Comparative analysis of Clustering Algorithms and an implementation of the most efficient of them all (K-Means Clustering)

ALGORITHM ANALYSIS AND DESIGN – CS262

Department of Applied Mathematics, DTU

TANISHKA SINGH
2K19/MC/129
tanishkasingh_2k19mc129@dtu.ac.in

SREYA MAJUMDER
2K19/MC/127
sreyamajumder_2k19mc127@dtu.ac.in

OVERVIEW

Clustering or cluster analysis is a Machine Learning technique that involves the grouping of query points. Given a set of query points, we can use a clustering algorithm to classify each query point into a specific class. In theory, query points that are in the same class should have similar properties and/or features, while query points in different classes should have highly dissimilar attributes and/or features. In simple words, the aim is to partition groups with similar attributes and allot them into clusters.

In Data Science, we can use clustering analysis to procure some valuable insights from our data by seeing what groups the data points fall into when we implement a clustering algorithm. With the advent of many data clustering algorithms and its extensive use in wide variety of applications, including image processing, computational biology, mobile communication, medicine and economics, has led to the popularity of this algorithms.

There are many types of clustering algorithms. Each algorithm offers a different approach to the challenge of discovering natural groups in data. A list of 10 of the different clustering algorithms is as follows:

- **Affinity Propagation**
- **Agglomerative Clustering**
- **BIRCH**
- **DBSCAN**
- **K-Means**
- **Mini-Batch K-Means**
- **Mean Shift**
- **OPTICS**
- **Spectral Clustering**
- **Mixture of Gaussians**

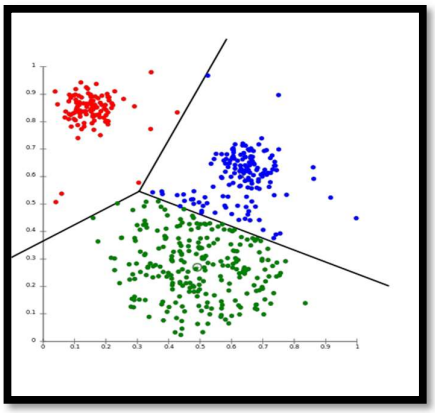
AIM

In this project we will be giving a comparative study of these clustering algorithm on the basis of their efficiency and time complexity. This analysis will further help us to conclude the best algorithm of them all, which theoretically is the K-Means clustering algorithm and we will get our analysis to assert this claim as well.
We will also be showing an implementation of the K-Means Clustering algorithm in a practical project which also makes use of the travelling salesman problem and helps us to find the shortest path using the Hamiltonian circuit.

K-MEANS CLUSTERING ALGORITHM

K-means clustering is one of the simplest and popular unsupervised machine-learning algorithms. Typically, such unsupervised algorithms make inferences from datasets using only input vectors without referring to known, or labelled, outcomes.

A cluster refers to a collection of query points clustered together because of certain similarities. We will define a target number *k*, which refers to the number of centroids you need in the dataset. A centroid is the imaginary or real location representing the centre of the cluster. Every query point is assigned to each of the clusters through reducing the in-cluster sum of squares.



In other words, the K-means algorithm identifies *k* number of centroids, and then assigns every data point to the nearest cluster, while keeping the centroids as small as possible. The *'means'* in the K-means refers to averaging of the data; that is, finding the centroid.

PRACTICAL IMPLEMENTATION

The practical implementation aims at recreating an emergency service optimization system, where we will simulate such a system with a few emergency services stations (police stations, NDFC stations). We will design a system which contains locations of these stations and also of citizens that are in need of emergency services during a natural disaster. Based on these locations it will help the emergency services stations to decide which station should cater to the needs of which citizens using K-Means Clustering Algorithm which will divide citizens into certain groups and help us carry out this task efficiently as without proper planning it might lead to loss of civilian life.

Even after every station is allocated citizens, they are responsible to efficiently extract them. Due to limited number of extraction vehicles, it is required that response team from each of the stations acts on extracting every citizen assigned to them in an efficient manner so that no time is wasted. This task will be carried out by applying the travelling salesman problem and finding the shortest Hamiltonian path.

SOFTWARE USED

- **Python**

REFERENCES

- **wikipedia.org**
- **geeksforgeeks.org**
- **tutorialspoint.com**
- **towardsdatascience.com**