of the Faculty of Science and Engineering, Saga University, vol. 36, issue. 1, pages 25-31, 2007

[6] VLeela, K.Sakthipriya, R.Manikandan, "A comparative analysis between k-mean and y-means Algorithms in Fisher's Iris data sets" , International Journal of Engineering and Technology, vol 5, issue 1, pages 245- 249,2013

[7] Stephen J. Redmon, Conor Heneghan, " A method for initializing the K-means clustering algorithm using kdtrees", Journal Pattern Recognition Letters, vol. 28, issue 8, pages 965-973,2007