Grenze International Journal of Engineering and Technology, Jan Issue

Clustering of Crime Data using Haversine K-means Clustering Algorithm

Sajna Mol H S[1] and Gladston Raj S[2]
[1]Research Scholar, CDIT, University of Kerala, Trivandrum
Email:sajnahsalam@gmail.com
[2]Professor, Department of Computer Science, Govt. College, Kariavattom, Trivandrum
Email: gladston@rediffmail.com

***Abstract*— In this world, the rate of crimes is increasing as well as challenging the capabilities of people who are investigating crimes. Proper crime analysis and clustering has to be done in those cases. Crime analysis is the analysis of crime patterns and trends. It also assists in the research and planning necessary for the functioning of tactical forces and administrative services. Crime data grouping and clustering is very important to analyse the crime patterns and trends. By identifying patterns of crime committed in the past and the most common types of crime, crimes can be prevented from recurring in an area. Machine learning plays a key role in efficiently clustering today's crime data.**

***Index Terms*— BIRCH, CLARA, Haversine K-means, K-means, K-medoid.**

I. INTRODUCTION

There is a wide variety of criminal activity taking place around the world today. It is essential to examine the traits and patterns of these criminal behaviors and spot potential correlations between them. Criminal activity is naturally concentrated in a few locations illegally organized by criminals, who engage in the most illegal activities in order to make money as profit. In the current scenario, offenders are becoming more technologically advanced in criminal activities, and the challenge for police and intelligence departments is the difficulty of assessing large quantities of data regarding criminal and terrorist activities. Thus, authorities must be aware of techniques to apprehend criminals and gain the upper hand in the perpetual competition between evildoers and law enforcement [1]. In such cases, crime analysis plays an important role in an efficient manner.

The main purpose of crime analysis is to catch criminals. The next goal is to prevent or curtail crime. The third is to reduce social turmoil. The ultimate goal is to assist in the development and evaluation of organizational procedures.

Crime analysis techniques are the techniques used to handle the crime data set in various situations. One of the most important crime analysis techniques is clustering techniques using machine learning.

Machine learning is a technology that enables computers to make decisions without any human intervention. This machine learning-based crime analysis requires data collection, sorting, pattern identification, forecasting and visualization. Algorithms based on machine learning have demonstrated their utility and effectiveness on various types of datasets and can acquire and make predictions with great accuracy.

Grenze ID: 01.GIJET.10.1.328_1

II. PROPOSED METHODOLOGY

Clustering is an unsupervised task that doesn't require prior knowledge by discovering groups of correlated documents [2]. This paper deals with the crime data using Haversine k-means clustering by geo-location and it is compared with k-medoid clustering, BIRCH clustering, CLARA and k-means algorithms with result parameters as clustering time.

The work is implemented on PyCharm IDE using Python language. The data is collected from the Github website and the dataset used is the Sanfrancisco crime dataset during the year 2016 and it contains 150501 records. Each entry has 13 featured attributes and they are- IncidntNum, Category, Descript, DayOfWeek, Date, Time, PdDistrict, Resolution, Address, X, Y, Location, PdId.

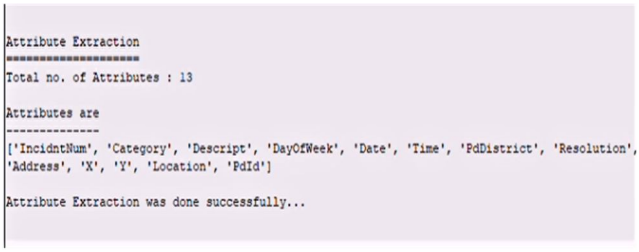

Figure 1. Attribute extraction

Then, the data is grouped based on district and category. On district based grouping, there are 10 districts in the dataset.

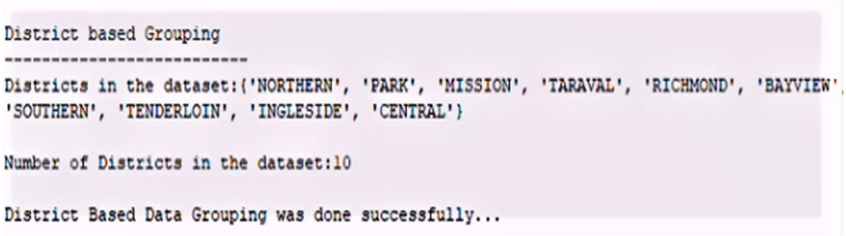

Figure 2. District based grouping

On category-based grouping, there are 32 categories in the dataset.

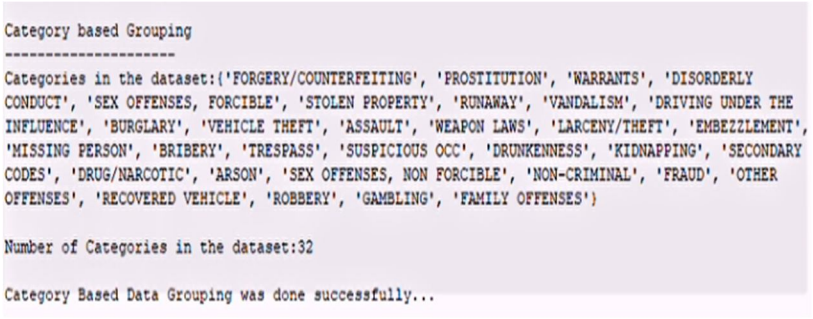

Figure 3. Category based grouping

After data grouping, clustering is performed using Haversine k-means clustering algorithm using geo-location.

A. K-Means Clustering Using Haversine Distance

K-means clustering is a technique for categorizing n observations into k different groups. Every observation belongs to the cluster with the nearest mean.

Longitude and latitude are used to calculate the distance between two objects. Haversine is an algorithm that computes the distance between two points on a sphere; the method is now refined with a straightforward formula. By utilizing a computer, this formula provides a very precise level of accuracy between the two objects [7].

Haversine distance is an appropriate distance metric on a spherical surface. This is the angular distance between two points on the sphere. The first coordinate of each point is taken as latitude and the second coordinate is taken as longitude, specified in radians. The data dimension must be 2.

$$D(x,y) = 2arcsin\left[\sqrt{sin^2\big((x1-y1)/2\big) + cos(x1)cos(y1)sin^2\big((x2-y2)/2\big)}\right] \quad (1)$$

The clustering time of this proposed algorithm is 36015ms. This algorithm is used to improve the clustering accuracy of the data by geo-location. Fig 4. shows the graph of clustering using Haversine k-means algorithm.

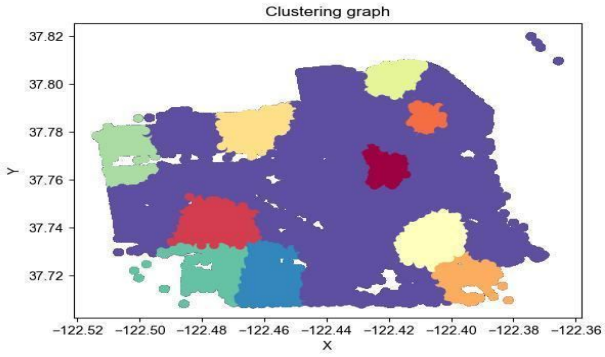

Figure 4. Clustering graph

B. Efficiency Checking Using Other Algorithms

a. K-medoid Algorithm

This is a classical cluster splitting method that divides a dataset of n objects into k clusters. Assume that it is known a priori. The medoid of a cluster is defined as the object in the cluster with the lowest average dissimilarity of all objects in the cluster, that is the most central point in the cluster. It is more robust to outliers and noise than k-means.

Algorithm:-

1. Initialization: select k random points from n data points as medoid.
2. Map each data point to its nearest medoid using one of the common distance metric measures.
3. If the cost decreases: For each medoid m, for each non-medoid data point o:

- Swap m and o, connect each data point to the nearest medoid, and recalculate the cost.
- Undo the swap, if the total cost is higher than the previous step.

K-medoid works by finding the midpoint of existing data without performing averaging like K-means [4].
The clustering time of this k- medoid algorithm on clustering the Sanfrancisco crime dataset is 60004 ms.

b. BIRCH Algorithm

Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH) is a method of decreasing big datasets by forming diminutive and compressed summaries of the large datasets that keep the most amount of information. This summary is then grouped instead of the original large dataset.
This method is based on the CF (clustering features) tree which is represented by three numbers:

- N: The number of elements in the subcluster.
- LS: The vector sum of data points.
- SS: The sum of squared data points.

The BIRCH algorithm has three parameters:

- *Threshold*- It is the maximum number of samples to be considered in the leaf nodes of the CF tree.
- *Branching factor*- It is the factor that shows how many CF sub- clusters a node can create.
- *N_clusters*- It is the number of clusters and is 8.

BIRCH's favorable position is its ability to cluster a growing dataset within a dataset, displaying multi-dimensional metric information and providing high-quality clustering for specific asset placements (memory and time imperatives) [5].
The clustering time of this BIRCH algorithm on clustering the Sanfrancisco crime dataset is 53003 ms.

c. CLARA Algorithm

CLARA (Clustering LARge Applications) is an extension of k-medoid. It takes only random samples of the input data (rather than the entire dataset) and computes the optimal medoid on those samples. Therefore, it performs better than k-medoid on crowded datasets.

Algorithm:-

- Randomly choose many fixed-sized collections from the original dataset.
- Calculate the k-medoid method for each subset and pick the related k symbol objects (medoids). Allocate each observation from the dataset to the nearest medoid.
- Determine the mean (or sum) of the dissimilarities of an observation to its closest medoid, which is used as a way to assess clustering quality.
- Save the subset with the least average (or total). Further assessment is done on the concluding partition.

In CLARA, medoids are used as their center points for a cluster, and the rest of the observations in a cluster are near to that center point [6].

The clustering time of this CLARA algorithm on clustering the Sanfrancisco crime dataset is 49006 ms.

d. K-Means Algorithm

K-means clustering is a type of unsupervised machine learning that segments datasets that are unlabeled into distinct clusters. In other words, it is a process that takes an unlabeled dataset and breaks it into k different clusters, with each dataset belonging to one group that has similar properties.

Algorithm:-

- Choose k clusters as the first step.
- Select a set of K instances as cluster centers.
- Each instance is assigned to the closest cluster by the algorithm.
- After all the instances are assigned a new cycle or after each instance is assigned, the cluster centroids are recalculated.
- This process is repeated [3].

The complexity of the k-means algorithm is O(tkn), which is relatively time-efficient since it involves n instances, c clusters, and t iterations. However, this method is prone to ending up with a local optimum. Since the mean must be predefined and c, the number of clusters must be specified in advance.

The clustering time of this k-means algorithm on clustering the Sanfrancisco crime dataset is 43001ms.

Fig 5. shows the comparison of Haversine K-means with other algorithms.

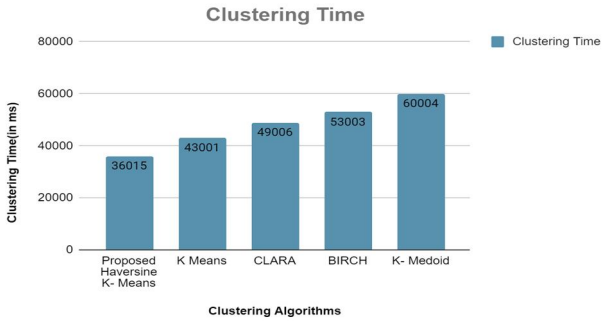

Figure 5. Comparison of algorithms

From the study, Haversine K- means algorithm is considered as a very convenient approach for making patterns and trends on the dataset as compared with other clustering algorithms such as k-medoids, BIRCH, CLARA and k-means algorithms.

The proposed work uses the Haversine -K-means algorithm to cluster the data by geo-location. Here, Haversine distance is used with the K-means algorithm to improve the clustering accuracy of the data.

III. CONCLUSION

In this paper, the crime data of Sanfrancisco is clustered using the Haversine K-means algorithm by geo-location and the clustering is done very efficiently with a lesser clustering time of 36015ms. It has been proven to be more

accurate compared to other algorithms such as k-medoid, BIRCH, CLARA and k-means. Efficient clustering using geo-location helps law enforcement investigators and authorities better understand and predict crime.

REFERENCES

[1] Khushabu A.Bokde, Tiksha P. Kakade, Dnyaneshwari S. Tumsare, Chetan G. Wadhai, Prof. Deepa Bhattacharya, "Crime Analysis Using K-Means Clustering" in International Journal of Engineering Research and Technology, Vol-7, Issue-4, ISSN:2278-0181, April 2018.

[2] Vineeth Jain, Yogesh Sharma, Ayush Bhatia, Vaibhav Arora, "Crime Prediction Using K-Means Algorithm" in Global Research and Development Journal for Engineering, Vol-2, Issue-5, ISSN:2455-5703, April 2017.

[3] Wasim A. Ali, Husam Alalloush, Manasa K. N, "Crime Analysis and Prediction using K-Means Clustering Technique" in EPRA International Journal of Research and Development, Vol-5, Issue-7, ISSN: 2455-7838, July 2020.

[4] Eduardus Hardika Sandy Atmaja, "Implementation of K-Medoids Clustering Algorithm to Cluster Crime Patterns in Yogyakarta" in International Journal of Applied Sciences and Smart Technologies, Vol-1, No-1, ISSN:2655-8564, 2019.

[5] Akshatha S R, Dr. Niharika Kumar, "Evaluation of BIRCH Clustering Algorithm for Big Data" in International Journal of Innovative Science and Research Technology, Vol-4, Issue-4, ISSN:2456- 2165, April 2019.

[6] Tanvi Gupta, Supriya P. Panda, "A Comparison of K-Means Clustering Algorithm and CLARA Clustering Algorithm on Iris Dataset" in International Journal of Engineering and Technology, Vol-7, Issue-4, 4766-4768, 2018.

[7] Dwi Arman Prasetya, Phong Thanh Nguyen, Rinat Faizullin, "Resolving the Shortest Path Problem using the Haversine Algorithm" in Journal of Critical Reviews, Vol-7, Issue-1, ISSN:2394-5125, 2020 .