Comparative Study Of Different Data Mining Techniques : A Review

Sudhir Singh
Deptt of Computer Science & Applications
M.D. University
Rohtak, Haryana
sudhirsingh24@yahoo.com

Nasib Singh Gill
Deptt of Computer Science & Applications
M.D. University
Rohtak, Haryana
nasib.gill@mdurohtak.ac.in

ABSTRACT
K-means and Incremental K-means are two very important and popular clustering techniques for today's large databases (Data warehouses, WWW and so on). The performance of the K-means and incremental K-means are different with each other based on their time analysis characteristics. Both algorithms are efficient compare to their existing algorithms with respect to time, cost and effort. In this paper, the performance evaluation of K-means clustering algorithm is implemented and most importantly it is compared with the performance of incremental K-means clustering algorithm and it also explains the characteristics of these two algorithms. This paper also explains some logical differences between these two most popular clustering algorithms. This paper uses a medicine database on which the experiment is performed.

KEYWORDS: Data mining, Knowledge management, K-means algorithm, Incremental K-means algorithm.

1. INTRODUCTION
Clustering is a method of grouping similar types of data. This is very useful method applied in various applications[1]. The K-means clustering and Incremental K-means clustering are the two most commonly used clustering techniques which grouped the data together based on different criteria. Incremental clustering is their extended version of K-means. Actual K-means suffers from several drawbacks, such as it needs predefined number of clusters and most importantly it does not has the capability to handle noisy data or outliers. Also it cannot form non-convex shapes clusters. Incremental K-means clustering is free from all these drawbacks and most importantly it does not require the number of cluster to be formed. In this paper K-means clustering and Incremental K-means clustering are applied on a medicine database and compare their performances.

2. CLUSTERING
Clustering can be considered the most important unsupervised learning problem, so, as every other problem of this kind, it deals with finding a structure in a collection of unlabeled data[7]. A loose definition of clustering could be "the process of organizing objects into groups whose members are similar in some way". A cluster is therefore a collection of objects which are "similar" between them and are "dissimilar" to the objects belonging to other clusters.
K-means clustering is a data mining algorithm used to cluster observations into groups of related observations without any prior knowledge of those relationships. The k-means algorithm is one of the simplest clustering techniques and it is commonly used in medical imaging, biometrics and related fields.

3. K-MEANS CLUSTERING
K-means algorithm is one of the partitioning based clustering algorithm. The general objective is to obtain the fixed number of partitions/clusters that minimize the sum of squared Euclidean distances between objects and cluster centroids.
Let $X=\{x_i|\ i=1,2,\ldots\ldots\ldots\ldots,n\}$ be a data set with n objects, k is the number of clusters, m_j is the centroid of cluster cj where $j=1,2,\ldots\ldots\ldots,k$. Then the algorithm finds the distance between a data object and a centroid by using the following Euclidean distance formula.
The Euclidean distance between two points/objects/items in a dataset, defined by point X and point Y is defined by Equation below[5].

$$\text{EUCLIDEAN DISTANCE}(X,Y) = (|X_1-Y_1|^2 + |X_2-Y_2|^{2+\ldots+} |X_{N-1}-Y_{N-1}|^{2+} |X_N-Y_N|^2)^{1/2}$$

OR

Euclidean distance formula=$\sqrt{\sum}|x_i-m_j|^2$ where **X** represents is the first data point, **Y** is the second data

point, **N** is the number of characteristics or attributes in data mining terminology.
Starting from an initial distribution of cluster centers in data space, each object is assigned to the cluster with closest center, after which each center itself is updated as the center of mass of all objects belonging to that particular cluster. The procedure is repeated until convergence.

3.1 K-MEANS ALGORITHM STEPS

The basic step of k-means clustering is to give the number of clusters k and consider first k objects from data set D as cluster & their centroid[4]. Then the k-means algorithm will do the three steps below until convergence.
Iterate until stable(=no object in the group matrix move):

1. Determine the centroid coordinate.
2. Determine the distance of each object to the centroids.
3. Group the object based on minimum distance.

3.2 FLOW CHART OF K-MEANS ALGORITHM

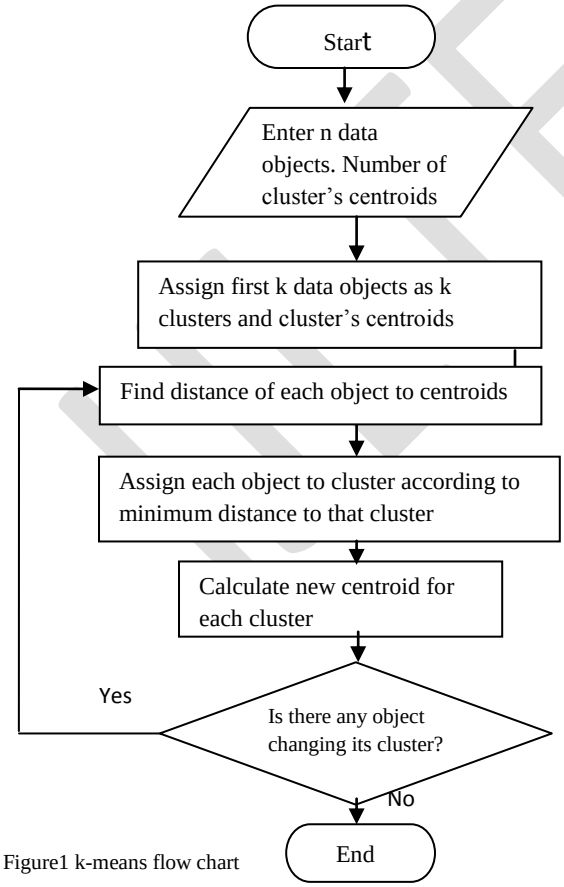

Figure1 k-means flow chart

3.3 K-MEANS Algorithm

k-means (D,K,C)

1. Repeat until(No change in centroid)
2. For i=1 to n do
3. Determine distance (m) between d_i and each centroid of any k_i in K such that m is minimum (1<=j<=k)
4. Assign d_i to cluster k_i.
5. Calculate new mean (centroid) for each cluster k_j. (1<=j<=k).

4. INCREMENTAL K-MEANS CLUSTERING

It improves the chances of finding the global optima with careful selection of initial cluster and mean. Number of group/cluster is unknown initially. Measuring minimum distance between two closest objects does the grouping. Objects are iteratively merged into the existing cluster or form a new cluster based upon the threshold limit. Thus the purpose of incremental k-means clustering is to classify the data on the basis of distance between the nearest object is less than threshold value.

4.2 FLOW CHART OF INCREMENTAL K-MEANS ALGORITHM

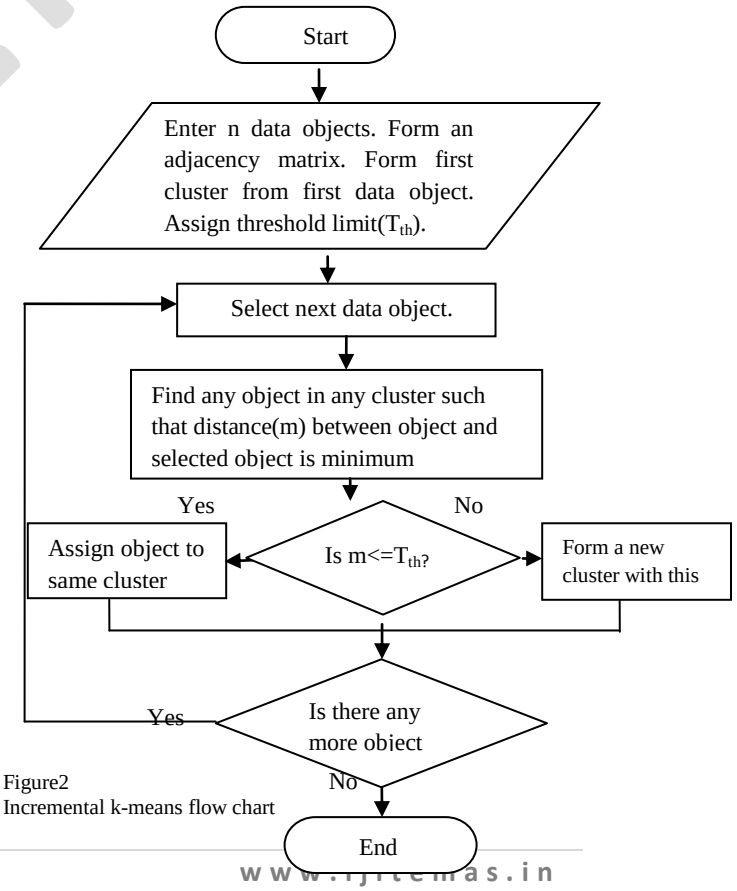

Figure2
Incremental k-means flow chart

5. EMPIRICAL ANALYSIS

To establish the practical efficiency of the k-means algorithm, We implemented and tested its performance on a medicine data set. In table 1 below there is sample of four medicines having two attributes weight and pH. Let these medicines given names A, B, C, D as data objects and its attributes weight and pH as X,Y.

Table 1 Data Object

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

5.1 IMPLEMENTATION THROUGH K-MEANS ALGORITHM

In the implementation k-means technique involve computation of centroids where these centroids will be used to cluster the data[4]. The K-means clustering, algorithm is applied on the medicine dataset and form clusters based on the nearest distance of the data from predefined centroids.

1st iteration

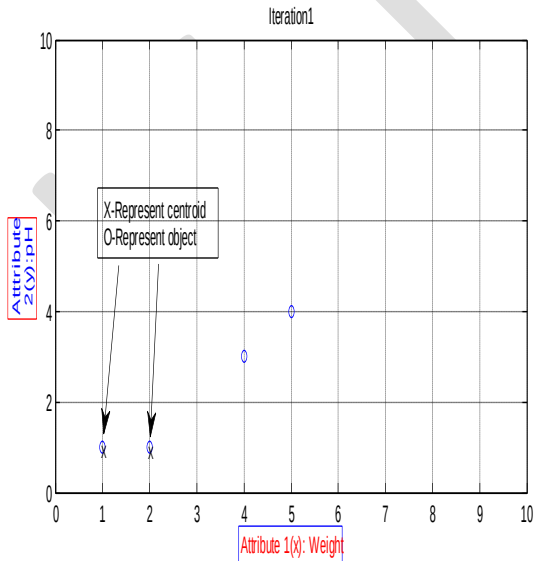

Figure3 Data points through first iteration

Distance Matrix using Euclidean distance is

$$D^0 = \begin{vmatrix} 0 & 1 & 3.61 & 5 \\ 1 & 0 & 2.83 & 4.24 \end{vmatrix} \quad \begin{matrix} C_1=(1,1)\ \text{group-1} \\ C_2=(2,1)\ \text{group-2} \end{matrix}$$

$$\begin{matrix} A & B & C & D \end{matrix}$$

$$\begin{vmatrix} 1 & 2 & 4 & 5 \\ 1 & 1 & 3 & 4 \end{vmatrix} \quad \begin{matrix} X \\ Y \end{matrix}$$

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 centriod 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.

On the basis of minimum distance we form group matrix

$$\begin{matrix} & A & B & C & D \end{matrix}$$

$$G^0 = \begin{vmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 \end{vmatrix} \quad \begin{matrix} \text{group-1} \\ \text{group-2} \end{matrix}$$

The element of group matrix below is 1 if and only if the object is assigned to that group.

2nd iteration

New centroid

$C_1=(1,1)$

C_2 =

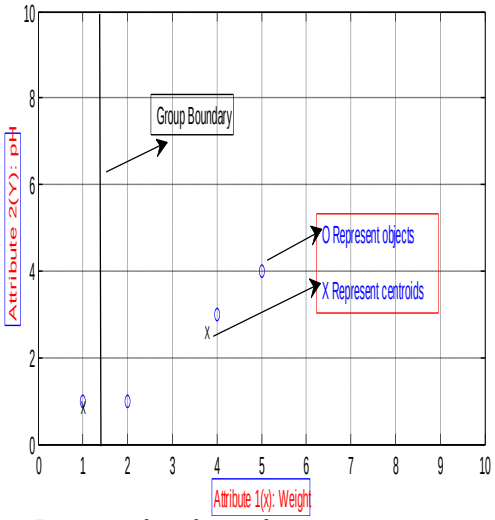

Figure4 Data points through second iteration

Distance Matrix using Euclidean distance is

$$D^1=\begin{vmatrix} 0 & 1 & 3.61 & 5 \\ 3.14 & 2.36 & 0.47 & 1.89 \end{vmatrix} \quad \begin{matrix} C_1=(1,1)\ \text{group-1} \\ C_2=(11/3,8/3)\text{group-2} \end{matrix}$$

Group Matrix

$$G^0=\begin{array}{c} \begin{matrix} A & B & C & D \end{matrix} \\ \left|\begin{matrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{matrix}\right| \end{array} \begin{matrix} \text{group-1} \\ \text{group-2} \end{matrix}$$

3rd iteration

New centroid

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

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

Distance Matrix using Euclidean distance is

$$D^1=\begin{vmatrix} 0.5 & 0.5 & 3.20 & 4.61 \\ 4.30 & 3.54 & 0.71 & 0.71 \end{vmatrix} \quad \begin{matrix} C_1=\left(1\tfrac{1}{2},1\right)\text{group-1} \\ C_2=\left(4\tfrac{1}{2},3\tfrac{1}{2}\right)\text{group-2} \end{matrix}$$

Group Matrix

$$G^0=\begin{array}{c} \begin{matrix} A & B & C & D \end{matrix} \\ \left|\begin{matrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{matrix}\right| \end{array} \begin{matrix} \text{group-1} \\ \text{group-2} \end{matrix}$$

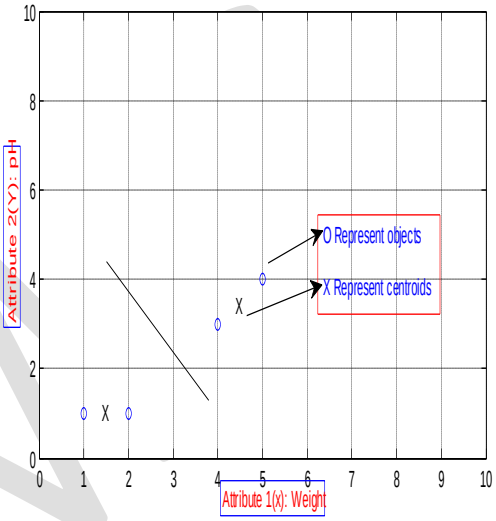

Figure5 Data point with new centroids using k-means

Table2 Final Grouping (using k-means clustering)

Object	Attribute 1(X): weight index	Attribute 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

5.2 IMPLEMENTATION THROUGH INCREMENTAL K-MEAN

Incremental k-means is based on Adjacency matrix which stores the distance between each pair of data object.

$$A[4][4]=\begin{array}{c} \begin{matrix} & A & B & C & D \end{matrix} \\ \begin{matrix} A \\ B \\ C \\ D \end{matrix} \left[\begin{matrix} 0 & 1 & 3.61 & 5 \\ 1 & 0 & 2.83 & 4.24 \\ 3.61 & 2.83 & 0 & 1.41 \\ 5 & 4.24 & 1.41 & 0 \end{matrix}\right] \end{array}$$

In this algorithm, initially we do not decide that how many cluster we have to be formed. Threshold limit (T_{th})=2.5 which is maximum distance allowed between two data object of same cluster.

First iteration.

Consider first data object i.e medicine A. K=1 where K denote the number of clusters.

Hence k_1=[A] i.e k_1=[1 1] (X Y)

Second iteration.

Consider second data object i.e medicine B(2,1). Distance between object B and A is "1", which is less than T_{th} , therefore object B is included in the same cluster in which object A is reside i.e k_1.

Hence $k_1=\begin{bmatrix} A \\ B \end{bmatrix}$ i.e $k_1=\begin{bmatrix} 1 & 1 \\ 2 & 1 \end{bmatrix}$ (X Y)

Third iteration.

Consider third data object i.e medicine C(4,3). Distance between object C and A is "3.61".
Distance between object C and B is "2.83". In both the cases distance (m) is greater than T_{th} , hence object C can't be element of 1st cluster.
Make a new cluster (k_2) with this object as an element.
Increase the value of K by one i.e K=2.

Hence $k_2=[C]$ i.e $k_2=[4\ 3]$ (X Y)

Fourth iteration.

Consider fourth data object i.e medicine D(5,4). Distance between object D and A is "5".
Distance between object D and B is "4.24".
and Distance between D and C is "1.41".
In 1st and 2nd distance is greater than threshold limit and in 3rd , distance is less than threshold T_{th}. So object D is closer to object C.
Cluster k_2 is updated with new object D as :

Hence $k_2=\begin{bmatrix} C \\ D \end{bmatrix}$ i.e $k_2=\begin{bmatrix} 4 & 3 \\ 5 & 4 \end{bmatrix}$ (X Y)

and K updates looks like

K=$\begin{bmatrix} [k_1] \\ [k_2] \end{bmatrix}$ i.e K=$\begin{bmatrix} \begin{bmatrix} 1 \\ 2 \end{bmatrix} \\ \begin{bmatrix} 4 \\ 5 \end{bmatrix} \end{bmatrix}$ (X Y)

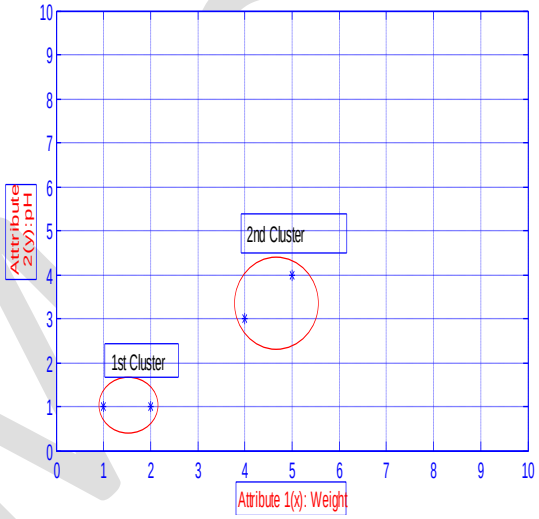

Figure6 Incremental k-means clustering

Table3 Final Grouping (Incremental k-means clustering)

Object	Attribute 1(X): weight index	Attribute 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

5.3 Performance Analysis

Time taken by an algorithm depends on the input, clustering a thousand data objects takes longer than clustering one object. Moreover, k-means and incremental k-means algorithms take different amount of time to cluster same data objects.

In this paper number of objects is represented by n, time taken by ith line is given by m_i , k is number of cluster and s is number of iteration.

Table4 Comparison of algorithm's running time

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	$O(nks)$	$O(nks)$	$O(nks)$

Comparative study of k-means and incremental K-means clustering algorithm is established. From the results it is observed that Incremental K-means Algorithm is efficient as compare to k-means algorithm.

6. CONCLUSION

In this paper we have reviewed k-means and incremental k-means clustering algorithm. we analyze the algorithm's running time not only as a function of k and n, but as a function of the degree to which the data set consists of well-separated clusters. Comparative study of k-means and incremental K-means clustering algorithm is established. From the results it is observed that Incremental K-means Algorithm is efficient as compare to k-means algorithm.

7. REFERENCES

[1] Han J. Kamber, Data Mining Concepts and Techniques, Morgan Kaufmann Publisher 2001.

[2] Chakraborty and Nagwani, S. and N.K.,2011, *Analysis and Study of Incremental K-Means Clustering Algorithm*, accepted by International conference on high performance architecture and grid computing (HPAGC).

[3] Dunham, M.H., 2003:Data Mining: Introductory And Advanced Topics, New Jersey: Prentice Hall,450.

[4] Data Mining, A Tutorial Based Primer, Richard J. Roiger and Michael W. Geatz : Pearson.

[5] Performance Evaluation of Incremental K-means Clustering Algorithm, IFRSA International Journal of Data Warehousing & Mining |Vol1|issue 1|Aug 2011

[6] M H Dunham, "Data Mining: Introductory and Advanced Topics," Prentice Hall, 2002.

[7] R C Dubes, A K Jain, "Algorithms for Clustering Data," Prentice Hall, 1988.

[8] B Zhang, M C Hsu, Umeshwar Dayal, "K-Means-A Data Clustering Algorithm,".

[9] Sanjay Chakraborty , N.K. Nagwani, IIJDWM Journal homepage: www.ifrsa.org

[10] Bharati M. Ramageri / Indian Journal of Computer Science and Engineering, Vol. 1 No. 4 301-305

[11] Khaled Alsabti,Syracuse University, An efficient K-Means Clustering Algorithm