International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-8 Issue-5, March 2019

An Optimal Approach of Initial Centroid Selection for Effective Clustering

T.Haribabu, I.Raju

Abstract*: Data is grouped together based on similarity this technique is called *clustering which very well known in datamining. For extracting use full data from cluster most of the people are using the algorithm K-Means. In K-Means approach selecting initial centroids is the problem & these centroids are selected randomly. Because of random centroids this algorithm re-iterate a many number of times. The K-Means algorithm Correctness depends much on the chosen central values. To enhance the performance of the K-Means one should not select the original centroids randomly these must be selected carefully. A new tactic to formulate the original centroids is proposed which improves the rapidity of clustering and cuts the computational complexity by reducing the number of iterations.

***Index Terms*: clustering, k- means, Euclidean distance.**

I. INTRODUCTION

Clustering is an unrestricted data inspection and data mining technique which group's objects such that objects within same clusters are similar to each other compared to the objects in other clusters. Grouping is an unaided learning and does not rely on default classes in these analysis, It first divide the group of data into bank based on data similarity and then attach the labels to the clusters. Clustering techniques have a wider implementation, in the areas of segmentation of images, information extracting, grouping of web pages, segmentation of market & engineering and scientific performance. There are innumerable approaches for clustering like hierarchical Approach, density-based approach, constraint-based approach, partitioning approach, model-based approach, grid-based approach,. In this project, we have chosen only partitioning clustering method.

The format of the remaining paper was organized as follows: next part 2 defined the traditional approach. Part 3 defines the proposed algorithm and its flowchart and continued to part 4 which shows the experimental results, dataset description, environmental setup followed by results analysis. Part 5 gives the conclusion and future work.

II. STANDARD K-MEANS

K-Means approach is a dividing up clustering manner which involves casual electing K original centroids where K is an end user desired clusters.

In clustering, it find the heterogeneity between points by determining the space between each pair of objects. These measures comprises of Manhattan, distance of Euclidean & distance of Minkowski.

Euclidean distance is defined as

$$\mathrm{dist}(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

Manhattan distance is defined as

$$d(x,y) = \left(\sum_{i=1}^{m}|x_i - y_i|^r\right)^{1/r}$$

Where x_i and y_i are dual non-dimensional objects.

r is a positive integer. This distance is known as L_r norm. It denotes the Manhattan distance when r=1 and Euclidean distance when r=2

A. Traditional Algorithm:

The K-means approach involves randomly selecting K initial groups here K is a user-defined number of preferred groups. Every point is assigned to the closest centroid and the collection of points close to a centroid form a cluster. The centroid gets updated according to the points in the group and these processes continue until the points stop changing their clusters. The algorithm can be summarized as follows.

Algorithm:

Input: Give the k value, no. of clusters,
D: a data set with n number of objects.

Outcome: K groups

Step 1: Initially take any k points as centroids.

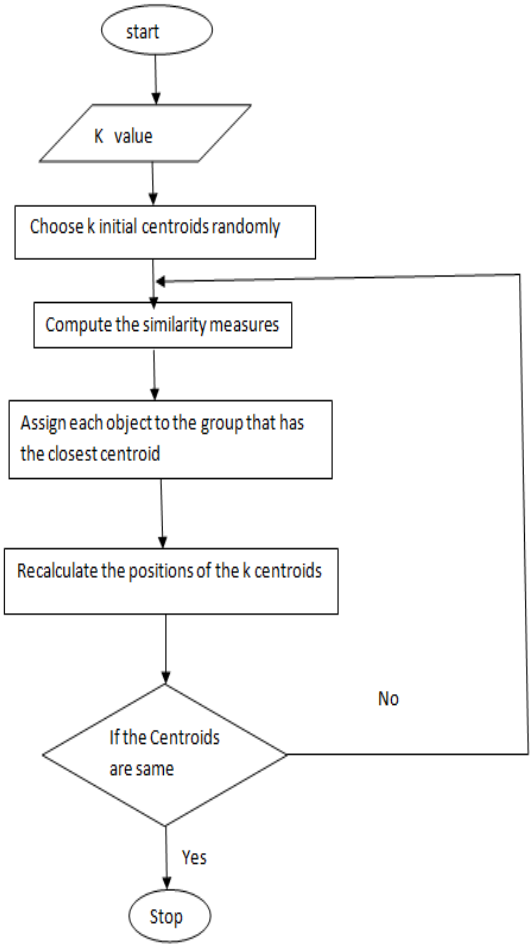
Step 2: Compute the Euclidean Distance between the tows of the data & k groups and cluster the data depended on the calculated distance.

Step 3: Calculate the average for respective co-ordinate element in each row for every group, the resultant average should be taken as initial centroids.

Step 4: Re-iterate step 2 and step3 until the centroids remain same.

B. Flowchart:

Manuscript published on 30 March 2019.
***Correspondence Author(s)**
T.Haribabu, Department of Computer Science and Engineering, VIEW College, Visakhapatnam, India.
I.Raju, Department of Computer Science and Engineering, VIEW College, Visakhapatnam, India.



III. PROPOSED METHOD:

The effectiveness of the K-Means approach depends immensely on the opted initial centroids. If the initial centroids are not chosen properly then it may result in increased number of algorithm iterations and time complexity. So this paper proposes an alternate approach of finding initial centroids which result in decreased number of iterations compared to traditional algorithm.

A. ALGORITHM:

Step 1: Select the dataset having n data rows.

Step 2: Add the elements in each and every row.

Step 3: Sort the dataset based on step 2 results in ascending order.

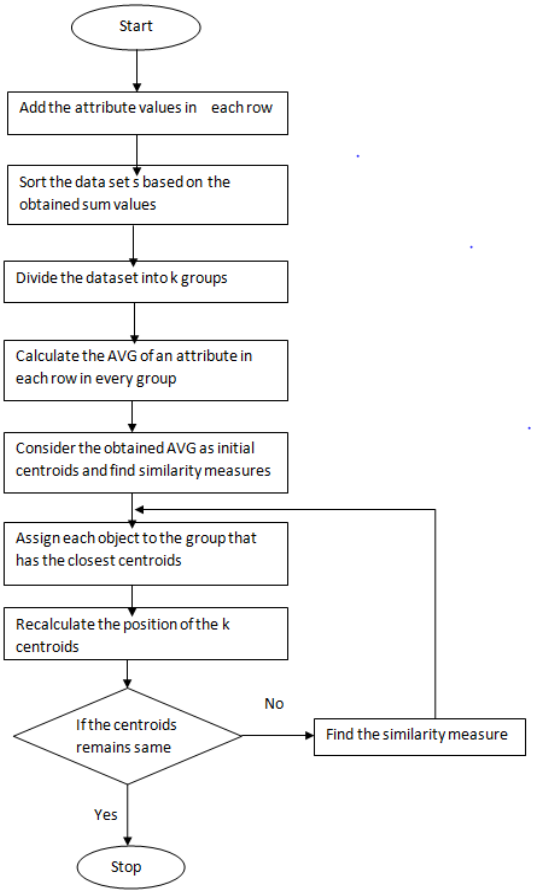
Step 4: Divide the datasets according to the user K value specification.

Step 5: Calculate the average for respective co-ordinate element in each row for every group, the resultant average should be taken as initial centroids.

Step 6: Compute the Euclidean Distance between the rows of the data& centroid results.

Step 7: Data can be grouped based on the calculated distance. Step 8: Repeat the process from step 5 until the clusters are similar with previous iteration.

B. Flowchart:



IV. EXPERIMENTAL RESULTS

A. *Datasets:*

Here we are using different data sets for correlate the performance of traditional k-means & the suggested model. These datasets are taken from UCI depository [9]. The representation of the datasets is displayed in Table1. Among different datasets, two of them are displayed here. The Iris dataset contains totally 150 data rows. The three varieties each having 50 data rows. The first 50 data rows belong to iris setosa, the next 50 data rows belong to iris versicolor, and the last 50 data rows belong to iris virginica. The wine dataset contains 178 data rows. Each data row contains the 14 features.

Dataset names	Attribute characteristics	No. of attributes	No. of instances
Iris	Integer, Real	4	150
Wine	Integer, Real	14	178

I Description of datasets

B. Experimental Setup

The new anticipated procedure is coded on the windows 10 desktop PC with Intel i3 processor, 2.20 GHz processor and 4GB RAM. We used an open source IDE called NetBeans to develop our program and it can run on any desktop installing NetBeans.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

C. Results Analysis

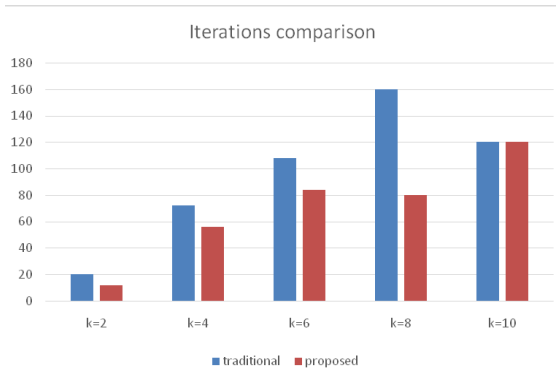


Fig 1: Result analysis of to Iris Dataset

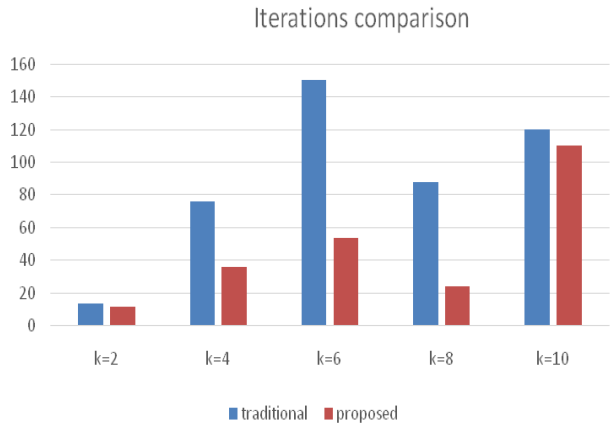


Fig 2: Result analysis of Wine Dataset

The iteration values are noted for the two algorithms with different k-values. Our proposed method forms the clusters with less number of iterations compared to the traditional method which shows that suggested model is better than traditional k-means method.

Fig.1. Present the representation of performance of proposed k-means and traditional k-means on different k-values for the chosen Iris Dataset.

Fig.2. shows the representation of performance of proposed k-means and traditional k-means on different k-values for the chosen Wine Dataset.

From fig.1 and fig.2 its clear that proposed is better than traditional k-means.

V. CONCLUSION AND FUTURE WORK

As the performance of the k-mean clustering method is more likely dependent on the chosen initial centroids, so there should be a standard method to choose the initial centroids which makes the k-mean algorithm to unique clustering results in fewer number of iteration. The proposed algorithm uses basic k-means to solve the problem of initial centroids. This paper comes up with new approach to find the initial centroids. We also compare the proposed k-means with basic k-means. The outcome shows that recommend procedure is having the less number of iteration values compared to traditional k-means. It concludes that proposed method is more efficient than traditional k-means. In Future work we use more standard datasets to evaluate the proposed algorithms.

REFERENCES:

1. Improvement of k-means based on weighted average by Md. Sohrav Mahmud, Md. Mostafizer and Md. Nasim Akhtar (20-22 December, 2012).
2. Amira Boukhdhir Oussama Lachiheb, Mohamed Sala Gouider. "An improved Map Reduce Design of Kmeans for clustering very large datasets", IEEE transaction.
3. V. Duon, M. Phayung. "Fast K-Means Clustering for very large datasets based on Map Reduce Combined with New Cutting Method (FMR KMeans)", Springer International Publishing Switzerland, 2015.
4. C. Xiaoli and al. "Optimized big data K-means clustering using Map Reduce", Springer Science + Business Media New York (2014).
5. Yugal Kumar and G. Sahoo, "A New Initialization Method to Originate Initial Cluster Centers for K-Means Algorithm", International Journal of Advanced Science and Technology Vol.62, (2014).
6. Yuan F, Meng Z.H, Zhang H.X and Dong C.R, "A New Algorithm get the initial Centroids," Proc. Of the 3rd international Conference on Machine learning and Cybernetics, pages 26-29, Aug 2004.
7. Neha Aggarwal et al., "A mid-point based k-means clustering algorithm" International Journal on Computer Science and Engineering (IJCSE), Vol 4 No. 06 June 2012, ISSN: 09753397.
8. Selection of initial centers for k-means algorithm by Anand M. Baswade and Prakash S. Nalwade (7 July, 2015).

AUTHORS PROFILE

Mr. T.Haribabu, Assistant professor in the department of Computer Science, Vignan's Institute Of Engineering For Women for the past 2 years and having efficient knowledge in Data Mining.

Mr. I.Raju, Assistant professor in the department of Computer Science, Vignan's Institute Of Engineering For Women for the past 3 years and having efficient knowledge in Data Mining and IoT.

.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)