An Improved and Customized I-K Means For Avoiding Similar Distance Problem

Pankaj Jadwal[1]
M.Tech Scholar(SGVU Jaipur)

Mrs.Ruchi Dave[2]
SGVU ,Jaipur

ABSTRACT

In k-means clustering algorithm, we have given n as number of points, k as number of clusters and t as number of iterations for having optimized clusters. But there are some problems associated with k means algorithm. A research paper [4] develops an incremental algorithm for solving sum-of-squares clustering problems in gene expression data sets. Clustering in gene expression data sets is a typical problem. A recent approach [5] to scaling the K means algorithm is based on discovering three kinds of regions .The problem which I am trying to solve is the similar distance problem. When we start clustering of elements, problem can be arise that point x1 which I want to clusterize can be at the same distance from more than one cluster. At that time we cannot decide that which cluster I should choose for that element. So I proposed a approach towards solving similar distance problem using improved k means clustering approach.

Keywords: k-means, clustering, optimized cluster.

I. INTRODUCTION

K Means is one of the simplest clustering algorithms that solve the famous clustering problem. The algorithm follows a simple and easy way to clusterize a given data set through a fixed number of clusters (assume k clusters). The main thought is to define k centroids, one for each cluster. These centroids should be placed in such a way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the closest centroid.There are so many cons with K Means and many researchers have published papers on that problems. A research paper named Selection of K in K-means clustering presented [1]. This paper first reviews existing methods for selecting the number of clusters for the algorithm. Parameters that affect this selection are then discussed and a new criterion to assist the selection is proposed. A research paper has been published named An Modified K Means algorithm for removing the problem of empty cluster[2]. In this paper, they have presented a improved *k*-means algorithm which removes the problem of generation of empty clusters (with some exceptions).But the problem on which I am working is the point which I want to clusterize, having same distance from 2 or more than centriods.so this is the point where the existing K Means fails. So I have proposed an improved and customized K Means so that this type of problem can be handled.

II. Problem Formulation

There are so many problems associated with K Means algorithm. A lot of research has been done on K Means clustering. So many research papers has been published in order to solve the problems associated with K Means Clustering. Problem on which I am working is similar distance problem in K Means Clustering algorithm. As we know that in K means algorithm, parameter on which we decide that a particular point will come in which cluster, is the distance of that point from the centroid of that cluster. We can say that a particular point will go into that cluster whose distance from centriod of any cluster is minimum. But if distance of any point p1 to 2 or more than 2 cluster's centroid is same than we are not in the position of finding that point p1 should go into which cluster. Than I am just improving K Means clustering algorithm so that Improved K Means will be able to face such type of problem.

Now I am doing slight modification in K Means Clustering algorithm so that we can solve such type of problem

I-KMeans algorithm:

NOP: Number of points in a cluster

D (ki): Density of cluster ki

1. Place K points into the space represented by the objects that are being clustered. These points represent initial group centroids.
2. Assign each object to the group that has the closest centroid.
3. If more than one point has same and smallest distance from the centroid then calculate NOP (Number of Points) for each cluster and find Min NOP. Object is assigned to the min NOP.

4. If NOP is same for these clusters than we calculate density for those clusters. D (ki) is the density of cluster ki where i is from 1 to number of clusters having same NOP problem.

D (ki) = (Min distance from centriod of the ki cluster+ Max distance from centriod of the ki Cluster)/2

Element will be inserted to that cluster whose D (ki) will be minimum.

5. When all objects have been assigned, recalculate the positions of the K centroids.

6. Repeat Steps 2, 3 and 4 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.

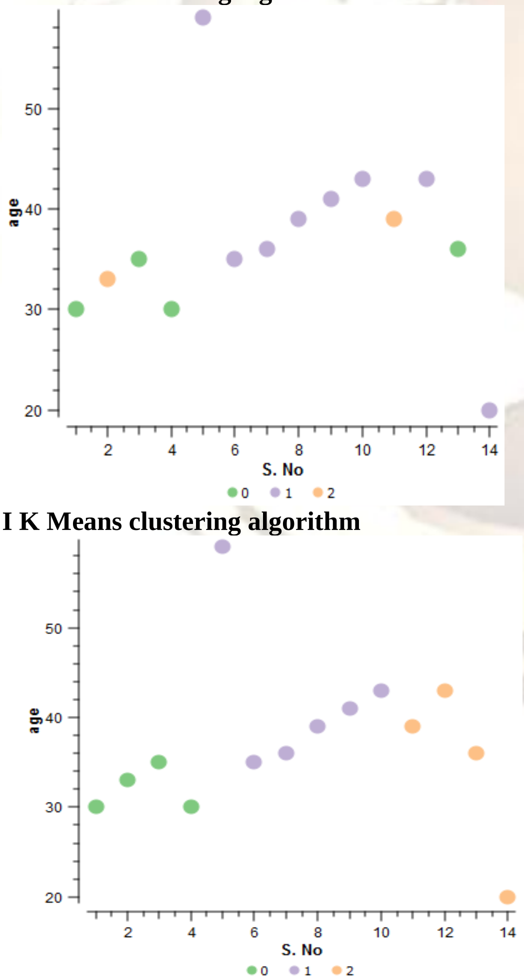
III. EXPERIMENTS AND RESULTS:

To evaluate the proposed algorithm, We have taken an example and apply K means and I Kmeans.I have compare the result in terms of Quality factor which is the ratio of inter class similarity by intra class similarity. I take dataset from uci machine repository [6] for I K-Means Clustering .Data is of the Banking sector.

IV. Data Set Information:

The data is related with direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be (or not) subscribed.

K Means Clustering algorithm



I K Means clustering algorithm

FIGURE 1 Comparing results on (a) K Means (b) I K Means

V. RESULTS AND DISCUSSION:

We have compared results in figure 1(a) and 1(b).Now to understand, which clustering is better, there should be equal distribution of points in each cluster as much as possible and quality factor should be maximum. Quality factor can be defined in terms of maximizing the inter class similarity and minimizing the intra class similarity. Quality factor is ratio of inter class similarity by inter class similarity. As from figure 1(a) and 1 (b) we came to know that in I K Means equal distribution of data compare to K Means and quality factor in I K Means is better than K Means.

Algorithm	Quality Factor
K-Means	.28
I K Means	.50

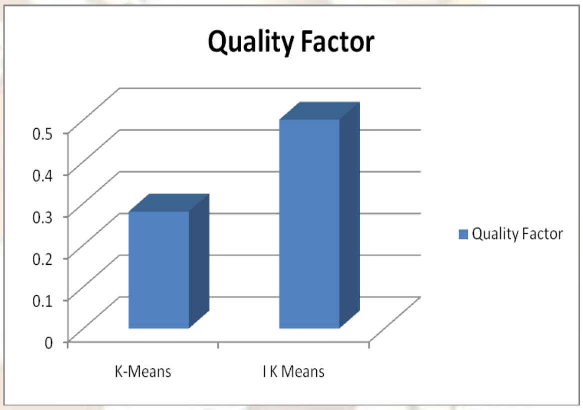


Figure 2(Quality factor of Algorithms)

VI. CONCLUSION

We have proposed an improved approach through which same distance problem can be solved and better result can be obtained. Better result depend on two factors, First

One is equal distribution of data in each cluster and quality factor. In both factors I K Means is better than K Means .Quality Factor of I K Means is better than K Means.

References:

1. Selection of K in K-means clustering, D T Pham, S S Dimov, and C D Nguyen, Manufacturing Engineering Centre, Cardiff University, Cardiff, UK.site name
2. A Modified *k*-means Algorithm to Avoid Empty Clusters, Malay K. Pakhira ,Kalyani Government Engineering College Kalyani, West Bengal, INDIA.
3. Improved K-means Algorithm Based on Fuzzy Feature Selection, Xiuyun Li, Jie Yang, Qing Wang, Jinjin Fan, Peng Liu

School of Information Management and Engineering, Shanghai University of Finance & Economics.

4. Modified global k-means algorithm for clustering in gene expression Data sets, Adil M. Bagirov and Karim Mardaneh.

5. Jiawei Han and Micheline Kamber ,Data Mining: Concepts and Techniques, 2nd edition, The Morgan Kaufmann Series in Data Management Systems, Jim Gray, Series Editor Morgan Kaufmann Publishers, March 2006. ISBN 1-55860-901-6

6. UCI Repository of machine learning databases, University of California. archive.ics.uci.edu/ml/datasets/**Bank** Marketing(accessed on line)