Performance Comparison of Various Partition based Clustering Algorithms

Sandeep¹, Priyanka², Renu Bansal³

¹Teaching Associates, Guru Jambheshwar University of Science and Technology, Hisar, Haryana, INDIA
³Teaching 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.

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 k₁ 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 k₁. i.e. k₁= [A]

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).

Let the threshold limit (Tₘₙ) = 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 k₁ 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. k₁. Now k₁ cluster have two objects object A and object B. It looks like this

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

K= [k₁] i.e. K= 

X Y

1 1

2 1

Table: Data Objects

<table>
<thead>
<tr>
<th>Object</th>
<th>attribute 1 (X) weight index</th>
<th>attribute 2 (Y) pH</th>
</tr>
</thead>
<tbody>
<tr>
<td>Medicine A</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Medicine B</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Medicine C</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>Medicine D</td>
<td>5</td>
<td>4</td>
</tr>
</tbody>
</table>
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, \(m = \sqrt{(4-3/2)^2 + (3-1)^2} = 3.20\). 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

\[
\begin{bmatrix}
k_1 \\
k_2
\end{bmatrix}
\]

i.e.

\[
\begin{bmatrix}
1 & 1 \\
2 & 1 \\
4 & 3
\end{bmatrix}
\]

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).

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

\[
\begin{align*}
1. m &= ((5-1.5)^2 + (4-1)^2)^{1/2} = 4.60 \text{(D and centroid (1.5,1))} \\
2. m &= ((5-4)^2 + (4-3)^2)^{1/2} = 1.41 \text{(D and centroid (4,3))}
\end{align*}
\]

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.

\[
\begin{bmatrix}
k_1 \\
k_2
\end{bmatrix}
\]

i.e.

\[
\begin{bmatrix}
1 & 1 \\
2 & 1 \\
4 & 3
\end{bmatrix}
\]

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 \(X = (4 + 5) / 2 = 4.5\) and \(Y = (3 + 4) / 2 = 3.5\). Centroid matrix changes to

\[
\begin{bmatrix}
3/2 & 1 \\
4.5 & 3.5
\end{bmatrix}
\]

Table: Final grouping (D-M clustering)

<table>
<thead>
<tr>
<th>Object</th>
<th>Feature 1 (X)</th>
<th>Feature 2 (Y)</th>
<th>pH</th>
<th>Group (result)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Medicine A</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Medicine B</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Medicine C</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Medicine D</td>
<td>5</td>
<td>4</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

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 centroid 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:

\[
D_0 = \begin{bmatrix}
0 & 3.61 & 5 & c_1=(1,1) \text{ group-1} \\
1 & 2.83 & 4.24 & c_2=(2,1) \text{ group-2}
\end{bmatrix}
\]

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. 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.

$$G^0 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 \end{bmatrix}$$

A B C D

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

$$D^1 = \begin{bmatrix} 2.36 & 1.31 & 5 \\ 3.14 & 2.36 & 0.47 & 1.89 \end{bmatrix}$$

A B C D

$$= \begin{bmatrix} 1 & 2 & 4 & 5 \\ 1 & 1 & 3 & 4 \end{bmatrix}$$

### 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

$$G^1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \end{bmatrix}$$

A B C D

### Iteration 2, determine centroids

Repeat step 4 to measure the new centroid’s coordinate based on the clustering of previous iteration. Group 1 and group 2 both has two members, thus the new centroid’s are

$$c_1 = \left( \frac{1+2}{2}, \frac{1+1}{2} \right) = (1 1/2, 1)$$
$$c_2 = \left( \frac{4+5}{2}, \frac{3+4}{3} \right) = (4 1/2, 3 1/3)$$

### Iteration-2, Objects-Centroids distances

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

$$D^2 = \begin{bmatrix} 0.5 & 0.5 & 3.20 & 4.61 \\ 4.30 & 3.54 & 0.71 & 0.71 \end{bmatrix}$$

A B C D

$$= \begin{bmatrix} 1 & 2 & 4 & 5 \\ 1 & 1 & 3 & 4 \end{bmatrix}$$

### Iteration-2, Objects clustering

Again, assign each object based on the minimum distance.

$$G^2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \end{bmatrix}$$

A B C D

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 cluster 1 & medicine C & medicine D belongs to cluster 2.

<table>
<thead>
<tr>
<th>Object</th>
<th>Feature 1 (X): weight index</th>
<th>Feature 2 (Y): pH</th>
<th>Group (result)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Medicine A</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Medicine B</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Medicine C</td>
<td>4</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>Medicine D</td>
<td>5</td>
<td>4</td>
<td>2</td>
</tr>
</tbody>
</table>

### 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
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 2\(^{nd}\) row and 1\(^{st}\) column is “1” which is taken according to Euclidean distance formula, i.e. \( m = \sqrt{(2-1)^2 + (1-1)^2} = 1 \) Compare 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 \( \times \) 2 having object A and object B as its element. Every new object in a cluster increases a row in the corresponding cluster matrix.

\[
k_1 = \begin{bmatrix} A \\ B \end{bmatrix}
\]

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

\[
K = \begin{bmatrix} [k_1] \\ [k_2] \end{bmatrix}
\]

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 \( 1^{st} \) cluster. Make a new cluster \( (k_2) \) 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 \( \times \) 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

\[
K = \begin{bmatrix} X \\ Y \end{bmatrix}
\]

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 1\(^{st}\), 2\(^{nd}\) distance is greater than threshold limit and in 3\(^{rd}\) 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 2\(^{nd}\) cluster. Now Cluster \( k_2 \) have two elements object D and object C. And \( K \) updates and looks like

\[
K = \begin{bmatrix} X \\ Y \end{bmatrix}
\]

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 1\(^{st}\) cluster and \( k_2 \) is 2\(^{nd}\) cluster. There are two clusters i.e. 1\(^{st}\) cluster as it is and 2\(^{nd}\) cluster is updated with new object, object D.

<table>
<thead>
<tr>
<th>Object</th>
<th>Feature 1 (weight index)</th>
<th>Feature 2 (Y): pH</th>
<th>Group (result)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Medicine A</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Medicine B</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Medicine C</td>
<td>4</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>Medicine D</td>
<td>5</td>
<td>4</td>
<td>2</td>
</tr>
</tbody>
</table>

| 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.
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+l times with cost m2. For each l and for each n, next statement runs k+l 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+l 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.

\[ T(n) = m_1 + m_2 \cdot \sum_{l=1}^{k} (n+l) + m_3 \cdot \sum_{l=1}^{k} l + m_4 \cdot \sum_{l=1}^{k} l + m_5 \cdot \sum_{l=1}^{k} (k+l) + m_6 \cdot \sum_{l=1}^{k} n \]

For worst case it will be \( O(n^2) \) where \( 2 <= k <= n \)

For best case it will be \( O(n) \)

For average case it will be \( O(nk) \)

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. 7th statement, for each k, 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.

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 k, 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.
VII. COMPARISON TABLE

<table>
<thead>
<tr>
<th>Name of algorithm</th>
<th>Worst case</th>
<th>Average case</th>
<th>Best case</th>
</tr>
</thead>
<tbody>
<tr>
<td>k-means</td>
<td>O(n^2)</td>
<td>O(n^2)</td>
<td>O(n)</td>
</tr>
<tr>
<td>Incremental k-means</td>
<td>O(dknz)</td>
<td>O(dknz)</td>
<td>O(dknz)</td>
</tr>
<tr>
<td>Dynamic clustering algorithm</td>
<td>O(n^2)</td>
<td>O(n^2)</td>
<td>O(n^2)</td>
</tr>
</tbody>
</table>

Table: Comparison of algorithm’s running time

REFERENCES

[14] “Introduction to Data Mining and Knowledge Discovery” 3rd Edition by Two
[17] Tapas Kanungo, Senior Member, IEEE, David M. Mount, Member, IEEE, Nathan S. Netanyahu, Member, IEEE, Christine D. Piatko, Ruth Silverman, and Angela 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
[75] Fred Ana, Finding Consistent Clusters in Data Partition,Institute de Telecomunicações, Institute Superior Técnico, Lisbon, Portugal.