K. J. Somaiya College of Engineering, Mumbai-77

(Autonomous College Affiliated to University of Mumbai)

Batch: **A4** **Roll No.:** **1211061**

Experiment / assignment / tutorial No.1

Grade: AA / AB / BB / BC / CC / CD /DD

Signature of the Staff In-charge with date

Title: Implementation of K-means Clustering algorithm.

AIM: To understand the Partitional clustering algorithm K-means.

Expected Outcome of Experiment:

CO: Learn data mining techniques as well as methods in integrating and interpreting the data set.

Books/ Journals/ Websites referred:

Data Mining Concepts and Techniques-Third Edition

Pre Lab/ Prior Concepts:

k-means clustering is a method of vector quantization, originally from signal processing, that is popular for cluster analysis in data mining. *k*-means clustering aims to partition *n* observations into *k* clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster.

The algorithm has a loose relationship to the *k*-nearest neighbour classifier, a popular machine learning technique for classification that is often confused with *k*-means because of the *k* in the name. One can apply the 1-nearest neighbour classifier on the cluster centres obtained by *k*-means to classify new data into the existing clusters

K-means Clustering algorithm :

Input:

```
Enter the no. of objects
9
Enter the objects
2
4
10
3
12
20
30
11
25
Enter the no. of clusters
3
```

Output:

```
Iteration no.:1
2   4   10  3   12  20  30  11  25
Cluster :
1   2   3   1   3   3   3   3   3
Means:
2   4   18
Iteration no.:2
2   4   10  3   12  20  30  11  25
Cluster :
1   2   2   1   3   3   3   2   3
Means:
2   8   21
Iteration no.:3
2   4   10  3   12  20  30  11  25
Cluster :
1   1   2   1   2   3   3   2   3
Means:
3   11  25
Iteration no.:4
2   4   10  3   12  20  30  11  25
Cluster :
1   1   2   1   2   3   3   2   3
Means:
3   11  25
```

Method:
Three Methods:

1. Main :
   - Randomly initialize three means from given values
   - Calculate the distance of other points from the above selected means.
   - Process continues till new calculated means is equal to old means.
2. Findmin :
   - Provide the minimum distance of given points with three means.
3. Check:
   - Provide Boolean answer about the change in the calculated means with respect to old means

Conclusion:

Thus K-means algorithm was used to classify clusters dynamically based on the no. of clusters given by the user.

k-means clustering was implemented and executed for both one dimension and two dimension points.

Post Lab Questions

1. Give real life application for data mining functionalities like classification, clustering, association and correlation, characterization and discrimination.

Ans: The list of areas where data mining is widely used −

1) Classify credit approval based on customer data

2) Target marketing of product

3) Medical diagnosis based on symptoms of patient

4) Treatment effectiveness analysis of patient based on the treatment given

5) Marketing: Clustering can be used for targeted marketing

6) Biology: In classifying plants and animals into different classes based on their features.

7) Libraries: Based on different details about books clustering can be used for book ordering

8) Insurance: Clustering different groups of policy holders can be identified.

2. Illustrate the strength and weakness of K-means clustering approach.

Ans:

Strengths of K-Means Algorithm:

1) Relatively efficient: O(tkn), where n is # objects, k is # clusters, and t is # iterations. Normally, k, t << n.

   Comparing: PAM: $O(k(n-k)^2)$, CLARA: $O(ks^2 + k(n-k))$

2) If variables are huge, then K-Means most of the times computationally faster than hierarchical clustering, if we keep k smalls.

3) K-Means produce tighter clusters than hierarchical clustering, especially if the clusters are globular.

Weakness of K-means Algorithm:

1. Applicable only when mean is defined, then what about categorical data
2. Need to specify k, the number of clusters, in advance
3. Difficult to predict K-Value.
4. With global cluster, it didn't work well.
5. Different initial partitions can result in different final clusters.
6. It does not work well with clusters (in the original data) of Different size and Different density
7. Unable to handle noisy data and outliers
8. Not suitable to discover clusters with non-convex shapes

Date: _____________ **Signature of faculty in-charge**