K-Mean Clustering Tutorial

By Kardi Teknomo

What is K-Mean Clustering?

Simply speaking it is an algorithm to cluster or to group your objects based on attributes/features into K number of group. K is positive integer number. The grouping is done by minimizing the sum of squares of distances between data and the corresponding cluster centroid. Thus the purpose of K-mean clustering is to classify the data.

Example: Suppose we have 4 objects and each object have 2 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

We also know before hand that these objects belong to two groups of medicine (cluster 1 and cluster 2). The problem now is to determine which medicines belong to cluster 1 and which medicines belong to the other cluster.

How the K-Mean Clustering algorithm works?

If the number of data is less than the number of cluster then we assign each data as the centroid of the cluster. Each centroid will have a cluster number. If the number of data is bigger than the number of cluster, for each data, we calculate the distance to all centroid and get the minimum distance. This data is said belong to the cluster that has minimum distance from this data.

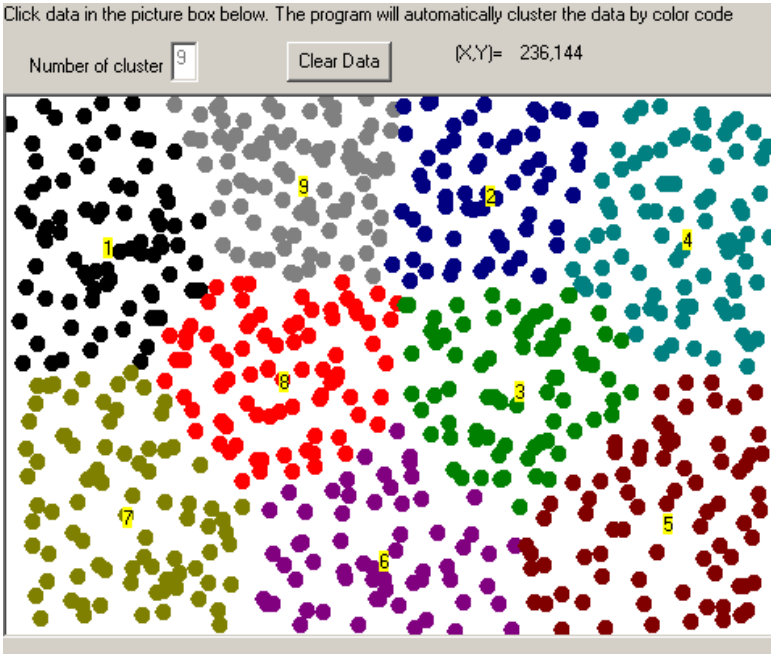
Since we are not sure about the location of the centroid, we need to adjust the centroid location based on the current updated data. Then we assign all the data to this new centroid. This process is repeated until no data is moving to another cluster anymore. Mathematically this loop can be proved to be convergent.

As an example, I have made a Visual Basic code. You may download the complete program in http://www.planetsourcecode.com/xq/ASP/txtCodeId.26983/lngWId.1/qx/vb/scripts/ShowCode.htm.
The number of features is limited to two only but you may extent it to any number of features. The main code is shown here.

```
Sub kMeanCluster (Data() As Variant, numCluster As Integer)
' main function to cluster data into k number of Clusters
' input: + Data matrix (0 to 2, 1 to TotalData); Row 0 = cluster, 1 =X, 2= Y; data in columns
'        + numCluster: number of cluster user want the data to be clustered
'        + private variables: Centroid, TotalData
' ouput: o) update centroid
'        o) assign cluster number to the Data (= row 0 of Data)
Dim i As Integer
Dim j As Integer
Dim X As Single
Dim Y As Single
Dim min As Single
Dim cluster As Integer
Dim d As Single
Dim sumXY()
Dim isStillMoving As Boolean
```

```
isStillMoving = True

If totalData <= numCluster Then
   Data(0, totalData) = totalData            ' cluster No = total data
   Centroid(1, totalData) = Data(1, totalData)  ' X
   Centroid(2, totalData) = Data(2, totalData)  ' Y
Else
   'calculate minimum distance to assign the new data
   min = 10 ^ 10                             'big number
   X = Data(1, totalData)
   Y = Data(2, totalData)
   For i = 1 To numCluster
      d = dist(X, Y, Centroid(1, i), Centroid(2, i))
      If d < min Then
         min = d
         cluster = i
      End If
   Next i
   Data(0, totalData) = cluster

   Do While isStillMoving
   ' this loop will surely convergent

      'calculate new centroids
      ReDim sumXY(1 To 3, 1 To numCluster)    ' 1 =X, 2=Y, 3=count number of data
      For i = 1 To totalData
         sumXY(1, Data(0, i)) = Data(1, i) + sumXY(1, Data(0, i))
         sumXY(2, Data(0, i)) = Data(2, i) + sumXY(2, Data(0, i))
         sumXY(3, Data(0, i)) = 1 + sumXY(3, Data(0, i))
      Next i
      For i = 1 To numCluster
         Centroid(1, i) = sumXY(1, i) / sumXY(3, i)
         Centroid(2, i) = sumXY(2, i) / sumXY(3, i)
      Next i


      'assign all data to the new centroids
      isStillMoving = False
      For i = 1 To totalData
         min = 10 ^ 10                         'big number
         X = Data(1, i)
         Y = Data(2, i)
         For j = 1 To numCluster
            d = dist(X, Y, Centroid(1, j), Centroid(2, j))
            If d < min Then
               min = d
               cluster = j
            End If
         Next j
         If Data(0, i) <> cluster Then
            Data(0, i) = cluster
            isStillMoving = True
         End If
      Next i
   Loop
End If
End Sub
```

The screen shot of the program is shown below.



When User click picture box to input new data (X, Y), the program will make group/cluster the data by minimizing the sum of squares of distances between data and the corresponding cluster centroid. Each dot is representing an object and the coordinate (X, Y) represents two attributes of the object. The colors of the dot and label number represent the cluster. You may try how the cluster may change when additional data is inputted.

What are the applications of K-mean clustering?

There are a lot of applications of the K-mean clustering, range from unsupervised learning of neural network, Pattern recognitions, Classification analysis, Artificial intelligent, image processing, machine vision, etc. In principle, you have several objects and each object have several attributes and you want to classify the objects based on the attributes, then you can apply this algorithm.

What are the weaknesses of K-Mean Clustering?

Similar to other algorithm, K-mean clustering has many weaknesses:

- When the numbers of data are not so many, initial grouping will determine the cluster significantly.
- The number of cluster, K, must be determined before hand.
- We never know the real cluster, using the same data, if it is inputted in a different way may produce different cluster if the number of data is a few.
- We never know which attribute contributes more to the grouping process since we assume that each attribute has the same weight.

One way to overcome those weaknesses is to use K-mean clustering only if there are available *many* data.

Are there any other resources for K-mean Clustering?

There are many books and journals or Internet resources discuss about K-mean clustering, your search must be depending on your application.