# Literature Survey on Criminal Identification in Mumbai using DBSCAN

**Akshay Rathod*, Rushikesh Sawant, Ashish Choudhary, Neha Singh**

University of Mumbai, Atharva College of Engineering, Mumbai

**Publication Info**

***Article history:***
*Received : 14 February 2020*
*Accepted : 17 May 2020*

***Keywords:***
*Cluster, Crime analysis, DBSCAN, K-Means, Hotspots.*

***Corresponding author:***
*Akshay Rathod*
***e-mail:*** *rathod.akshay654321@gmail.com*

**Abstract**

*Crime rates are increasing every day in India, with Mumbai being the third among the 19 cities for 3 consecutive years; security against crime needs to be given increased priority by the government as well as individuals. In this paper, literature survey of crime analysis using DBSCAN clustering on crime dataset is done. The clustering helps clients in selecting a better, safer route from their current location to a desired location. The literature study focusses primarily on the algorithms used previously for similar systems, and comparing them with DBSCAN.*

## 1. INTRODUCTION

Criminals are a menace to the world. With all measures taken for controlling crime rates, crimes still happen very frequently in all parts of the world and we all are vulnerable to it. In 2020, India's crime index is 44.16, being among the top 60 countries affected most due to criminals. Our goal is to eradicate crime from our society. For the most part, the crime investigation start after a complaint has been filed, after the crime has already occurred. Such a system is good for fighting crime, but the key to decrease the crime rates would be crime avoidance. Our goal is to use technology and computer science to offer crime prevention. Clustering algorithms can help determine crime prone areas based on the history of criminal incidences. With such an application, anti-crime organizations will have strong knowledge about the crime prone areas, types of crime that may happen and the parties that might involve. With this knowledge, crimes can be expected to happen beforehand, and necessary precautions to avoid crime incidences can be taken. Clustering algorithm helps find clusters which tell where the possibility of crime happening is the most. Clusters with dense populations of crime incidences are classified as crime Hotspots, and the rest are considered as noise and are neglected. Crime datasets have a huge scope for data mining too. Hidden factors that might support criminals, such as lack of CCTV cameras in an area, absence of street lights, etc. may be highlighted due to information mining. Dealing with such factors will make executing crimes nearly impossible, thus promoting better safety. Such information will clearly define where the anti-crime agencies need to work upon. Many clustering algorithms are used for such applications in the past. One of the most common clustering algorithms is K-Means. Since K-Means faces problems with noisy data, this paper focuses upon an algorithm DBSCAN. DBSCAN is a clustering algorithm which is ideal for geo-spatial datasets such as crime data, which can handle noise efficiently.

## 2. LITERATURE STUDY

In the past, there have been many such systems, where crime data is analysed using different algorithms, mainly K-Means, K-Medoids, KNN etc. Some of the models and methodology are explained:

The authors, Jain et al. in their paper "Crime Prediction using K-Means Algorithm" [2], have used K-means clustering algorithm to find out patterns from the crime dataset. K-Means clustering algorithm is distance-based algorithm. The Euclidean distance metric is used to find the distance of a point from the nearest centres and decides if that point should belong to the cluster or not. The number of clusters cannot be determined at the start of the algorithm. Hence various iterations of K-Means have to be performed.

The authors have used Rapid Miner tool for analysis because of its flexibility and scalability. The main aim of the analysis was to understand which year was the crime rate highest and lowest. Supporting this information, bar graphs are plotted for each cluster.

The paper was published in 2013, after which, a lot of better results using better technologies were introduced. Crime dataset is an England based record of crimes from the year 1990 to 2012. The analysis is also based on, no



2010-2019 S-JPSET : Vol. 12, Issue Supplementary-1, ISSN : 2229-7111 (Print) and ISSN : 2454-5767 (Online) 

information about the actual location was used. Homicide was the prime target to tackle.

The authors, Tayal et al. In their paper "Crime detection and criminal identification in India using data mining techniques" [4] have used data mining using the WEKA® tool for applying K means on the crime dataset. After the cluster identification, Crime prediction is also done. For prediction, KNN classification is used. The classifier accuracy observed is 93.62 and 93.99 %. Google maps representation is done to show the clusters over a map. However, the representation is not useful enough, since only markers over the map give information about the number of crimes took place in a vast area around the marker, not giving much information about the exact location of crime.

The authors, Atmaja et al. in their paper "Implementation of k-Medoids Clustering Algorithm to Cluster Crime Patterns in Yogyakarta" [5], have implemented K-Medoids clustering algorithm for grouping the crime dataset of Yogyakarta, published in 2019. The crime data owned by the Yogyakarta Police is still stored in the manual form such as register books and excel. The data is only stored and is not used to produce any information. Where the data can be processed and analyzed to produce valuable information in efforts to prevent crime. Prediction of crime includes finding support and confidence for each group for extracting association rules from dataset. Probability of 'IF Theft THEN Embezzlement' is found which is then used for giving advice.

Research on the use of Euclidean distance in K-means algorithm has been successfully done. The aim of his study was to cluster crime data into three categories, namely high, medium and low crime level. Although the objective of the research was achieved, K-means algorithm is classified as an ineffective algorithm because it involves too much noise and outliers caused by the average selection of clusters [6].This study tried to improve previous study by replacing K-means algorithm with K-medoids algorithm. K-medoids algorithm is a clustering algorithms that can handle outliers or other extreme variables [6]. K-medoids work by determining the center point of existing data without performing an average calculation as in K-means.

The location where the probability of crime incidence happening is predicted to be the most is termed as crime hot spot. [8][7]. The authors, E. Eftelioglu et al., in their paper "Crime hotspot detection: A computa-tional perspective" [8], discuss the limitations of clustering for hotspot detection. They have used statistical approach for hotspot detection. SatScan software is used for hotspot detection over clustering. Circular and linear statistical scanning is done using SatScan.

The author, Sumanta Das et al. int their paper "A Geo-Statistical Approach for Crime hot spot Prediction" are predicting the hot spots of the next year based on the present ones. Clustering on an Indian crime dataset is done, with records of the major cities of India, in the year 2010 to 2014. Various clustering algorithms are compared on the basis of outputs they produced. Among Nearest neighbour hierarchical spatial clustering (NNHSC) and Kmeans, STAC(Spatial and Temporal Analysis of Crime) produced better output.

## 3. PROPOSED MODEL

DBSCAN is used which is a very accurate and application suited algorithm for identification of hotspots.

DBSCAN is Density-based spatial clustering of applications with Noise, which will find the crime hotspots based on the density of crime instances in an area, in other words, finding a set of points minimum 'x' points which are populated in an area at least 'y' square units. The points that do not follow the criteria are useless and are classified as 'noise'. DBSCAN classification is based two primary factors –

- The maximum distance between two samples for one to be considered as in the neighborhood of the other.
- The number of samples (or total weight) in a neighborhood for a point to be considered as a core point. This includes the point itself.

The core points define the cluster. If the amount of crime instances in a specific radius exceed a specific defined number, then the algorithm will classify the points. Since DBScan can easily detect outliers, noise can be eliminated. Major previous papers used K-means for the same purpose. KMeans is a distance based clustering technique. The clustering algorithm uses the distance to compute which data point will be the part of which cluster, or groups of other data points. The distance measures used by KMeans is Euclidean distance which is computed by the f ormula-

$$\sqrt{((x1-y1)^2 + (x2-y2)^2} \quad (1)$$

Disadvantages of k-means are

- Same sized clusters are attempted to find by the algorithm.
- Problems are faced dealing with non-globular structures.
- K-Means does not consider the density of data points, since it is a distance-based clustering approach.
- Unable to handle noisy data.
- K-Means is affected by curse of dimensionality, especially when dealing with such high dimensional data.

Over K-means, DBScan can easily find density connected regions. Different sizes of hotspots can be identified.


 2010-2019 S-JPSET : Vol. 12, Issue Supplementary-1, ISSN : 2229-7111 (Print) and ISSN : 2454-5767 (Online)

## 4. CONCLUSION

There are a lot of clustering algorithms and none are perfect. Each has its own advantages and disadvantages. No single algorithm can work for every application. DBSCAN is a very simple algorithm which gives great performance in most applications, and the results have very close resemblance with the human intuitional clustering.

The project focuses on analyzing crime data by implementing clustering algorithm DBSCAN on a crime dataset. We have done crime analysis and the results are plotted on the map, which will not only help us understand the crime trends, but also apply this knowledge directly for helping the users. Especially in India, where such systems have not been implemented, and orthodox practices such as file system being used, an intelligent system like the proposed one has a huge scope.

## 5. REFERENCES

[1] Jain, V., Sharma, Y., Bhatia, A.K., & Arora, V. (2017). Crime Prediction using K-means Algorithm.

[2] Agarwal, J., Nagpal, R., & Sehgal, R. (2013). Crime Analysis using K-Means Clustering.

[3] https://www.gov.uk/government/publications/offencesrecorded-by-the-police-in-england-and-wales-byoffence-and-police-force-area-1990-to-2011-12

[4] Tayal, D.K., Jain, A., Arora, S. et al. Crime detection and criminal identification in India using data mining techniques. AI & Soc 30, 117–127 (2015). https://doi.org/10.1007/s00146-014-0539-6

[5] Atmaja, Eduardus. (2019). Implementation of k-Medoids Clustering Algorithm to Cluster Crime Patterns in Yogyakarta. International Journal of Applied Sciences and Smart Technologies. 1. 33-44. 10.24071/ijasst.v1i1.1859.

[6] J. Han, "Data mining: concepts and techniques second edition," Morgan Kaufmann, San Francisco, 2006.

[7] Sumanta Das, Malini Roy Choudhury," A Geo-Statistical Approach for Crime hot spot Prediction", Department of Civil Engineering.,2016.

[8] E. Eftelioglu, S. Shekhar, and X. Tang, "Crime hotspot detection: A computa-tional perspective," in Data Mining Trends and Applications in Criminal Scienceand Investigations. IGI Global, 2016, pp. 82–111.

[9] Ester, Martin; Kriegel, Hans-Peter; Sander, Jörg; Xu, Xiaowei (1996). Simoudis, Evangelos; Han, Jiawei; Fayyad, Usama M. (eds.). A density-based algorithm for discovering clusters in large spatial databases with noise. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96).

 