Assignment No : 6

Roll no : 4351

1 Title

K Means Clustering

2 Problem Statement

Implement a simple approach for k-means/ k-medoids clustering using C++.

3 Learning Objectives

1. To understand Data Mining Concepts.
2. To develop problem solving abilities using Mathematical Modelling
3. To develop time and space efficient algorithms

4 Learning Outcome

1. After successfully completing this assignment, you should be able to Understand Implement efficient design, analysis and testing of algorithmic assignments.
2. Also we will learn a very effective clustering technique, which will help us in studying further techniques

5 Theory

In statistics and machine learning, k-means clustering is a method of cluster analysis which aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean.

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. This results in a partitioning of the data space into Voronoi cells.

The problem is computationally difficult (NP-hard); however, there are efficient heuristic algorithms that are commonly employed and converge quickly to a local optimum. These are usually similar to expectation-maximization algorithm for mixtures of Gaussian distributions via an iterative refinement approach employed by both algorithms. Additionally, they both use cluster centers to model the data; however, k-means clustering tends to find clusters of comparable spatial extent, while the expectation-maximization mechanism allows clusters to have different shapes.

The algorithm has a loose relationship to the k-nearest neighbor classifier, a popular machine learning technique for classification that is often confused with k-means because of the k in the name.

6 Mathematical Model

Let S be the solution perspective of the system such that
S={ s, e , X, Y, Fme , DD, NDD ,Ffriend | ϕs }
where
DD = Deterministic Data
NDD = Non - Deterministic Data = User Input

ϕs = Constraints on System S

s is start state

e is end state

X = x1, x2, x3, ..., xn — 'X' n instances of data.

A = a1, a2, a3, ..., ap — A p attributes of each instance X.

Operations Performed : initializecluster(X, Y), assigncluster(X, Y), updatecluster(X, Y)

initializecluster(X)1 = 'C' randomly selecting k cluster centres.
C = m — m is the center of cluster C

assigncluster(X, C) = B'; calculate the nearest cluster center to each Xi.

Cj mj — center of cluster Cj

if min d(Xi , m1), d(Xi , m2), , d(Xi , mk) = d(Xi , mj) then Xi mj

Xj Xj,1 , Xj,2 , , Xj,p — p attributes of Xj

mj mj,1 , mj,2 , , mj,p — p attributes of mj

d(Xi , mj) = — Xi,1 mj,1 — + — Xi,2 mj,2 — + + — Xi,p mj,p —

updatecluster(C) = C'; calculate the new cluster center. = m — no of instances within cluster

Y=Output
Y= C1, C2, C3, CK —O outcome of operations performed by the system k clusters where, Y1=Name
Y2 = Email
Y3 = Phone no

Success Case = Partition Successful

Failure Case = Partition Unsuccessful

7 Conclusion

Thus, after successfully completing this assignment, we understood and implemented k - mean clustering technique to partition instance data into k clusters in C++.