भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

FINDING NEAREST HOSPITAL USING K-MEANS CLUSTERING

By
Potnuru Srilakshmi
(EP20BTECH11016)
Nitta Tamaswi Tania
(ME20BTECH11032)

Contents

1 Introduction

Finding the closest hospital based on location can be completed by grouping the available hospital data into categories based on their geographic coordinates using latitude and longitude information. K-means clustering, a well-liked unsupervised learning algorithm used in data science, is the clustering method used to group similar data points into clusters based on their similarity or proximity to each other.

The purpose of this project is to demonstrate the use of K-means clustering to find the nearest hospitals based on their latitude and longitude coordinates. By clustering the hospital data, we can easily identify the group of hospitals that are closest to a particular location, allowing for quick and efficient decision-making in emergency situations.

2 K-Means Clustering

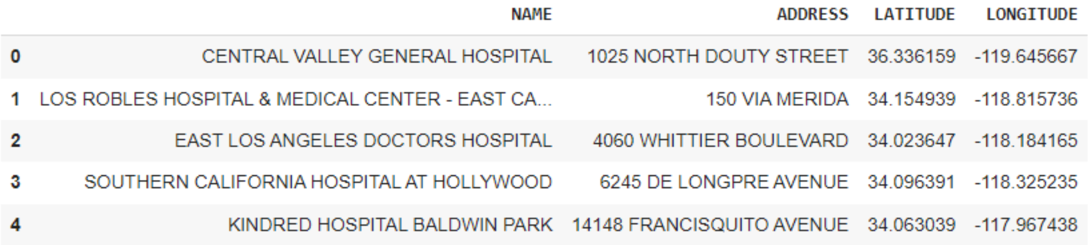
The current project proposes the utilization of the K-means clustering algorithm as an effective and straightforward unsupervised machine learning tool for grouping similar data points .

Cluster analysis could be divided into hierarchical clustering and non-hierarchical clustering techniques. Examples of hierarchical techniques are single linkage, complete linkage, average linkage, median, and Ward. Non-hierarchical techniques include k-means, adaptive k-means, k-medoids, and fuzzy clustering. A good clustering algorithm ideally should produce groups with distinct non-overlapping boundaries.

In this project, we implemented traditional kmeans clustering algorithm and Euclidean distance measure of similarity was chosen to be used in the analysis of hospital's data set.

3 Data Set

The information about the geographical coordinates of some Hospitals in USA is obtained from This kaggle dataset. The provided dataset consists of hospital information, including hospital name, address, and their respective latitude and longitude coordinates.



	NAME	ADDRESS	LATITUDE	LONGITUDE
0	CENTRAL VALLEY GENERAL HOSPITAL	1025 NORTH DOUTY STREET	36.336159	-119.645667
1	LOS ROBLES HOSPITAL & MEDICAL CENTER - EAST CA...	150 VIA MERIDA	34.154939	-118.815736
2	EAST LOS ANGELES DOCTORS HOSPITAL	4060 WHITTIER BOULEVARD	34.023647	-118.184165
3	SOUTHERN CALIFORNIA HOSPITAL AT HOLLYWOOD	6245 DE LONGPRE AVENUE	34.096391	-118.325235
4	KINDRED HOSPITAL BALDWIN PARK	14148 FRANCISQUITO AVENUE	34.063039	-117.967438

Figure 1: Sample dataset

4 Methodology

4.1 Development of k-mean clustering algorithm

Given a dataset of n data points x1, x2, . . . , xn such that each data point is in R^4,the problem of finding the minimum variance clustering of the dataset into k clusters is that of finding k points m_j (j=1, 2, . . . , k) in R^d such that

$$\frac{1}{n}\sum_{l=1}^{n}\left[min_j d^2(x_l, m_j)\right] \qquad (1)$$

is minimized, where d(x_i,m_j) denotes the Euclidean distance between xi and m_j.The points m_j (j=1, 2, . . . ,k) are known as cluster centroids. The problem in Eq.(1) is to find k cluster centroids, such that the average squared Euclidean distance (mean squared error, MSE) between a data point and its nearest cluster centroid is minimized.

An accessible approach to implement an approximation of the solution to Eq. (1) is the k-means algorithm. The popularity of k-means is due to its straightforward implementation, scalability, quick convergence, and capacity to handle sparse data

The k-means approach can be compared to a gradient descent procedure that iteratively updates the starting cluster centroids to reduce the objective function in Eq. (1). K-means consistently reach a local minimum. The beginning cluster centroids determine the specific local minimum that is discovered. Finding the global minimum is an NP-complete issue.

4.2 Traditional k-means algorithm

```
MSE = largenumber;
Select initial cluster centroids $\{m_j\}_j$ K = 1;
Do
    OldMSE = MSE;
    MSE1 = 0;
    For j = 1 to k
        $m_j = 0$; $n_j = 0$;
    endfor
    For i = 1 to n
        For j = 1 to k
            Compute squared Euclidean distance $d^2\ (x_i,\ m_j)$;
        endfor
        Find the closest centroid $m_j$ to $x_i$;
        $m_j = m_j + x_i$; $n_j = n_j + 1$;
        MSE1 = MSE1+ $d^2(x_i,\ m_j)$;
    endfor
    For j = 1 to k
        nj = max($n_j$, 1); $m_j = m_j/n_j$;
    endfor
    MSE=MSE1; while (MSE<OldMSE)
```

4.3 Psuedocode

- Step 1: Accept the number of clusters to group data into and the dataset to cluster as input values
- Step 2: Initialize the first K clusters
  - Take first k instances or
  - Take Random sampling of k elements
- Step 3: Calculate the arithmetic means of each cluster formed in the dataset.
- Step 4: K-means assigns each record in the dataset to only one of the initial clusters
  - Each record is assigned to the nearest cluster using a measure of distance (e.g Euclidean distance).
- Step 5: K-means re-assigns each record in the dataset to the most similar cluster and re-calculates the arithmetic mean of all the clusters in the dataset.

4.4 Source code

The source code of the project is avaliable here

5 Results

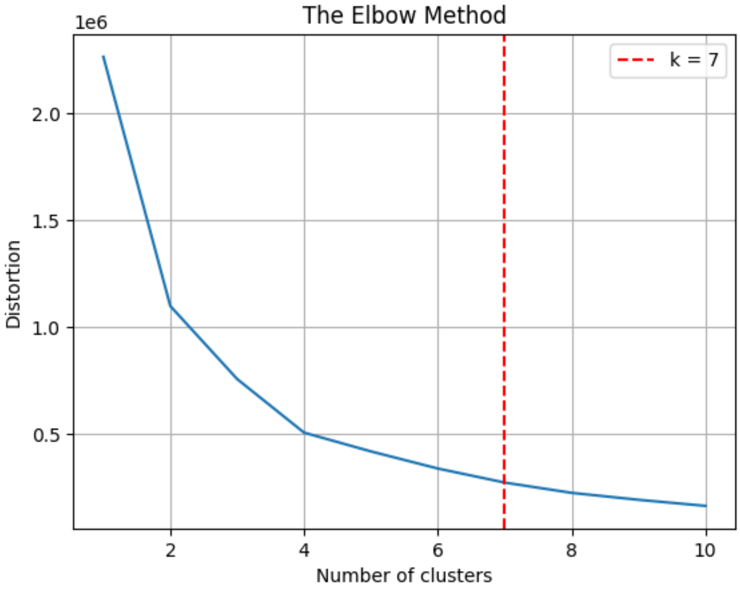


Figure 2: Elbow plot to get the optimal number of clusters.

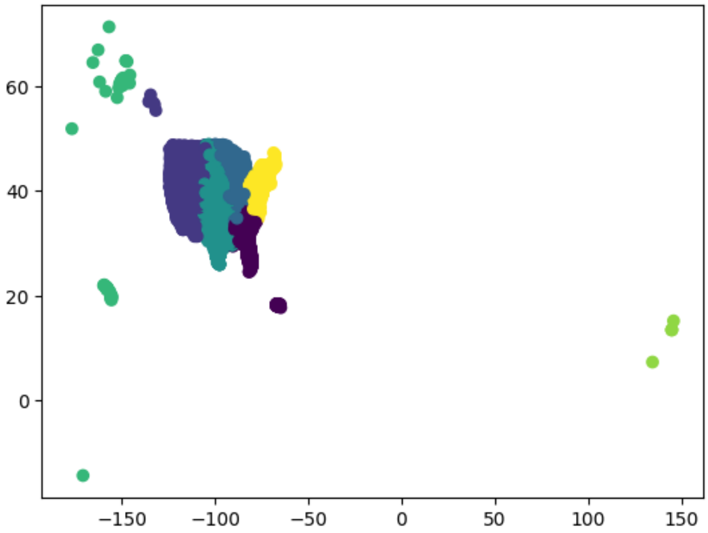


Figure 3: scatter plot of the data points with color code for the clusters.

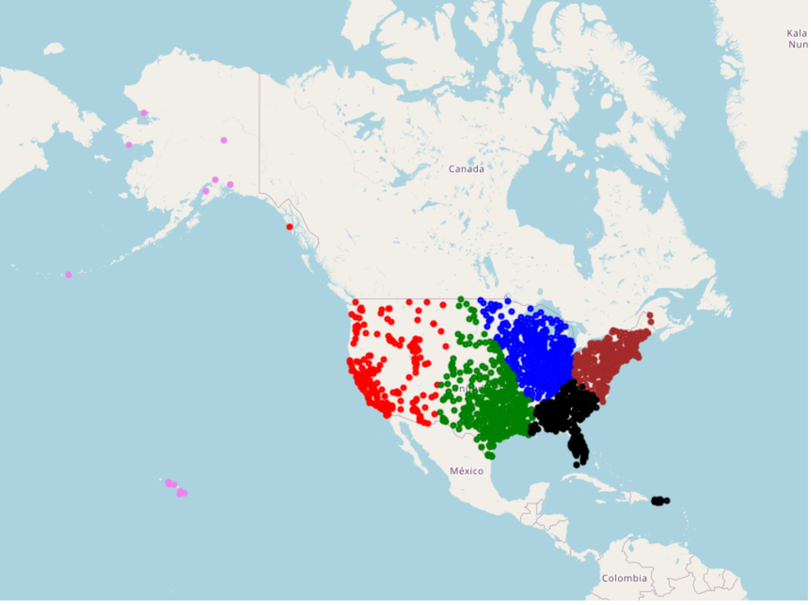


Figure 4: Geographical Visual of the hospital data using Folium module

6 Discussion and Conclusion

In this paper, we provided a simple and qualitative methodology to compare the predictive power of clustering algorithm and the Euclidean distance as a measure of similarity distance. We demonstrated our technique using kmeans clustering algorithm on a data set of Hospitals in USA with geographical cordinates.

In conclusion, K-means clustering is an excellent and efficient approach for emergency response services, medical experts, and patients seeking medical care to locate the closest hospitals based on latitude and longitude coordinates. This project has shown how the clustering technique can be used to solve a practical issue and how it can be implemented using the scikit-learn library and the Python programming language. The project's findings are encouraging, and the technique's use can be expanded to include industries other than medicine, such as marketing, logistics, and transportation.

It is crucial to highlight that this method has drawbacks, including the requirement for precise and current hospital location data, the need to choose an appropriate value for K, and the potential for biassed conclusions as a result of missing or biassed data. However, K-means clustering can offer a useful tool for decision-making and problem-solving in a variety of applications by carefully analysing the results and addressing these limitations.

7 References

- https://www.kaggle.com/datasets/andrewmvd/us-hospital-locations?resource=download
- arxiv.org/pdf/1002.2425
- https://github.com/iamtekson/geospatial-machine-learning