An Improvement in K-mean Clustering Algorithm Using Better Time and Accuracy

Er. Nikhil Chaturvedi[1] and Er. Anand Rajavat[2]

[1]Dept.of C.S.E, S.V.I.T.S, Indore (M.P)and [2] Asst. Prof, S.V.I.T.S, Indore

Abstract:

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters).K-means is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. The process of k means algorithm data is partitioned into K clusters and the data are randomly choose to the clusters resulting in clusters that have the same number of data set. This paper is proposed a new K means clustering algorithm we calculate the initial centroids systemically instead of random assigned due to which accuracy and time improved.

Key words: *Data mining, Cluster, Basic K-means algorithm, Improved K-means algorithm.*

I INTRODUCTION

data mining refers to using a variety of techniques it identify suggests of information or decision -making knowledge in the database and extracting these in such a way they can use different area such as decision support forecasting. The data is often voluminous. It is the hidden information in the data that is useful. Data mining depends on effective data collection and warehousing as well as computer processing. Data mining involves the use of sophisticated data analysis tools to discover previously unknown, valid pattern and relationship in large data set [1].

Clustering is one of the tasks of data mining. Clustering [2] is useful technique for the discovery of data distribution and patterns in the underlying data. The aim of clustering is to discover both dense and sparse regions in data set. The main two approaches to clustering - hierarchical clustering and partitioning clustering [3] K-mean algorithm is main categories of partitioning algorithm. The main difference between partitioned and hierarchical clustering is that, in partitioned clustering algorithm data is partitioned into more than two subgroups in hierarchical clustering algorithm data is divided into two subgroups in each step. K-mean clustering is a partitioning clustering technique in which clusters are formed with the help of centroids. On the basis of these centroids, clusters can vary from one another in different iterations. Moreover, data elements can vary from one cluster to another, as clusters are based on the random numbers known as initial centroids.

A new algorithm is introduced and implemented in research. The whole paper is organized in this way. First the basic K-mean clustering algorithm is discussed and then proposed K-mean

DOI : 10.5121/ijpla.2013.3402

clustering algorithm is explored. The implementation work and the results of experiments are followed by the comparison of both algorithms.

II RELATED WORK

In [4] the research of k-Means clustering algorithm is one of the clustering algorithms which have lot of used in applications because of its simplicity and implementation. K-Means algorithm's is less accuracy because of selection of k initial centers is randomly. Therefore, in this paper surveyed different approaches for initial centers selection for k-Means algorithm. Comparative analysis of Original K-Means and improved k-Means Algorithm.

In [5] this paper study of different approaches to k- Means clustering, and analysis of different datasets using Original k-Means and other modified algorithms.

In [6] this paper main aim to reduce the initial centroid for k mean algorithm. This paper proposed Hierarchical K-means algorithm. It uses all the clustering algorithm results of K-means and reaches its local optimal. This algorithm is used for the complex clustering cases with large numbers of data set and many dimensional attributes because Hierarchical algorithm in order to determine the initial centroids for K-means.

In [7] researchers introduced K mean clustering algorithm.
This paper proposes method for the making K-means clustering algorithm more efficient and effective. In this paper time complexity improve using the unique data set.

III BASIC K-MEAN CLUSTERING ALGORITHM

K means clustering [8] is a partition-based cluster analysis method. According to this algorithm we firstly select k data value as initial cluster centers, then calculate the distance between each data value and each cluster center and assign it to the closest cluster, update the averages of all clusters, repeat this process until the criterion is not match.

K means clustering aims to partition data into k clusters in which each data value belongs to the cluster with the nearest mean. Figure 1 shows how to process of the basic K mean clustering algorithm [9] in steps.

Basic K-mean algorithm:

Initially, we are chose K number of clusters in algorithm.

The first step is to choose a set of K objects as centres of the clusters. Often chosen such that the objects are in distance basis how to one or more further away.

In the next step of the algorithm considers each object and assigns it to the cluster which is closest.

After that the cluster centroids are recalculated after each Object assignment, or after the whole cycle are completed.

This process is repeat until the all object are assign to its clusters.

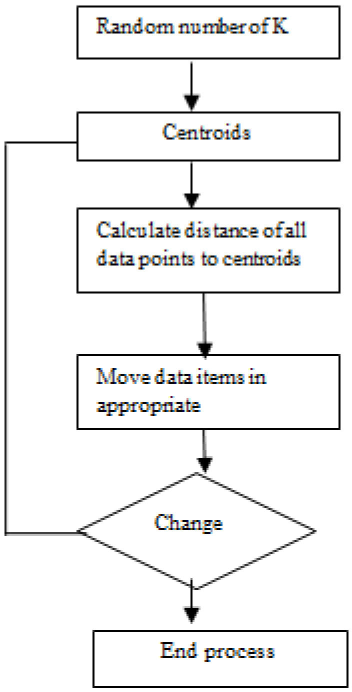
Random number of K
Centroids
Calculate distance of all data points to centroids
Move data items in appropriate
Change
End process

Fig. 1 K mean Algorithm Process

IV PROPOSED ALGORITHM

In the proposed clustering method discussed in this paper, for the original k-means algorithm is modified to improve the time and accuracy.

Input:
D is the set of all the data items
k is number of clusters

Output:
A set of k clusters.
Steps:

Phase 1: For the initial centroids

Input:
D // set of n data
k // Number of clusters

Output: A set of k initial centroids.

```
Steps:
1. Set p = 1;
2. Measure the distance between each data and all other data in the set D;
3. Find the closest pair of data from the set D and form a data set Ap (1<= p <= k) which contains these two data, Delete these two data from the set D;
4. Find the data in D that is closest to the data set Ap, Add it to Ap and delete it from D;
5. Repeat step 4 until the number of data in Ap reaches all data in D;
6. If p<k, then p = p+1, find another pair of data  from D between which the distance is the small form another data set Ap and delete them from D, Go to step 4;
7. for each data-point set Ap (1<=p<=k) find the mean of data in Ap.
     These means will be the initial centroids.
```

Phase 2: Data to the clusters

```
Input:
D // set of n data.
C // set of k centroids
Output:
A set of k clusters
Steps:
1. Compute distance between each data to all the centroids
2. for each data di find the closest centroid ci and assing to cluster j.
3. Set ClusterCL[i] = j; // j:CL of the closest cluster
4. Set Shorter_Dist[i] = d (di, cj);
5. For each cluster j (1<=j<=k), recalculate the centroids;
6. Repeat
7. for each data di,
    7.1 Compute the distance from the centroids of the closest cluster;
    7.2 If distance is less than or equal to the present closest distance, the data-point stays in cluster;
    Else
            7.2.1 For every centroids compute the distance.
End for;
7.2.2 Data di assign to the cluster with the closest centroid cj
7.2.3 Set ClusterCL[i] =j;
7.2.4 Set Shorter_Dist[i] = d (di, cj);
End for;
8. For each cluster j (1<=j<=k), recalculate the centroids; until the criteria is met.
```

In the first phase, we determine initial centroids systematically. The second phase use of functions of the clustering method. It starts by the initial clusters based on the distance of each data from the initial centroids. These clusters are finding by using a heuristic approach, thereby improving the accuracy. In this phase compute the distance between the data and all other data from the data set. Then find out the closest pair of data and form a set A1 consisting of these two data, and delete them from the data set D. After that evaluate the data which is closest to the set A1, add it to A1 and delete it from D. Repeat this process until the all the element in the set A1 completed. Then go back to the second step and form another data set A2. Repeat this till 'k' such sets of data are obtained. Finally the initial centroids are obtained by averaging all the data in each data set. The Euclidean distance is used for determining the close of each data to the cluster centroids. The initial centroids of the clusters are used as input for the second phase, and assigning data to appropriate clusters.

The first step in the second phase is to determine the distance between each data and the initial centroids of all the clusters. After that the data are assigned to the clusters having the closest centroids. This gives the results as initial grouping of the data. For each data the cluster to which it is assigned (ClusterCL) and its distance from the centroid of the nearest cluster (Shorter_Dist) are noted. For each cluster calculated the mean of the data values for the centroids. Until this step, the process is similar as original k-means algorithm except that the initial centroids are computed systematically.

At time of iteration, the data may get redistributed to different clusters. The method involves distance between each data and the centroid of its present nearest cluster. At the time of starting the iteration, the distance of each data from the new centroid of its present nearest cluster is determined. If present distance is less than or equal to the previous nearest distance, that is an indication that the data point stays in that cluster itself and there is no need to computation of distance. This result save our time for required computing the distances to k-1 clusters centroids. On the other hand, if the new centroid of the present closest cluster is more distance from the its previous centroid, there is a chance for the data getting included in another nearer cluster. In that case, it is required to computation of distance from the centroids. New nearest cluster and record are identify for the new value of the nearest distance. The loop is repeated until no more data cross cluster boundaries.

V EXPERIMENTAL RESULTS

We apply both the algorithms original and proposed for the different number of records. Both the algorithms original and proposed need number of clusters as an input. In the basic K-means clustering algorithm set of initial centroids are required. The proposed method finds initial centroids systematically. The proposed method requires only the data and number of clusters as inputs.

The basic K-means clustering algorithm is executed sevan times for the different data values of initial centroids . In each experiment the time was computed and taken the average time of all experiments. Table 1 shows the performance comparison of the Basic and proposed k-mean clustering algorithms. The experiments results show that the proposed algorithm is producing better results in less amounts of computational time and accuracy compared to the basic k-means algorithm.

Table 1. Performance Comparison

No. of Records	No. Of Cluster	Algorithm	Run	Accuracy	Execution time in sec
100	**K=3**	**Basic k-mean**	**7**	**70.14**	**0.103**
		Proposed k-mean	**1**	**82.66**	**0.083**
200	**K=3**	**Basic k-mean**	**7**	**70.10**	**0.128**

		Proposed k-mean	1	82.11	0.098
300	K=4	Basic k-mean	7	70.57	0.144
		Proposed k-mean	1	82.31	0.122
400	K=2	Basic k-mean	7	70.44	0.162
		Proposed k-mean	1	82.34	0.142
500	K=2	Basic k-mean	7	70.12	0.206
		Proposed k-mean	1	82.19	0.186

Figure 2 depicts the performances of the original k-means algorithm and the proposed algorithm in terms of the accuracy and time.

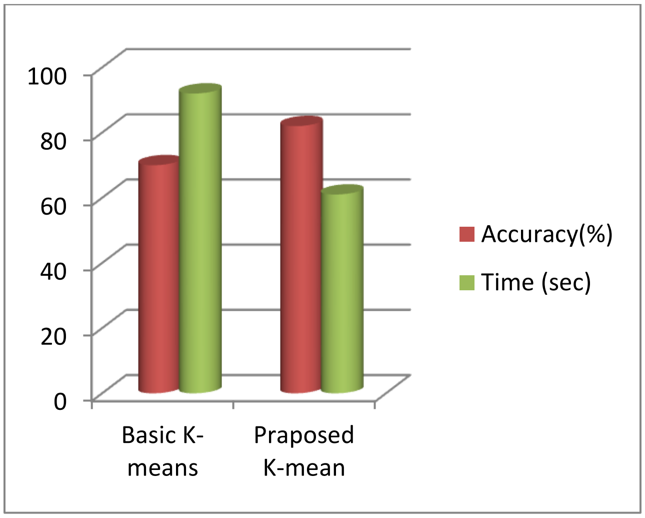

Fig. 2Time and Accuracy of the algorithms

VI TECHNOLOGY USED

For the experiment of proposed K- mean clustering algorithm we are used the NetbeansIDE6.7 and the output of the results show with the help of text area present in the window developed for the experiment. All the experiments held in tools will be performed on a 2.40GHz Intel(R) Core (TM) i5-2430 MB memory, 2 GB RAM running on the windows XP Professional OS. Programs will be coded in the java programming language.

VII CONCLUSION

In this paper, the improved algorithm of K-means clustering algorithm is proposed to overcome the deficiency of the classical K-means clustering algorithm. The classical K-means algorithm use the selecting the initial centroids approach .This algorithm performs well only when the data sets are shorts and it suffers from increased number of data are more initial centroid problem. The new proposed method use the systemically finds initial cenroid which reduces the number of data base scans and it is useful for large amount of data base scan. This method ensures the entire process of clustering time will be reduced in execution process.

REFERENCES

[1] A. K. Jain and R. C. Dubes, Algorithms for Clustering Data. Englewood Cliffs, NJ: Prentice-Hall, 1988.

[2] S. Z. Selim and M. A. Ismail, “K-means type algorithms: a generalized convergence theorem and characterization of local optimality,” in IEEE Transaction on Pattern Analysis and Machine Intelligence, vol. 6, No. 1, pp. 81--87, 1984.

[3] Jieming Zhou, J.G. and X. Chen, "An Enhancement of K-means Clustering Algorithm," in Business Intelligence and Financial Engineering, BIFE '09. International Conference on, Beijing, 2009.

[4] M.P.S Bhatia, Deepika Khurana Analysis of Initial Centers for k-Means Clustering Algorithm International Journal of Computer Applications (0975 – 8887) Volume 71– No.5, May 2013

[5] Dr. M.P.S Bhatia1 and Deepika Khurana Experimental study of Data clustering using k- Means and modified algorithms International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.3, No.3, May 2013

[6] Kohei Arai and Ali Ridho Barakbah Hierarchical K-means: an algorithm for centroids initialization for K-means Rep. Fac. Sci. Engrg. Reports of the Faculty of Science and Engineering, Saga Univ. Saga University, Vol. 36, No.1, 2007 36-1 (2007),25-31

[7] Napoleon, D. and P.G. Lakshmi An efficient K-Means clustering algorithm for reducing time complexity using uniform distribution data points , 2010 in Trendz in Information Sciences and Computing (TISC), Chennai

[8] S. Ray, and R. H. Turi, “Determination of number of clusters in kmeans clustering and application in colour image segmentation, ”In Proceedings of the 4th International Conference on Advances in Pattern Recognition and Digital Techniques, 1999, pp.137-143.

[9] 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.

[10] Dong, J. and M. Qi, "K-means Optimization Algorithm for Solving Clustering Problem," in Second International Workshop on Knowledge Discovery and Data Mining (WKDD), Moscow 2009.