Vandana Publications

Volume-4, Issue-1, February-2014, ISSN No.: 2250-0758

International Journal of Engineering and Management Research

Available at: www.ijemr.net

Page Number: 216-223

Performance Comparison of Various Partition based Clustering Algorithms

Sandeep[1], Priyanka[2], Renu Bansal[3]

[1]Teaching Associates, Guru Jambheshwar University of Science and Technology, Hisar, Haryana, INDIA

[3]Teaching Associates, Guru Jambheshwar University of Science and Technology, Hisar, Haryana, INDIA

ABSTRACT

In this paper a review of k-means, incremental k – means and D-M algorithm is presented. How the objects are clustered based on the three partitioning algorithms is shown. The complexities are calculated and compared. Implementations of the algorithms are clearly presented with their objects grouping.

I. IMPLEMENTATION VIEW OF D-M CLUSTERING

An example is taken to understand D-M clustering algorithm. In table below there is sample of four medicines having two attributes weight and pH. Let these medicines are given name A, B, C, D as data objects and its attributes weight and pH as X, Y. Here example is taken only for four data object with two attributes although it is applicable to n data objects with n attributes. The goal is to group these objects into groups (clusters) based upon attributes.

Object	attribute 1 (X): weight index	attribute 2 (Y): pH
Medicine A	1	1
Medicine B	2	1
Medicine C	4	3
Medicine D	5	4

Table: Data Objects

1. Initialization

In dynamic means clustering algorithm, there is no information about how many clusters have to be formed. So value of k is taken one initially. Consider object/medicine A as the first cluster and value of k is initialized to 1, where k denotes the number of clusters.K1 is name of first cluster. In other words k_1 is first cluster matrix. It is represented by a matrix of order 1 × 2 having object/medicine A as its row with its two attributes X and Y as columns. There is only one element in $k_{1.}$ k_1= [A] i.e. k_1= [1 1]. C denotes centroid matrix which keeps information about each cluster's centroids. Each cluster has centroid. Till now only one cluster is formed so only one centroid. Moreover, There is only one object in that cluster therefore cluster's centroid is that object i.e. object A (1, 1).

$$C = \begin{array}{cc} X & Y \\ [1 & 1] \end{array}$$

Let the threshold limit (T_{th}) = 2.5 which is maximum distance allowed between a cluster's centroid and its objects.

2. Select next object: (until all data objects are examined)

Select next object i.e. medicine B (2, 1) and calculate the distance (m) between object B and centroid of each cluster. There is only one cluster k_1 thus only one centroid (1, 1). So using **Euclidean distance formula**, m= $((2-1)^2 + (1-1)^2)$ ½ = 1.Distance (m) is "1" which is less than threshold limit. Therefore object B is included in the same cluster in which object A is i.e. $k_{1.}$ Now k_1 cluster have two objects object A and object B. It looks like this

$$K = [[k_1]] \quad \text{i.e.} \quad K = \left[\begin{array}{c} \begin{array}{cc} X & Y \end{array} \\ \begin{bmatrix} 1 & 1 \\ 2 & 1 \end{bmatrix} \end{array}\right]$$

Centroid matrix also updates. Attributes X and Y is recalculated for object A (1, 1) and object B (2, 1). Attribute X is mean value of X coordinates of all objects in that cluster and attribute Y is mean value of Y coordinates of all objects in that cluster. The coordinates of centroid are

$$X = (1 + 2) / 2 = 3/2$$

$$Y = (1 + 1) / 2 = 1$$

3. Select next object

Select next object i.e. medicine C (4, 3) and measure the distance between object C and each centroid of the clusters. Till now, there is only one cluster k_1 thus only one centroid (3/2, 1). So using Euclidean distance formula, $m = ((4-3/2)^2 + (3-1)^2)^{½} = 3.20$.Distance between object C and centroid is "3.20" which is greater than threshold (T_{th}). So make a new cluster (k_2) and increase the value of k by one i.e. k=2.And K updates with a new row

$$K = \begin{bmatrix} [k_1] \\ [k_2] \end{bmatrix} \quad \text{i.e.} \quad K = \begin{bmatrix} \begin{bmatrix} 1 & 1 \\ 2 & 1 \end{bmatrix} \\ \begin{bmatrix} 4 & 3 \end{bmatrix} \end{bmatrix}$$

Centroid matrix also updates with new row. This new row is 2nd cluster's centroid. 2nd cluster have single object so its centroid is object C (4, 3).

$$C = \begin{array}{cc} X & Y \\ \end{array} \begin{bmatrix} 1.5 & 1 \\ 4 & 3 \end{bmatrix}$$

4. Select next object

Select next object i.e. medicine D (5, 4) and measure the distance between object D and each centroid of the clusters. There are two clusters k_1 and k_2. k_1 and k_2 have centroids (3/2, 1) and (4, 3) respectively. There is need to find distance between object D and centroid (3/2, 1) and distance between object D and centroid (4, 3). So using Euclidean distance formula

$$1.\ m = ((5 - 1.5)^2 + (4 - 1)^2)^{½} = 4.60 (\text{D and centroid}(1.5,1))$$

$$2.\ m = ((5 - 4)^2 + (4 - 3)^2)^{½} = 1.41 (\text{D and centroid } (4, 3))$$

In case 1st distance is greater than threshold limit and in case 2nd distance (m) is less than threshold (T_{th}). So object D is closer to object C. Now Cluster k_2 have two elements object D and object C.

$$K = \begin{bmatrix} [k_1] \\ [k_2] \end{bmatrix} \quad \text{i.e.} \quad K = \begin{bmatrix} \begin{bmatrix} 1 & 1 \\ 2 & 1 \end{bmatrix} \\ \begin{bmatrix} 4 & 3 \\ 5 & 4 \end{bmatrix} \end{bmatrix}$$

Now K is matrix with two clusters k_1 and k_2, k_1 have two objects object A and object B and k_2 having two objects object C and object D. There is no addition of any object to cluster k_1 so its centroid is as it is. Centroid of cluster k_2 changes to X = (4 + 5) / 2 = 4.5 andY = (3 + 4) / 2 = 3.5.Centroid matrix changes to

$$C = \begin{array}{cc} X & Y \\ \end{array} \begin{bmatrix} 1.5 & 1 \\ 4.5 & 3.5 \end{bmatrix}$$

Object	Feature 1 (X): weight index	Feature 2 (Y): pH	Group (result)
Medicine A	1	1	1
Medicine B	2	1	1
Medicine C	4	3	2
Medicine D	5	4	2

Table: Final grouping (D-M clustering)

There are two clusters, showing that medicine A & medicine B belongs to cluster1 & medicine C & medicine D belongs to cluster2.

II. IMPLEMENTATION VIEW OF K-MEANS

1. Initial value of centroids

In k-Means *k* is number of clusters to be formed. Let its value is two. Consider Object A and Object B as two clusters. As there is only one element in each cluster so let $c_1=(1,1)$ and $c_2=(2,1)$ denote the coordinate of the centroids of both clusters.

2. Objects-Centroids distance

Next step of this algorithm is to measure the distance between cluster centroid to each object. Using Euclidean distance, distance matrix at iteration 0 is:

$$D^0 = \begin{bmatrix} 0 & 1 & 3.61 & 5 \\ 1 & 0 & 2.83 & 4.24 \end{bmatrix} \begin{array}{l} c_1 = (1, 1)\ \text{group-1} \\ c_2 = (2, 1)\ \text{group-2} \end{array}$$

$$\begin{array}{cccc} A & B & C & D \end{array}$$

$$\begin{bmatrix} 1 & 2 & 4 & 5 \\ 1 & 1 & 3 & 4 \end{bmatrix} \begin{array}{l} X \\ Y \end{array}$$

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. For example, distance from object C = (4, 3) to the first centroid $c_1=(1,1)$ is $\sqrt{(4-1)^2 + (3-1)^2} = 3.61$, and its distance to the second centroid $c_2=(2,1)$ is $\sqrt{(4-2)^2 + (3-1)^2} = 2.83$, etc.

3. Objects clustering

Assign each object based on the minimum distance. Thus, object A is assigned to group 1, object B to group 2, object C to group 2 and object D to group 2. The element of Group matrix below is 1 if and only if the object is assigned to that group.

$$G^0 = \begin{array}{cc} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 \end{bmatrix} & \begin{array}{l} \text{group-1} \\ \text{group-2} \end{array} \\ \begin{array}{cccc} A & B & C & D \end{array} & \end{array}$$

***4.* Iteration-1, determine centroids**

Knowing the members of each group, now compute the new centroid of each group based on these new memberships. Group 1 only has one member thus the centroid remains in $c_1=(1,1)$. Group 2 now has three members, thus the centroid is the average coordinate among the three members:

$$C_2 = \left(\frac{2+4+5}{3} , \frac{1+3+4}{3} \right) = (11/3, 8/3)$$

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

$$D^1 = \begin{array}{cc} \begin{bmatrix} 0 & 1 & 3.61 & 5 \\ 3.14 & 2.36 & 0.47 & 1.89 \end{bmatrix} & \begin{array}{l} c_1 = (1, 1) \text{ group-1} \\ c_2 = (11/3, 8/3) \text{ group-2} \end{array} \\ \begin{array}{cccc} A & B & C & D \end{array} & \\ \begin{bmatrix} 1 & 2 & 4 & 5 \\ 1 & 1 & 3 & 4 \end{bmatrix} & \begin{array}{l} X \\ Y \end{array} \end{array}$$

5. Iteration-1, Objects clustering

Similar to step 3, assign each object based on the minimum distance. Based on the new distance matrix, move the object B to Group 1 while all the other objects remain in Group 2.The Group matrix is shown as

$$G^1 = \begin{array}{cc} \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix} & \begin{array}{l} \text{group-1} \\ \text{group-2} \end{array} \\ \begin{array}{cccc} A & B & C & D \end{array} & \end{array}$$

Iteration 2, determine centroids

Repeat step 4 to measure the new centroid's coordinate based on the clustering of previous iteration. Group1 and group 2 both has two members, thus the new centroid's are

$$C_1 = \left(\frac{1+2}{2} , \frac{1+1}{2} \right) = (1\tfrac{1}{2}, 1) \text{ and}$$

$$C_2 = \left(\frac{4+5}{2} , \frac{3+4}{3} \right) = (4\tfrac{1}{2}, 3\tfrac{1}{2})$$

Iteration-2, Objects-Centroids distances

Repeat step 2 again, there is new distance matrix at iteration 2 as

$$D^2 = \begin{array}{cc} \begin{bmatrix} 0.5 & 0.5 & 3.20 & 4.61 \\ 4.30 & 3.54 & 0.71 & 0.71 \end{bmatrix} & \begin{array}{l} c_1 = (1\tfrac{1}{2}, 1) \text{ group-1} \\ c_2 = (4\tfrac{1}{2}, 3\tfrac{1}{2}) \text{ group-2} \end{array} \\ \begin{array}{cccc} A & B & C & D \end{array} & \\ \begin{bmatrix} 1 & 2 & 4 & 5 \\ 1 & 1 & 3 & 4 \end{bmatrix} & \begin{array}{l} X \\ Y \end{array} \end{array}$$

Iteration-2, Objects clustering

Again, assign each object based on the minimum distance.

$$G^2 = \begin{array}{cc} \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix} & \begin{array}{l} \text{group-1} \\ \text{group-2} \end{array} \\ \begin{array}{cccc} A & B & C & D \end{array} & \end{array}$$

But Obtained result that $G^2=G^1$. Comparing the grouping of last iteration and this iteration reveals that the objects does not move group anymore. Thus, the computation of the k-Means clustering has reached its stability and no more iteration is needed. Here is the final grouping in table below. There are two clusters showing that medicine A & medicine B belongs to cluster1 & medicine C & medicine D belongs to cluster 2.

Object	Feature 1 (X): weight index	Feature 2 (Y): pH	Group (result)
Medicine A	1	1	1
Medicine B	2	1	1
Medicine C	4	3	2
Medicine D	5	4	2

Table: Final grouping (k-Means clustering)

III. IMPLEMENTATION VIEW OF INCREMENTAL K-MEANS

1. Value of adjacency matrix

In this algorithm there is no information about how many clusters have to be formed. So, value of k is taken one initially. Preprocessing is done to form an adjacency matrix of order $n \times n$ which stores the distance between each pair of data object, where n is the number of objects. Consider first data object i.e. medicine A as first cluster and value of k is initialized to 1.k_1 is name of first cluster. In other words k_1 is first cluster matrix. It is represented by a matrix of order 1 × 2 having object/medicine A as its row with its two attributes X and Y as columns. k_1 is shown below. Till now there is one cluster matrix so there is only one element in K.

$$k_1 = [A] \quad \text{i.e.} \quad k_1 = \begin{array}{cc} X & Y \\ \end{array} \\ k_1 = \begin{bmatrix} 1 & 1 \end{bmatrix}$$

Let Threshold limit (T_{th}) = 2.5 which is maximum distance allowed between two data object of same cluster.

2. Select next object (until all data objects are examined) select next object i.e. medicine B (2, 1) and measure the distance (m) between object B and each object in the clusters. As there is only one cluster k_1 having only one object i.e. object A. So there is need to find distance between object B and object A. Distance between object B and object A is "1" i.e. entry in 2nd row and 1st column is "1" which is taken according to Euclidean distance formula, i.e. m= $((2\text{-}1)^2 + (1\text{-}1)^2)$ ½ = 1Compare distance (m) with threshold limit which is 2.5. It is less than our threshold limit. Therefore object B is included in the same cluster in which object A reside i.e. k_1. Now k_1 cluster have one more object i.e. medicine B. Cluster k_1 is updated with a new order 2 × 2 having object A and object B as its element. Every new object in a cluster increases a row in the corresponding cluster matrix.

$$k_1 = \begin{bmatrix} A \\ B \end{bmatrix} \quad \text{i.e.} \quad k_1 = \begin{array}{c} \begin{array}{cc} X & Y \end{array} \\ \begin{bmatrix} 1 & 1 \\ 2 & 1 \end{bmatrix} \end{array}$$

Similarly matrix K changes its order automatically. It looks like this

$$K = \begin{bmatrix} [k_1] \end{bmatrix} \quad \text{i.e.} \quad K = \begin{array}{c} \begin{array}{cc} X & Y \end{array} \\ \begin{bmatrix} \begin{bmatrix} 1 & 1 \\ 2 & 1 \end{bmatrix} \end{bmatrix} \end{array}$$

3. Select next object

Select next object i.e. medicine C (4, 3) and measure the distance between object C and each object in the clusters. There is only one cluster k_1 with two objects object A and object B. So there is need to find distance between object C and object A and between object C and object B. By using adjacency matrix- 1. Distance between object C and object A is "3.61" and Distance between object C and object B is "2.83"So in both cases distance (m) is greater than threshold (T_{th}). Object C can't be element of 1st cluster. Make a new cluster (k2) with this object as its an element. Increase the value of k by one i.e.

k=2.k_2 is the name of second cluster. Its cluster matrix is represented by a matrix of order 1 × 2 having object/medicine C as its row with two attributes X and Y as its column. There is only one element in k_2. And K updates with new row i.e. [k_2] and looks like

$$K = \begin{bmatrix} [k_1] \\ [k_2] \end{bmatrix} \quad \text{i.e.} \quad K = \begin{array}{c} \begin{array}{cc} X & Y \end{array} \\ \begin{bmatrix} \begin{bmatrix} 1 & 1 \\ 2 & 1 \end{bmatrix} \\ \begin{bmatrix} 4 & 3 \end{bmatrix} \end{bmatrix} \end{array}$$

4. Select next object

Select next object i.e. D (5, 4) and measure the distance between object D and each object in the clusters. There are two clusters k_1 and k_2. k_1 have two objects, object A and object B, and k_2 having only one object, object C. So there is need to find distance between object D object A, between object D and object B and between object D and object C. By using adjacency matrix- Distance between object D and object A is "5" and Distance between object D and object B is "4.24" and Distance between D and C is "1.41".In case 1st, 2nd distance is greater than threshold limit and in 3rd case distance (m) is less than threshold (T_{th}). So object D is closer to object C i.e. object D is an element of 2nd cluster. Now Cluster k_2 have two elements object D and object C.And K updates and looks like

$$K = \begin{bmatrix} [k_1] \\ [k_2] \end{bmatrix} \quad \text{i.e.} \quad K = \begin{array}{c} \begin{array}{cc} X & Y \end{array} \\ \begin{bmatrix} \begin{bmatrix} 1 & 1 \\ 2 & 1 \end{bmatrix} \\ \begin{bmatrix} 4 & 3 \\ 5 & 4 \end{bmatrix} \end{bmatrix} \end{array}$$

K is matrix with two clusters k_1 and k_2 , k_1 have two objects A(1, 1) and object B(2, 1) and k_2 having two objects C(4, 3) and object D(5, 4).All the data objects have been selected one by one. So set K is final set of clusters. k_1 is 1st cluster and k_2 is 2nd cluster.A .There are two clusters i.e. 1st cluster as it is and 2nd cluster is updated with new object, object D.

Object	Feature 1 (X): weight index	Feature 2 (Y): pH	Group (result)
Medicine A	1	1	1
Medicine B	2	1	1
Medicine C	4	3	2
Medicine D	5	4	2

Table: Final grouping (Incremental K-Means clustering)

IV. K-MEANS COMPLEXITY

To calculate the running time of k-Means algorithm it is necessary to know the number of times each statement run and cost of running.

Running Time

Algorithm:

k-means *(D, K, C)*

sr.no.		*cost*	*time*
1	Repeat until (No change in centroid)	m1	$l>=1$
2	for i= 1 to *n* do	m2	${}_1\sum^{l}(n+1)$
3	Determine distance (m) between d_i and each centroid c_j of any k_j in K such that m is minimum. (1<=j<=k)	m3	${}_1\sum^{l}{}_1\sum^{n}(k+1)$
4.	Assign d_i to cluster k_j	m4	${}_1\sum^{l}{}_1\sum^{n} 1$
5	Calculate new mean (centroid) for each cluster kj (1<=j<=k)	m5	${}_1\sum^{l}(k+1)$.

Fig.: Complexity of K-means

Sometimes number of steps is not known so it has been assumed. For example let number of times first statement runs with cost m1 is *l (>=1)*. For each *l*, next statement, for *i=1, 2 . . . n,* where *n* is number of data objects, runs *n+1* times with cost m2. For each *l* and for each *n*, next statement runs *k+1* times where *k* is number of clusters with cost m3. 4th statement runs one time for each *l* and for each *n* with cost m4. Calculating new mean for each cluster requires *k+1* runs for each *l* with cost m5 as shown.

Running time of algorithm is the sum of running time for each statement executed i.e.

$$T(n) = m1*l + m2*{}_1\sum^{l}(n+1) + m3*{}_1\sum^{l}{}_1\sum^{n}(k+1) + m4*{}_1\sum^{l}{}_1\sum^{n}1 + m5*{}_1\sum^{l}(k+1).$$

$$= m1*l + m2*l*(n+1) + m3*l*n*(k+1) + m4*l*n*1 + m5*l*(k+1)$$

$$= m1*l + m2*l*n + m2*l + m3*l*n*k + m3*l*n + m4*l*n + m5*l*k + m5*l$$

$$= (m1 + m2 + m5)*l + (m2 + m3 + m4)*l*n + m3*l*n*k$$

*For **worst case** it will be $O(n^i)$ where $2<=i<3$*

*For **best case** it will be $O(n)$*

*For **average case** it will be $O(n^2)$*

V. INCREMENTAL K-MEANS

In incremental k-means, number of times each statement runs is known. 1st, 2nd, 3rd, and 4th statement runs one time only with cost m1, m2, m3, m4 respectively. Next statement, for *i= 2, 3 . . . n* where *n* is number of data objects, runs *n* times with cost m5. 6th statement, for each *n*, scans each object in each cluster with cost m6. To understand running time for this statement let there are *k* clusters and in each cluster there are *s* objects. So running time of this statement, for each *n* and for each *k* is *s*+1. 7th statement runs *n*-1 times. Rest of statements is part of if-then-else body. Let if-then part body run for *r* times with cost m8 and then else part runs for *n*-1-*r* times with cost m9, m10, m11 as shown in figure.

Running Time

Algorithm:

Incremental k-Means (D, A)	*cost*	*time*
1. k=1	m1	1
2. k_k={ d_k}	m2	1
3. K= { k_k }	m3	1
4. Assign some constant value to T_{th}	m4	1
5. for i= 2 to *n* do	m5	n
6. Determine d_m in some cluster k_j in K such that distance (m) between d_m and d_i is minimum. (1<=j<=k)	m6	${}_{i=2}\sum^{n}{}_{j=1}\sum^{k}(s+1)$
7. if (m<= T_{th}) then	m7	n-1
8. k_j= k_j U d_i	m8	r
9. else k= k+1	m9	n-1-r
10. k_k= d_i	m10	n-1-r
11. K= K U k_k	m11	n-1-r

Figure: Complexity of Incremental K-means

$$T(n) = m1*1 + m2*1 + m3*1 + m4*1 + m5*n + m6*{}_{i=2}\sum^{n}{}_{j=1}\sum^{k}(s+1) + m7*(n-1)$$
$$+ m8*r + m9*(n-1-r) + m10*(n-1-r) + m11*(n-1-r)$$

$$= m1 + m2 + m3 + m4 + (m5 + m7 + m10 + m11)*n - (m7 + m10 + m11) + (m8 - m9 - m10 - m11)*r + m6*{}_{i=2}\sum^{n}{}_{j=1}\sum^{k}(s+1).$$

*For **worst case** it will be $O(nks)$, **best case** $O(nks)$ and **average case** $O(nks)$*

VI. D-M CLUSTERING ALGORITHM

In D-M clustering algorithm, like incremental k-Means, number of times each statement runs is known. 1st, 2nd, 3rd, 4th and 5th statement runs one time only with cost m1, m2, m3, m4 and m5 respectively. Next statement, for *i= 2, 3 . . . n* where *n* is number of data objects, runs *n* times with cost m6. 7th statement, for each *n*, scans centroid of each cluster with cost m7. So it runs *k*+1 times where k is number of clusters. 8th statement runs *n*-1 times with cost m8. Rest of statements is part of if-then-else body. Let if-then part body runs for *r* times with cost m9, m10 and then else part body runs for *n*-1-*r* times with cost m11, m12, m13, m14 as shown.

Running Time

Algorithm:	cost	time
D-M clustering algorithm *(D)*	*cost*	*time*
1. let k=1	m1	1
2 $k_k=\{ d_k\}$	m2	1
3. $K= \{ k_k \}$	m3	1
4. $C_k=d_k$	m4	1
5. Assign some constant value to T_{th}	m5	1
6. for i= 2 to *n* do	m6	n
7. Determine distance (m) between d_i and each centroid c_j of any k_j in K such that m is minimum. (1<=j<=k)	m7	$_{i=2}\sum^{n} (k+1)$
8. if ($m<= T_{th}$) then	m8	n-1
9. $k_j= k_j \cup d_i$	m9	r
10. Calculate new mean (Centroid c_j) for cluster k_j;	m10	r
11. else k= k+1	m11	n-1-r
12. $k_k= d_i$	m12	n-1-r
13. $K= K \cup k_k$	m13	n-1-r
14. $C_k= d_i$	m14	n-1-r

Complexity of D-M Clustering algorithm

$T(n) = m1 *1 + m2 *1 + m3 *1 + m4 * 1+ m5 * 1+ m6 *n + m7 * {}_{i=2}\sum^{n} (k+1) + m8 * (n-1) + m9 * r + m10 * r + m11 * (n-1-r) + m12 * (n-1-r) + m13 * (n-1-r) + m14 * (n-1-r)$ [10]

$=m1+ m2+ m3 +m4 + m5+ (m6 +m8 +m11+ m12 +m13+ m14) * n - (m8 +m11+ m12 + m13+ m14)+ (m9 + m10 - m11 - m12 - m13 - m14) * r+ m7 * {}_{i=2}\sum^{n} (k+1)$

[11]

*For **worst case** let p increases with increase in i then*

$${}_{i=2}\sum^{n} (k+1) =2+3....n$$
$$= n*(n+1)/2 -1$$

So $T(n)= m1+ m2+ m3 +m4 + m5+ (m6 +m8 +m11+ m12 +m13+ m14) * n - (m8 +m11+ m12 + m13+ m14)+ (m9 + m10 - m11 - m12 - m13 - m14) * r +m7* (n*(n+1)/2 -1)$ [12]

$T(n)=O(n^2)$

*For **best case** let p=1 for 2<=i<=n then* ${}_{i=2}\sum^{n} (k+1) = 2*n$

So $T(n)= m1+ m2+ m3 +m4+ m5+ (m6 +m8 +m11+ m12 +m13+ m14) * n - (m8 +m11+ m12 + m13+ m14)+ (m9 + m10 - m11 - m12 - m13 - m14) * r +m7*2*n$ [13]

$T(n)=O(n)$

*For **average case** it will be* $O(n^i)$ $1<= i<=2$

VII. COMPARISON TABLE

Name of algorithm	Worst case	Average case	Best case
k-means	$O(n^i)$ *where* $2<= i <3$	$O(n^2)$	$O(n)$
Incremental k-means	*O(nks)*	*O(nks)*	*O(nks)*
Dynamic clustering algorithm	$O(n^2)$	$O(n^i)$ *where* $1<=i<=2$	$O(n)$

Table: Comparison of algorithm's running time

REFERENCES

[1] J. Han and M. Kamber K. *Data Mining: Concepts and Techniques*. Morgan Kaufman, 2000.

[2] Margrett H. Dunham: *Data Mining: Introductory concepts and techniques*.

[3] P.S Bradley, Usama M. Fayyad., "*Initial Points for K-Means Clustering.*", Advances in Khowledge Discovery and Data Mining. MIT Press.

[4] Ramkrishnan and Gehrke: *Database Management Systems, 2nd ed*.

[5] Keim D. A.: "*Databases and Visualization*", Proc. Tutorial ACM SIGMOD Int. Conf. on management of Data, Montreal, Canada, 1996, p. 543.

[6] META Group Application Development Strategies: "*Data Mining for Data Warehouses: Uncovering* from Ralf Hart Hartmut Guting's VLDB Journal v3, n4, October 1994. *Hidden Patterns*." 7/13/95.

[7] "*Introduction to Data Mining and Knowledge Discovery" 3rd* Edition by Two Crows Corporation.

[8] Teknomo, kardi. "*K-Means clustering tutorial.*" http://www.people.revoledu.com/kardi/tutorial/kmeans.index.html

[9] Anderberg, M.R. "*Cluster analysis for applications*", Academic Press, New York, 1973, pp. 162-163

Costa, Luciano da Fontoura and Cesar, R.M., *Shape Analysis and Classification, Theory and Practice*, CRC Press, Boca Raton, 2001, pp 577-615.

[10] "*AUTOCLUST: Automatic clustering via boundary extraction for mining massive point-data sets*": Vladimir Estivill-Castro and Ickjai Lee Department of Computer Science & Software Engineering, The University of Newcastle, Callaghan, NSW 2308, Australia.

[11] Osmar R. Zaïane: "*Principles of Knowledge Discovery in Databases* - Chapter 8: Data-Clustering" http://www.cs.ualberta.ca/~zaiane/courses/cmput690/slides/Chapter8/index.html.

[12] TariqRashid:"Clustering" http://www.cs.bris.ac.uk/home/tr1690/documentation/fuzzy_clustering_initial_report/node11.html

[14] "*Introduction to Data Mining and Knowledge Discovery" 3rd* Edition by Two

[15] Shi Na, Liu Xumin Guan Yong.Research on k-means Clustering Algorithm: An Improved k-means Clustering Algorithm ,2010

[16] Gerhard M¨unz, Sa Li, Georg Carle Traffic Anomaly Detection Using K-Means Clustering. Computer Networks and Internet Wilhelm Schickard Institute for Computer Science University of Tuebingen, Germany,2005.

[17] Tapas Kanungo, Senior Member, IEEE, David M. Mount, Member, IEEE,Nathan S. Netanyahu, Member, IEEE, Christine D. Piatko, Ruth Silverman, andAngela Y. Wu, Senior Member, IEEE. An Efficient k-Means Clustering Algorithm:Analysis and Implementation, IEEE Transaction On Pattern Analysis And Machine Intelligence, VOL. 24, NO. 7, JULY 2002

[18] D T Pham, S. S Dimov, and C D Nguyen Selection of K in K-means clustering. Manufacturing Engineering Centre, Cardiff University, Cardiff, UK, 2004. DOI: 10.1243/095440605X829

[19] Anil K. Jain. Data Clustering: 50 Years Beyond K-Means.Department of Computer Science & Engineering

,Michigan State University East Lansing, Michigan 48824 USA,2002.

[20] Bryan Orme and Rich Johnson. Improving K-Means Cluster Analysis: Ensemble Analysis Instead of Highest Reproducibility Replicates, Saw tooth Software, 2008.

[21] Malay K. Pakhira A Modified *k*-means Algorithm to Avoid Empty Clusters. Kalyani Government Engineering College, Kalyani, West Bengal, INDIA
Email: malay_pakhira@yahoo.com .International Journal of Recent Trends in Engineering, Vol 1, No. 1, May 2009.

[22] Adil M.Bagirov, JulienUgon, DeanWebb .Fast modified global k-means algorithm for incremental cluster construction Centre for Informatics and Applied Optimization,Graduate School of Information Technology and Mathematical Sciences, University of Ballarat,Victoria 3353,Australia,2005.

[23] J. O. Omolehin, J. O. Oyelade, O. O. Ojeniyi and K. Rauf. Application of Fuzzy logic in decision making on students' academic performance.Bulletin of Pure and Applied Sciences, vol. 24E(2), pp. 281-187, 2005

[24] XindongWu , Vipin Kumar , J. Ross Quinlan , Joydeep Ghosh , Qiang Yang , Hiroshi Motoda , Geoffrey J. McLachlan , Philip S. Yu , Zhi-Hua Zhou , Michael Steinbach , David J. Hand , Dan Steinberg. Top 10 algorithms in data mining,Received: 9 July 2007 / Revised: 28 September 2007 / Accepted: 8 October 2007Published online: 4 December 2007.

[25] Zhexue Huang.CSIRO Mathematical and Information Sciences ,GPO Box 664, Canberra 2601, email:Zhexue.Huang@cmis.csiro.au,1996.

[26] C. Zahn. *Graph-theoretical methods for detecting and describing gestalt clusters*. In *IEEE Transactions on Computers*, pages 20:68–86, 1971.

[27] Clustering Algorithms at:
http://www.cs. uregina.ca/~hamilton/courses/831/notes/clustering/clustering.html.

[28] V. Brailovsky. *A probabilistic approach to clustering*. Pattern Recognition Letters, 12(4): 193–198, 1991.

[29] J. W. Shavlik and T. G. Dietterich. *Readings in Machine Learning*. Morgan Kaufmann, 1990.

[30] Hoppner, F., Klawonn **F.,** Kruse, R., and Runkler, T., *Fuzzy Cluster Analysis,* John **Wiley** and Sons,1999.

[31] Kantardzic, M. *Data Mining: Concepts, Models, Method, And Algorithms,* New
Jersey: IEEE Press, 2003.

[32] Steinbach, M., Karypis, G., Kumar, V., *A* Comparison *of Document Clustering Techniques,* University of Minnesota, Technical Report #00-034, 2000, at
http://www.cs.umn. edu/tech_reports/

[33] U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy. *Advances in Knowledge Discovery and Data Mining*. AAAI/MIT Press, 1996. [45]

[34] Sauravjyoti Sarmah and Dhruba K. Bhattacharyya. An Effective Technique for Clustering Incremental Gene Expression data. Department of Computer Science & Engg., Tezpur University Tezpur-784028, Assam, India,2008.

[35] T.Chandrasekhar, K.Thangavel and E.Elayaraja.An Introduction to Gene Expression Data, Research Scholar, Bharathiar university, Coimbatore, Tamilnadu, India - 641 046. IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 6, No 3, November 2011 ISSN (Online): 1694-0814,2011.

[36] Adil M. Bagirov Karim Mardaneh .Modified global k-means algorithm for clustering in gene expression data sets.Centre for Informatics and Applied Optimization,School of Information Technology and Mathematical Sciences,University of Ballarat, Victoria, 3353, Australia, 2003. Email: a.bagirov@ballarat.edu.au

[37] Sanjay Chakraborty, N. K. Nagwani . Analysis and Study of Incremental K-Means Clustering Algorithm,2011.High Performance Architecture and Grid Computing Communications in Computer and Information Science Volume 169, 2011, pp 338-341 .

[38] A.M.Sowjanya† and M.Shashi. Cluster Feature-Based Incremental Clustering Approach
(CFICA) For Numerical Data .Department of Computer Science. & Systems Engineering , College of Engineering ,Andhra University, Visakhapatnam.IJCSNS International Journal of Computer Science and Network Security, VOL.10 No.9, September 2010

[39] Taoying Li and Yan Chen Fuzzy *K*-Means Incremental Clustering Based on *K*-Center and Vector Quantization .Transportation Management College, Dalian Maritime University, Dalian 116026, P.R. China,2010.

[40] M.D. Boomija, M.C.A., M.Phil. Comparison Of Partition Based Clustering Algorithms,Lecturer, Department of MCA ,Prathyusha Institute of Technology and Management, Poonamallee -Tiruvallur High Road, Aranvoyalkuppam, Chennai – 602 025,2006.

[41] J.A. Lozano, P. Larra~naga. An empirical comparison of four initialization methods for the K-Means algorithm,Department of Computer Science and Artificial Intelligence, Intelligent Systems Group, University of the Basque Country, P.O. Box 649, E-20080,1999.

[42] Michael Steinbach, George Karypis, Vipin Kumar. A Comparison of Document Clustering Techniques,Department of Computer Science and Egineering, University of Minnesota Technical Report #00-034,1998.

[43] Dr. Latesh G. Malik, Pramod Prasad. Generating Customer Profiles for Retail Stores Using Clustering Techniques, Research Scholar Department of Computer Science and Engineering .G. H. Raisoni College of Engineering Nagpur, India,2011.Email: pmpramod@gmail.com

[44] J. Han : *OLAP Mining: An integration of OLAP with data mining*. Proc. of IFIP ICDS, 1997.

[45] Han J., Chee S., Chiang J. Y. : *Issues for On-Line Analytical Mining of Data Warehouses.* Proc. of 1998, SIGMOD'96 Workshop on Research Issues on DMKD'98 , 1998, pp. 2:1-2:5

[46] MacQueen, J.: "*Some Methods for Classification and Analysis of Multivariate Observations*", 5th Berkeley Symp. Math. Statist. Prob., Vol. 1, pp. 281-297.

[47] Press W. H.,Teukolsky S. A., Vetterling W. T., Flannery B. P.: "*Numerical Recipes in C*", 2nd ed., Cambridge University Press, 1992.

[48] L.Kaufmann and P.J. Rousseeuw. *Finding groups in data; an inroduction to cluster analysis*. Newyork;john wiley and sons, 1990

[49] Schikuta E., Erhart M.: "*The bang-clustering system: Grid-based data analysis*". Proc. Sec. Int. Symp. IDA-97, Vol. 1280 LNCS, London, UK, 1997.

[50] Huang, *extensions to the k-means algotithm for clusterig large data sets with categorical values.* Data mining and knowledge disvovery, 2:283-304,1998. [71]

[51] J.Gennari,P.Langley, and D.Fisher. *models of incremental concept formation*. Artificial intelligence, 40:11-61,1989. [72]

[52] E..Knorr and R.N.G *algorithms for mining distance-based outliers in large datasets*. In proc. 1998 int. conf. Very large data bases (vldb '98), pages 392-403, new York, aug.1998. [73

[53]http://www.newagepublishers.com/samplechapter/000896.pdf

[54] R. Agarwal, C.Faloutsos and A. Swami, *Efficient* [74] *similarity search in sequence databases*. In proceeding of international conference of foundation of data organization and algorithm. 1993. [75]

[55] http://www.newagepublishers.com/research methodology

[56] Matheus C.J.; Chan P.K.; and Piatetsky-Shapiro G. *Systems for Knowledge Discovery in Databases*, IEEE Transactions: 903-913.

[57] R.O. Duda, P.E. Hart and D.G.Stork , *Pattern Classification* (2nd edition) wiley New York 2001

[58] P. S. Heckbert. *Graphics Gems(4)*. Academic Press, 1994.

[59] L. Ertoz, M.Steinbach and V.Kumar. finding clusters of different size, shapes, and densities of noisy, high dimensional data. In proceeding of 3rd SIAM international conference on data mining, 2003.

[60] A. K. Jain, M. N. Murty, and P. J. Flynn. *Data clustering: a review*. ACM Computing Surveys, 31(3): 264–323, 1999.

[61] T. Kohonen. Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43:59–69, 1982.

[62] P.S.Bradley, U Fayyad, C. Reina, *initialization of iterative refinement clustering algorithm procedding* of the 4th ACM SIGKDD international conference on knowledge discovery and data mining. 1998

[63] García J.A., Fdez-Valdivia J., Cortijo F. J., and Molina R. 1994. *A Dynamic Approach for Clustering Data*. Signal Processing, Vol. 44, 1994, pp. 181-196.

[64] H. Ralambondrainy. *A conceptual version of the k-means algorithm. Pattern Recognition Letters*, 16(11):1147–1157, 1995.

[65] E. H. Ruspini. *A new approach to clustering*. Information and Control, 22–32.

[66] *K-means clustering algorithm data mining tutorial* started by KINGSLEY TAGBO at 12-14-2004.

[67] P. Sneath and R. Sokal. Numerical taxonomy, 1973.

[68] A. K. H. Tung, R. T. Ng, L. V. S. Lakshmanan, and J. Han. *Constraint-based clustering in large databases*. In ICDT, pages 405–419, 2001.

[69] F. Can, E. A. Fox, C. D. Snavely, and R. K. France. *Incremental clustering for very large document databases*: Initial MARIAN experience. Inf. Sci., 84:101–114, 1995.

[70] M. Charikar, C. Chekuri, T. Feder, and R. Motwani. *Incremental clustering and dynamic information retrieval*. In STOC, pages 626–635, 1997.

[71] A. Cornu´ jols. *Getting order independence in incrementale learning*. In European Conference on Machine Learning, pages 196–212, 1993.

[72] M. Ester, H. P. Kriegel, J. Sander, M. Wimmer, and X. Xu. *Incremental clustering for mining in a data warehousing environment.* In VLDB, pages 323–333, 1998.

[73] J. Lin, M. Vlachos, E. J. Keogh, and D. Gunopulos. *Iterative incremental clustering of time series*. In EDBT, pages 106–122, 2004.

[74] Cormen. H Thomas, Leiserson E. Charles, Rivest L.Ronald, Stein Clifford "*Introduction to algorithms- 2nd editiono*" Printice-Hall India.

[75] Fred Ana, *Finding Consisitent Clusters in Data Partition*,Institute de Telecomunicações, Institute Superior Técnico, Lisbon, Portugal.

.