Enhancing the Exactness of K-Means Clustering Algorithm by Centroids

Mahesh Singh, Sangeeta Raheja, Aashima

Abstract— **paper reflect the comparative study of k-means clustering algorithm with the mean based initial centroids. The original k-means algorithm helps to calculate the distance between the data objects, but the difficulty is that the k-means algorithm is not efficient and accurate to calculate distance and henceforth make clusters. K-means algorithm needs lots of iterations to make clusters. In this paper, centroids based algorithm is used to avoid lots of iterations. This systematic method will help to give the global optimal result for any set of data objects. The limitation of k-means algorithm will be removed by using initial centroids.**

Index Terms— **Clustering,Centroids, Data Mining, K-mean**

I. INTRODUCTION

Clustering is one of the most useful tasks in data mining process for discovering groups and identifying interesting distributions and patterns in the underlying data. Clustering problem is all about partitioning a given data set into groups (clusters) such that the data points in a cluster are more similar to each other than points in different clusters. For example, consider a retail database records containing items purchased by customers. A clustering method could group the customers in such a way that customers with similar buying patterns are in the same cluster. Thus, the main point in the clustering process is to reveal the organization of patterns into “sensible” groups, which allow us to find out similarities and differences, as well as to get useful conclusions about them.

This idea is applicable in many fields, like in life sciences, medical sciences and engineering. Clustering may be found under different names in different contexts, such as unsupervised learning (in pattern recognition), numerical taxonomy (in biology, ecology), typology (in social sciences) and partition (in graph theory).

Existing clustering algorithms are broadly classified into Hierarchical and Partitioning clustering algorithms. Hierarchical algorithms decompose a database D of n objects into several levels of nested partitionings (clusterings), represented by a dendrogram, a tree that iteratively splits D into smaller subsets until each subset consists of only one object. There are two types of hierarchical algorithms; an *Agglomerative* that builds the tree from the leaf nodes up, whereas a *Divisive* builds the tree from the top down. Partitioning algorithms makes a single partition of a database D of n objects into a set of k clusters, so that the objects in a cluster are more similar to each other than to objects in different clusters.

In *k*-means algorithm, the prototype, called the center, is the mean value of all objects belonging to a cluster. The *k*-modes algorithm extends the *k*-means paradigm to categorical domains. For *k*-medoid algorithms, the prototype, called the “medoid”, is the most centrally located object of a cluster. The algorithm CLARANS, is an improved k-medoid type algorithm restricting the huge search space by using two additional user-supplied parameters. It is significantly efficient than the well-known *k*-medoid algorithms PAM and CLARA.

We have analyzed different variants of k-means clustering algorithm. The results of different variants are analyzed with a scenario. The k-means method has been shown to be effective in producing good clustering results for many practical applications. However, the k-means algorithm generally converges at a local optima. This affects the quality of end clustering results. This is so because the k-means algorithm is heavily dependent on the selection of initial centroids, which are actually selected randomly in the beginning of the algorithm.

We have developed k-means clustering algorithm with Mean-Based Initial Centroids. This algorithm removes the limitation of terminating of the k-means algorithm at local optima. It also makes the k-means algorithm to be applied to a wide variety of input data sets. We have taken two scenarios for comparing the results of the proposed algorithm with the Midpoint-Based Initial Centroid algorithm. We find that the proposed algorithm works better

II. BASIC CONCEPT OF DATA MINING

Data mining refers to extracting or “mining” knowledge from huge amounts of data stored in either databases, data warehouses or other information repositories.

The result of data mining is useful information or knowledge which can be used for applications ranging from business management, production control, market analysis to engineering design and science exploration.

A. DATA MINING TECHNIQUES

The kinds of patterns that can be discovered depend upon the data mining tasks employed. There are two types of data mining tasks: **descriptive** data mining tasks that describe the

Manuscript received May 20, 2014.

Mahesh Singh, Assoc. Member IEEE AITM Students Branch, Branch Code: 09831, Asst. Professor CSE Department. Advanced Institute of Technology & Management, Palwal

Sangeeta Raheja, Student, M.Tech (CSE), Advanced Institute of Technology & Management, Palwal

Aashima, Student, M.Tech (CSE), Advanced Institute of Technology & Management, Palwal.

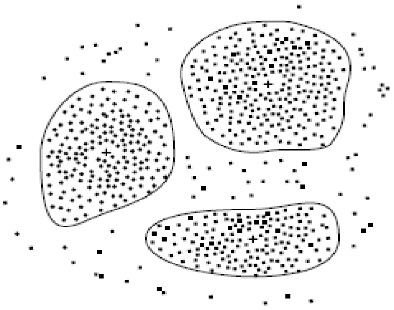
general properties of the existing data, and **predictive** data mining tasks that attempt to do predictions based on inference on available data.

In some cases, we may have no idea regarding what kinds of patterns in their data may be interesting, and hence may want to search for several different kinds of patterns in parallel. Therefore it is important to have a data mining system that can mine multiple kinds of patterns to accommodate different user expectations or applications. Data mining systems should be able to discover patterns at various granularity (i.e., different levels of abstraction).

- ▶ Association Analysis
- ▶ Classification and Prediction
- ▶ Cluster Analysis

Cluster Analysis

- ▶ Unlike classification and prediction, which interpretes class-labeled data objects, clustering analyzes data objects without consulting a known class label.



Cluster Analysis

- ▶ A 2-D plot of customer data with respect to customer locations in a place, showing three data clusters. Each cluster “center” is marked with a “+”.
- ▶ In general, the class labels are not present in the training data simply because they are not clustered or grouped based on the principle of maximizing the intraclass similarity and minimizing the interclass similarity. clusters of objects are formed so that objects within a cluster have high similarity in comparison to one another, but they are very dissimilar to objects in other clusters. Each cluster that is formed can be seen as a class of objects,from which rules can be derived. Clustering can also facilitate taxonomy formation, that is, the organization of observations into a hierarchy of classes that group similar events together.

III. PROBLEM FORMULATION/ AREA OF STUDY

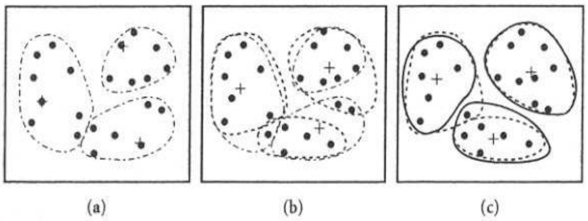
The work done in my paper is based on k-means Clustering Algorithm. Clustering is a technique of Data Mining which aims at grouping a set of data objects into a certain number of clusters according to some similarity/dissimilarity measure. K-means is a well known and widely used algorithm used for clustering, but it has certain limitations. In my work I have tried to remove one of the prime limitations of k-means algorithm. My algorithm proves to be better than the original k-means algorithm and a further research done in this direction.

IV. CLUSTERING ALGORITHMS CATEGORIES

A clustering algorithm is an algorithm that takes a data set, and produces some clustering of the data set. The basic goal of any clustering algorithm is to generate clusters that contain similar objects. A multitude of clustering methods are proposed in the literature. Clustering algorithms can be divided according to:

- The type of data input to the algorithm.
- The clustering criterion defining the similarity between data points.
- The theory and fundamental concepts on which clustering analysis techniques are based (e.g. fuzzy theory, statistics).
- **k-Means Clustering Algorithm:** k-Means is a clustering algorithm that deals with numerical attribute values (NAs) , although it can also be applied to categorical datasets with binary values, by seeing the binary values as numerical. The k-Means clustering technique for numerical datasets requires the user to specify the number of clusters to be produced and the algorithm builds and refines the specified number of clusters. During the k-Means clustering algorithm for numerical datasets, the following generic loop is performed:

```
1. Insert the first k objects into k new clusters.
2. Calculate the initial k means for k clusters.
3. For each object o
        a. Calculate the dissimilarity between o and the means of all clusters.
        b. Insert o into the cluster C whose mean is closest to o.
4. Recalculate the cluster means so that the cluster dissimilarity between mean and objects is minimized.
5. Repeat 3 and 4 until no or few objects change clusters after a full cycle test of all the objects.
```



Clustering using the k-Means Method.

Limitations

- A weakness is that it often terminates at a local optimum rather than a global optimum.
- The number of clusters k needs to be specified in advance by the user.
- K-means is unable to handle noisy data and outliers and it is not suitable to discover clusters with non-convex shapes. Finally,
- The results depend on the order of the objects in the input dataset as different orderings will produce different results.

V. SCENARIO BASED ON K-MEAN ALGORITHM :

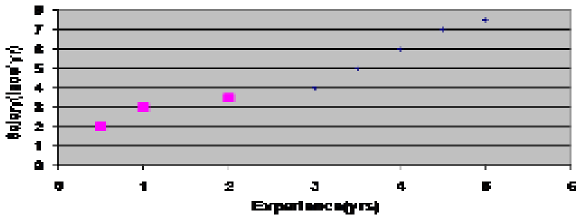
Following is an example of original k-mean clustering in which the centroids are taken randomly.

Suppose we have several objects (8 Employees of an Organization) and each object has two attribute features as shown in table below. Our goal is to group these objects into K=3 groups based on the two features (Experience in no. of yrs and Annual Salary).

Employee Data Set

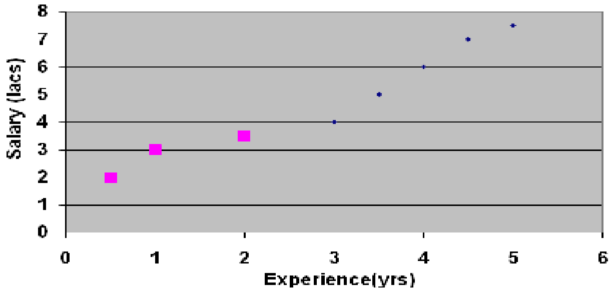
EMPLOYEE	ATTRIBUTE1:X (Experience in No. of yrs)	ATTRIBUTE2:Y (Salary in Lacs/annum)
Emp1	0.5	2
Emp2	1	3
Emp3	2	3.5
Emp4	3	4
Emp5	3.5	5
Emp6	4	6
Emp7	4.5	7
Emp8	5	7.5

Each employee represents one point with two attributes (X, Y) that we can represent in coordinate in an attribute space as shown in figure 4.1



Data Set with three initial random centroids

Initial value of centroids : Suppose we use Emp1, Emp2 and Emp3 as the first centroids. Let $c_{1,}$ c_2 , c3 denote the coordinate of the centroids, then c_1 (0.5, 2) ,c_2 (1,3) , c3(2, 3.5).



Dataset with initial centroids

1. **Objects-Centroids distance:** we calculate the distance between cluster centroid to each object. Let us use Euclidean distance, then we have distance matrix at iteration 0 is

D0=

$$\begin{bmatrix} 0 & 1.118 & 2.12 & 3.2 & 4.24 & 5.315 & 6.4 & 7.1 \\ 1.118 & 0 & 1.118 & 2.23 & 3.2 & 4.24 & 5.315 & 6.02 \\ 2.12 & 1.118 & 0 & 1.118 & 2.12 & 3.2 & 4.35 & 5 \end{bmatrix}$$

Each column in the distance matrix symbolizes the object. The first row of the distance matrix corresponds

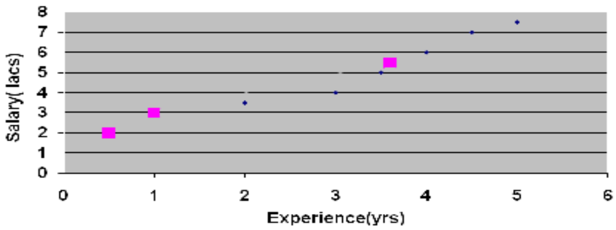
to the distance of each object to the first centroid and the second row is the distance of each object to the

second centroid. Similarly for third.

2. **Objects Clustering:** We assign each object based on the minimum distance. Thus, Emp1 is assigned to group 1, Emp2 to group 2, Emp3, Emp4 …..Emp8 to group 3 . The element of Group matrix below is 1 if and only if the object is assigned to that group.

$$G1=\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

3. **Iteration-1, determine Centroids:** Knowing the members of each group, now we compute the new centroid of each group based on these new memberships. Group 1 only has one member thus the centroid remains in c_1 (0.5,2). Group 2 also has one member, thus the centroid remains C2 (1,3). Group 3 has 6 members so the cenroid becomes the average coordinate among the six members:C3= ((2+3+3.5+4+4.5+5)/6 , (3.5+4+5+6+7+7.5)/6)) = (3.66 , 5.5)



Dataset after Iteration 1

4. **Iteration-1, Objects-Centroids distances:** The next step is to compute the distance of all objects to the new centroids. Similar to step 2, we have distance matrix at iteration 1 is

D1=

$$\begin{bmatrix} 0 & 1.118 & 2.12 & 3.20 & 4.24 & 5.315 & 6.4 & 7.1 \\ 1.118 & 0 & 1.118 & 2.236 & 3.2 & 4.24 & 5.315 & 6.02 \\ 4.71 & 3.65 & 2.76 & 1.64 & 0.525 & 0.604 & 1.72 & 2.41 \end{bmatrix}$$

5. **Iteration-1, Objects clustering :**Similar to step 3, we assign each object based on the minimum distance. Based on the new distance matrix, we move Emp3 to Group 2 while all the other objects remain as before. The Group matrix is shown below

$$G1= \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

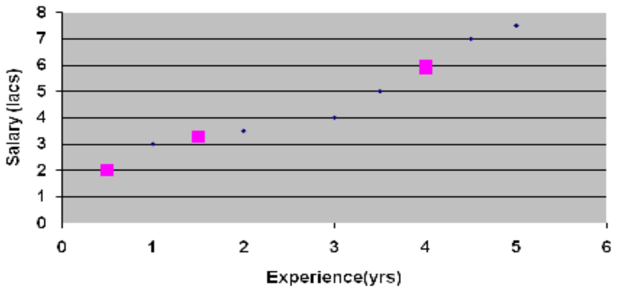
6. **Iteration 2, determine centroids:** Now we repeat step 4 to calculate the new centroids coordinate based on the clustering of previous iteration.

C1= (.5 ,2) C2= (1.5 , 3.25) C3= (4 , 5.9)

7. **Iteration-2, Objects-Centroids distances:** Repeat step 2 again, we have new distance matrix at iteration 2 as

D2=

$$\begin{bmatrix} 0 & 1.118 & 2.12 & 3.2 & 4.24 & 5.315 & 6.4 & 7.1 \\ 1.6 & 0.56 & 0.56 & 1.67 & 2.66 & 3.72 & 4.8 & 5.5 \\ 5.24 & 4.17 & 3.12 & 2.15 & 1.03 & 0.1 & 1.21 & 1.88 \end{bmatrix}$$



Dataset after Iteration2

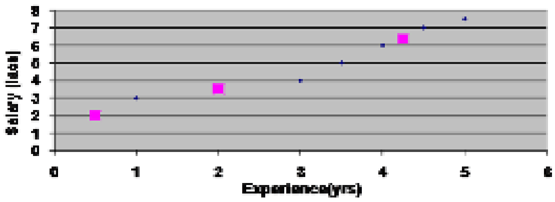
8. **Iteration-3, determine centroids:**

C1 = (0.5 ,2) C2 = (2, 3.5) C3 = (4.25 , 6.37)

9. **Iteration-3, Objects-Centroids distances:**

D2=

$$\begin{bmatrix} 0 & 1.118 & 2.12 & 3.2 & 4.24 & 5.315 & 6.4 & 7.1 \\ 2.12 & 1.118 & 0 & 1.118 & 2.12 & 3.2 & 4.3 & 5 \\ 5.76 & 4.66 & 3.65 & 2.68 & 1.56 & 0.45 & 0.67 & 1.35 \end{bmatrix}$$



Dataset after Iteration 3

11. **Iteration-3, Objects clustering:**

$$G3= \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{bmatrix}$$

We obtain result that **G3 = G2**. Comparing the grouping of last iteration and this iteration reveals that the objects does not move group anymore. Thus, the computation of the k-mean clustering has reached its stability and no more iteration is needed. We get the final grouping as the results shown in Table

Clustering Results

EMPLOYEE	ATTRIBUTE1:X (Experience in yrs)	ATTRIBUTE2: Y (Salary Lacs/annul)	GROUP
Emp1	0.5	2	1
Emp2	1	3	2
Emp3	2	3.5	2
Emp4	3	4	2
Emp5	3.5	5	3
Emp6	4	6	3
Emp7	4.5	7	3
Emp8	5	7.5	3

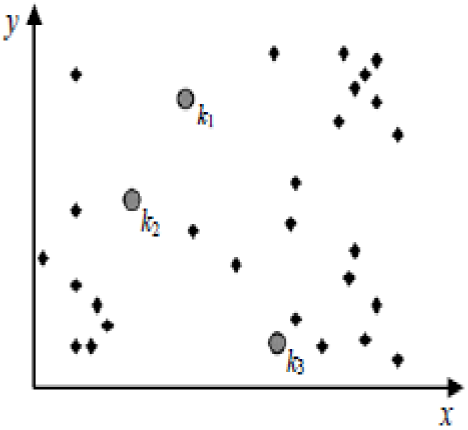
VI. THE LIMITATIONS OF ORIGINAL K-MEANS ALGORITHM

- It is computationally very expensive as it involves several distance calculations of each data point from all the centroids in each iteration.
- The final cluster results heavily depends on the selection of initial centroids which causes it to converge at local optimum.

VII. AN EFFICIENT ENHANCED K-MEAN CLUSTERING TECHNIQUE

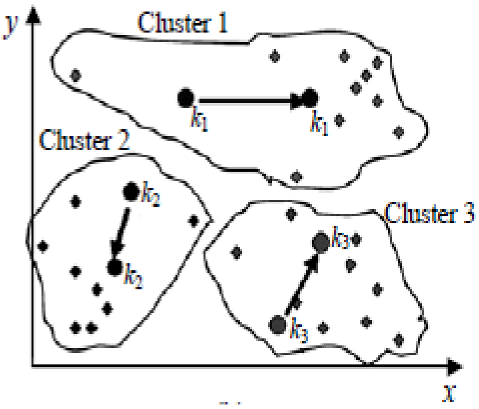
The following algorithm makes k-means more efficient by removing the first limitation i.e. it limits the number of computations to some extent. The idea makes k-means more efficient, especially for dataset containing large number of clusters. Since, in each iteration, the k-means algorithm computes the distances between data point and all centers, this is computationally very expensive especially for huge datasets. Therefore, we do can benefit from previous iteration of k-means algorithm. For each data point, we can keep the distance to the nearest cluster. At the next iteration, we compute the distance to the previous nearest cluster. If the new distance is less than or equal to the previous distance, the point stays in its cluster, and there is no need to compute its distances to the other cluster centers. This saves the time required to compute distances to k−1 cluster centers.

Following fig. explains the idea.



Initial Centroids to a dataset

shows points distribution over the initial 3 centroids, and the new centroids for the next iteration.



Recalculating the position of the centroids

VIII. ENHANCING K-MEANS WITH IMPROVED INITIAL CENTER USING MID-POINT METHOD

In this algorithm a systematic method to determine the initial centroids is explained. This method is quite efficient to produce good clusters using k-mean method, as compared to taking the initial centroids randomly.

Steps:

1. In the given data set D, if the data points contain both the positive and negative attribute values then goto step 2, else goto step 4.
2. Find the minimum attribute value in the given dataset D.
3. For each data point attribute, subtract with the minimum attribute value.
4. For each data point calculate the distance from origin.
5. Sort the distances obtained in step 4. Sort the data points in accordance with the distances.
6. Partition the sorted data points into k equal sets.
7. In each set , take the middle point as the initial centroid.

In the above algorithm, if the input data set contains the negative value attributes, then all the attributes are transformed to positive space by subtracting each data point attribute with the minimum attribute value in the data set. This transformation is required because in the algorithm the distance from origin to each data point is calculated. So if there are both positive and negative values, then for different data points same Euclidean distance will be obtained which will result in incorrect selection of initial centroids. After transforming all attribute values to positive, next step is to calculate the distance of each point from the origin. Then the original data points are sorted into k equal sets. In each set, the mid-point is calculated. All the mid-points are taken as the initial centroids.

Example: Scenario of the systematic method of determining initial centroids

Following is the example of k-mean algorithm using the enhanced method. The input data set contains 16 entities(condensing machines) which are described by two attributes:- Condensing temperature and corresponding pressure. The input parameter k is taken as 4. i.e. all the 16 entities have to be categorized into 4 clusters based on their efficiency.

Data Set of Condensing Machines

MACHINES	X ATTRIBUTE	Y ATTRIBUTE
Machine 1	15	58
Machine 2	50	93
Machine 3	25	130
Machine 4	40	130
Machine 5	25	165
Machine 6	50	170
Machine 7	25	225
Machine 8	60	220
Machine 9	40	250
Machine 10	43	270
Machine 11	50	280
Machine 12	50	320
Machine 13	43	360
Machine 14	60	360
Machine 15	60	405
Machine 16	60	540

Step 1: There are no negative values in the given data set, so goto step 4.

Step 4 , 5 , 6 ,7 : After calculating the distance of each data point from origin, the distances and the corresponding data points are sorted. Then they are divided into 4 equal groups. Then the mid-point of each group is taken

Mid-point of each subset

Distance from origin	X attribute	Y attribute	Mid-point
59.91 105.59 132.38 136.01	15 50 25 40	58 93 130 130	(27.5,94)

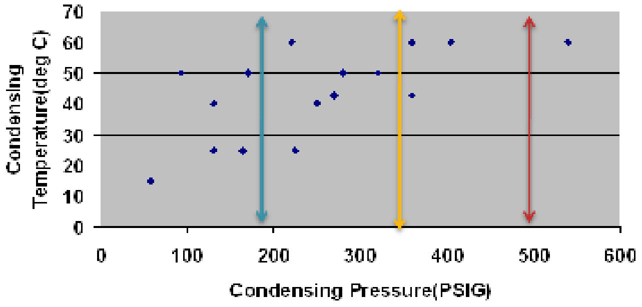
166.88	25	165	(42.5,192.5)
177.20	50	170	
226.38	25	225	
228.04	60	220	
253.18	40	250	(45,285)
273.40	43	270	
284.43	50	280	
323.88	50	320	
362.56	43	360	(51.5,450)
364.97	60	360	
409.42	60	405	
543.32	60	540	

Now using the calculated mid-points for each group as the initial 4 centroids, apply the k-mean algorithm on the input data. After three iterations of the k-mean algorithm, stability was achieved. The resulting clusters are as under

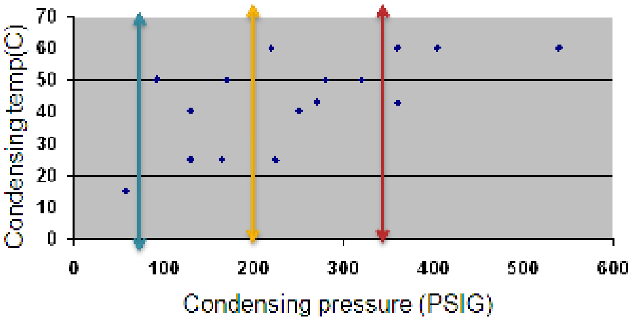
Clustering Results

MACHINES	X ATTRIBUTE	Y ATTRIBUTE	RESULTING CLUSTER
Machine 1	15	58	1
Machine 2	50	93	1
Machine 3	25	130	1
Machine 4	40	130	1
Machine 5	25	165	2
Machine 6	50	170	2
Machine 7	25	225	2
Machine 8	60	220	2
Machine 9	40	250	2
Machine 10	43	270	3
Machine 11	50	280	3
Machine 12	50	320	3
Machine 13	43	360	3
Machine 14	60	360	3
Machine 15	60	405	4
Machine 16	60	540	4

1. Comparison between Original K-Means, Midpoint Based K-Means and Mean Based K-Means



2. Original K-Means Method



Mean Based K-Means Method

IX. CONCLUSION AND FUTURE SCOPE

The proposed enhanced algorithm is easy to implement and it proves to be a better method to determine the initial centroids to be used in the k-means clustering algorithm. As the end clustering results of the k-mean clustering method are highly dependent on the selection of initial centroids , so there should be a systematic method to determine the initial centroids which makes the k-mean algorithm to converge in global optima and unique clustering results. This requirement is fulfilled by the proposed algorithm. Besides solving the problem of non-unique results, our proposed algorithm is also widely applicable to different types to problems. The problems with uniform as well as the problems with non-uniform distribution of data points are better addressed by our proposed algorithm.

Our proposed algorithm tries to enhance the k-means clustering algorithm by eliminating one of its drawback. But still lots of work needs to be done to enhance the k-means algorithm to a greater extent. K-means can be applied on numerical data only. But in day to day life we encounter scenarios with a combination of both numerical and categorical data values. So future work can be carried out in the direction of making the k-means algorithm applicable for mixed type of data.

REFERENCES

[1] Jiawei Han, MichelineKamber; Data Mining: Concepts and Techniques

[2] M. Halkidi, Y.Batistakis, M. Vazirgiannis; Clustering algorithms and validity measures : 0-7695-1218-6/01 2001 IEEE

[3] Rui Xu; Survey of Clustering Algorithms : IEEE Tansactions on Neural Networks, Vol 16, No. 3, May 2005

[4] Tian Zhang, Raghu Ramakrishnan, and MironLivny; BIRCH: An Efficient Data Clustering Method for Very Large Databases: Technical report, Computer sciences Dept., Univ. of Wisconsin Madison, 1996.

[5] Vladimir Estivill-Castro; Why so many clustering algorithm- A Position Paper: SIGKDD Explorations: Vol 4, Issue 1

[6] HesamIzakian;Clustering Categorical data using a Swarm-based method: 978-1-4244-5612-3/09 2009 IEEE

[7] AristidisLikas; the global k-means clustering algorithm: Pattern Recognition 36(2003).

[8] Rodrigo G.F. Soares; An Evolutionary approach for the Clustering data Problem : 978-1-4244-1821-3/08 2008 IEEE

[9] Yinghua Zhou, Hong Yu; A Novel k-means Algorithm for clustering and outlier detection: Second International conference on future information technology and management engineering: 2009 IEEE

[10] SudiptoGuha; ROCK: A robust clustering algorithm for categorical attributes: 0-7695-0071-4/99 1999 IEEE

[11] ĐinhQuangHuy and ĐinhMạnhTường :LINK-CONNECTED: A New Approach Of Clustering Algorithm For Categorical Attributes : Department of Computer Science and Engineering, Harbin institute of Technology, P.R.China, 2005.

[12] Maria Halkidi, YannisBatistakis, MichalisVazirgiannis; On Clustering Validation Techniques: Journal of Intelligent Information Systems, Vol. 17, pp. 107–145, 2001.

[13] T,Chiu, D.Fang, J.Chen, Y.Wang : A Robust and Scalable Clustering Algorithm for Mixed type attributes in large Database environment : Int.Conf. on Knowledge Discovery and Data Mining, pp. 263-268, 2001.

[14] Li, G. Biswas; Unsupervised Learning with Mixed Numeric and Nominal Data: IEEE Transaction On Knowledge and Data Engineering, Vol. 14, no. 4, 2002.

[15] SushmitaMitra, Sankar K. Pal and PabitraMitra; Data Mining in Soft Computing Framework: A Survey: in IEEE Transactions on Neural Networks, Vol. 13, No. 1, January 2002.

[16] P. Pantel; Clustering by Committee: Ph.d. dissertation, Department of Computing Science, University of Alberta, 2003

Mahesh Singh, Assoc. Member IEEE AITM Students Branch, Branch Code: 09831, Asst. Professor CSE Department. Advanced Institute of Technology & Management, Palwal

Sangeeta Raheja, Student, M.Tech (CSE), Advanced Institute of Technology & Management, Palwal

Aashima**,** Student, M.Tech (CSE), Advanced Institute of Technology & Management, Palwal