Criminal Identification in Mumbai using DBSCAN

Akshay Rathod[1], Rushikesh Sawant[2], Ashish Choudhary[3] and Prof.Neha Singh[4]

[1,2,3]Student, University of Mumbai, Atharva College of Engineering, Mumbai

[4] Assistant Professor , University of Mumbai, Atharva College of Engineering, Mumbai

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, crime analysis is done by performing DBSCAN clustering on crime dataset, synthetically developed using make-blobs, to help clients in selecting a better, safer route from their current location to a desired location. The primary aim is to find out major patterns of criminal activities which will help the anti-criminal agencies to take actions beforehand.

Keywords: *Cluster, Crime analysis, DBSCAN, Leaflet, Hotspots.*

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. Existing Systems

Previous such systems mainly used algorithms such as K-Means and K-Medoids. These clustering algorithms depend upon the distance metrics and not the density of points in an area. The distance between the data point and the centre points is calculated with the distance measures. The clustering algorithm then uses this distance to compute which data point will be the part of which cluster, or groups of other data points. One of the distance measures is Euclidean distance which is computed by the formula-

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

K-Means is clustering algorithm that uses the Euclidean distance measure for clustering. A similar system developed in 2013 by the authors Jyoti Agarwal et al. in their paper “Crime analysis using K-Means” [2] had used K-Means algorithm. The system uses RapidMiner for updating the dataset. WEKA tool was used for clustering. Crime analysis is done on the resultant clusters.

Another paper from 2015, “Crime detection and criminal identification in India” [17] by authors Devendra Kumar Tayal et al., have used K-Means for clustering. The results include plots, number of crimes versus year, and clustering is done over the plot for analyzing the years with highest crime probability. GMAPI representation of clusters is shown.

Apart from clustering, neural networks and classification techniques have been used previously. Authors Eugenio Cesario et al. in the paper “Forecasting Crimes using Autoregressive Models” [18] have done analysis using neural networks and Naïve Bayes classifier in 2014. A comparison is done and the results show that Naïve Bayes performed better with an accuracy of 90.22%.

Only bar graphs and plots are presented in similar projects, which give trends in crime, which have not been implemented for developing a helpful system for users. Bar plots representing the number of crime instances in a city/state over a period of month/years were shown. Very little information of the exact geographical location is taken into consideration. Previous implementations provide raw implementations of clustering algorithms providing results, although very useful for the police jurisdiction, serve little use to the common individual who is concerned about the safety value of a particular street he has to travel through.

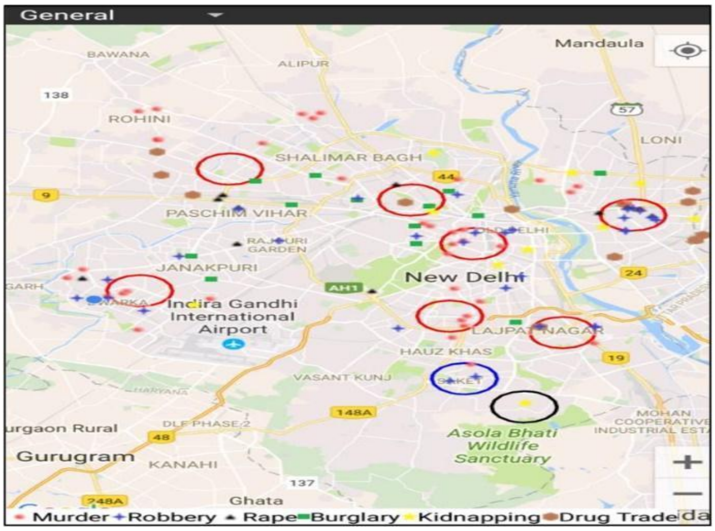


Figure 1. Existing systems sample [1]

2.1. Drawbacks

- Standard deviation from the mean is high, resulting in a scattered plotting which cannot be considered true to the data [4].
- Mapping of route from a starting point to end point is not done [12].
- Database updation is not present.
- Active feedback from the users is not present.
- Data mining techniques are obsolete [9]

3. Proposed System

User can interact with the system for the following:

3.1. Safe travel

The system can be used by any user who wants to travel safely. The user’s current location is monitored using the phone GPS. The user will then enter the destination. With the help of the Leaflet javascript Routing Machine API, the shortest route will be visible to the user. By dragging the route given over the map, the user can maneuver the route, avoiding any crime clusters on the map, according to the safety requirements. The UI will provide all functionalities related to directions of the route selected, safety value of the streets, enroute destinations etc

3.2. Feedback

After the safe completion of the journey, the user will be prompted to fill out a feedback form where additional information required to the system is taken which will help in improving the output. This information is fed back into the database, which will improve the further iterations of the clustering function.

Information taken would be regarding-

- The safety of the route,
- Any suspicious activity that may have occurred,
- Presence of any CCTV cameras over the route,
- Condition of street lights and
- Overall experience with the system such as UI design and usability.

4. Model Description

4.1. Creating the Dataset

The basis for the accurate functioning of the project was generated synthetically, since Mumbai Police crime data is very confidential. The support for this was provided by a form, filled by individuals, which gave us a brief idea of locations with high crime rate in Mumbai. A python library- ‘makeblobs’ from sklearn was used to create datapoints around the actual datapoints provided by individuals via the form.

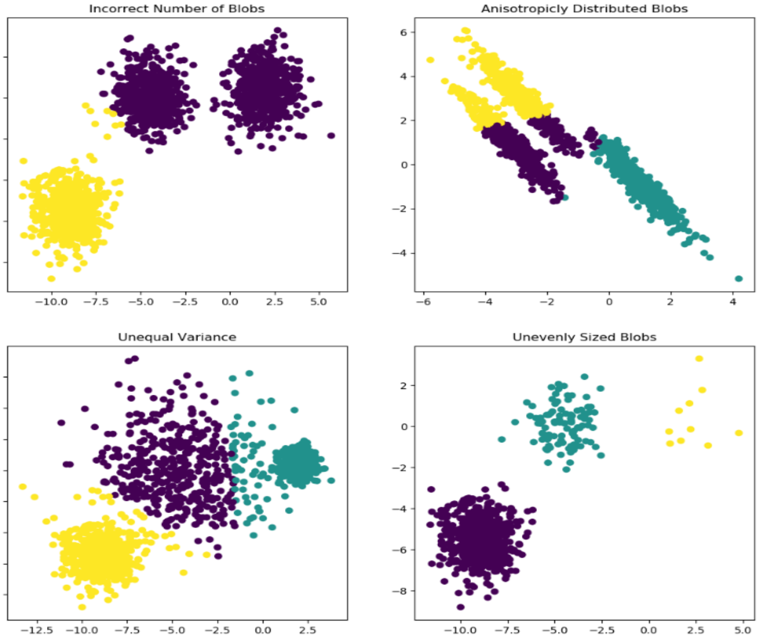


Figure 2. MakeBlobs Sample [15]

The above figure shows sample synthetic data created using 'makeblobs' and k-means result. It requires the cluster center, the cluster standard deviation and the number of samples in each cluster.

For creation of crime database, the cluster centers were fed from a google form result, which formed the basis for the application.

Table 1. Dataset Snapshot

Date	**Time**	**Latitude**	**Longitude**	**Crime**
1/1/2014	0:39	19.1394	72.9231	LarcenyMV,HitnRun,MissingPerson
1/1/2014	0:35	19.2091	72.8204	Assault(Simple),Harrassment,Threats
1/1/2014	1:05	19.0207	72.8322	Assault(Simple),LarcenyMV,ChainSnatching,Robbery
1/1/2014	2:07	19.1404	72.923	LarcenyMV,HitnRun,MissingPerson
1/1/	3:53	19.04	72.8	Assault(Simple),LarcenyMV,ChainSnatching,Robbery

2014		09	662	
1/1/2014	4:07	19.2095	72.8213	Assault(Simple),Harrassment,Threats
1/1/2014	7:35	19.0409	72.8658	Assault(Simple),LarcenyMV,ChainSnatching,Robbery
1/1/2014	11:23	19.11686819	72.8217336	Assault(Aggravated)
1/1/2014	12:40	19.1098	72.9119	Assault(Simple),Harrassment,LarcenyMV,Robbery

The crimes have been categorized into following – Assault (Simple), Assault (Aggravated), Larceny MV, Chain Snatching, Pocket Picking, Robbery, Hit and Run, Harassment, Missing Person, Sex Offender Violation, Housebreak, Threats. This will help to identify each crime instance, and maintaining a level of abstraction by hiding the details.

4.2. The Routing Mechanism

Leaflet, a javascript based API is used for map functions on the page. The leaflet 'Routing machine' provides all routing requirements for the project.

4.3. Clustering

Density-based spatial clustering of applications with noise (DBSCAN) is a data clustering algorithm proposed by Martin Ester, Hans-Peter Kriegel, Jörg Sander and Xiaowei Xu in 1996 [14]. DBSCAN classification is based two primary factors –

- EPS: the maximum distance between two samples for one to be considered as in the neighborhood of the other.
- N: 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.

A brief comparison of DBSCAN and K-Means with respect to relevance for geospatial data clustering is shown.

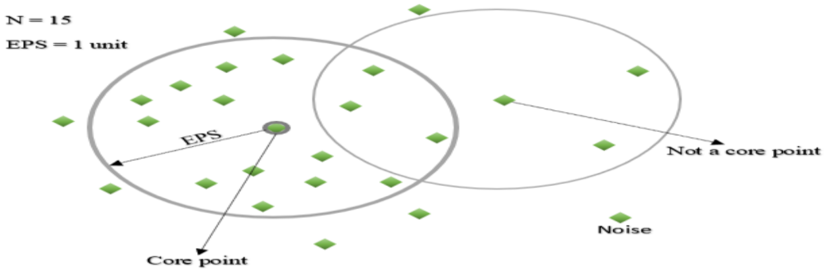


Figure 3. DBSCAN Algorithm

The DBSCAN classification factors have been explained in the Figure-3. Each data point is visited by the algorithm. For each data point, the total number of data points around the visited node in a radius of EPS is calculated. If this number is greater than or equal to N, then the visited node is declared as the “Core point”. In the Figure, the visited node is the one with thick grey border. The EPS radius includes 15 data points. Since N=15, this node is Core point. A cluster is formed for each Core point, which are separated by the distance EPS. In other words, if 2 core points exist within the EPS distance, then both the points will be a part of a single cluster. The points outside the radius are noise or outlier points.

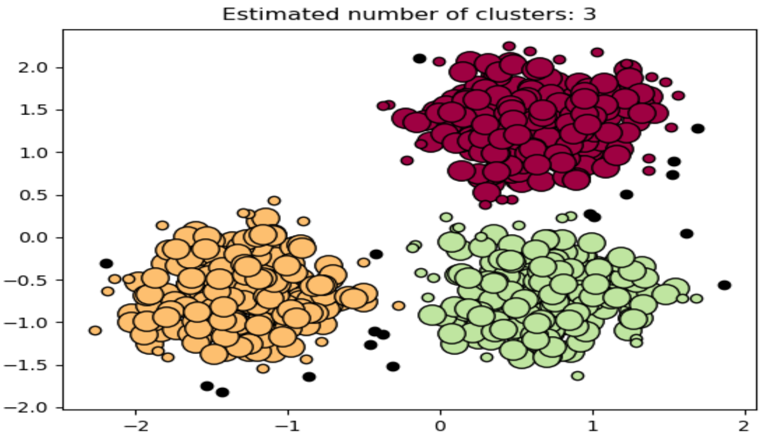


Figure 4. DBSCAN Sample Clustering Output

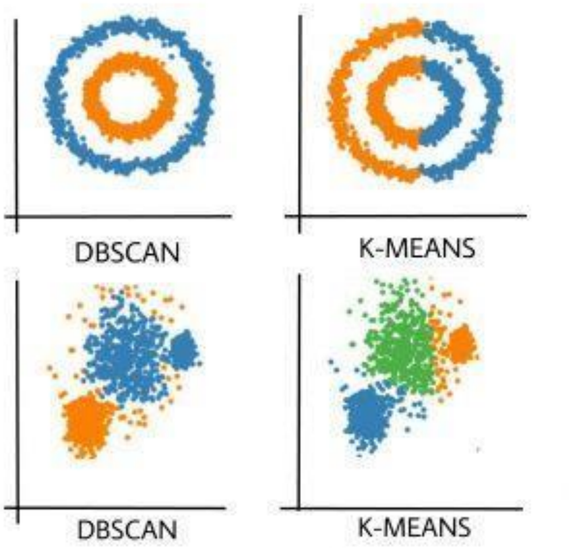


Figure 5. K-Means Sample Clustering Output

DBSCAN can identify each cluster by separating the core samples (shown with bigger coloured markers) and the noise (shown with small black markers) in Figure-4.

The above figure-5 shows clustering over the same dataset, with different algorithms. Since there is a lot of noise points involved, K-Means does not classify the exact cluster. Also, KMeans does not cluster spherical data well. The number of clusters is unknown at the beginning of the algorithm.

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 those as core points. Since DBSCAN can easily detect outliers, noise can be eliminated. Major previous papers used K-means for the same purpose. K-Means clustering algorithm is distance based, and not density based. 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.

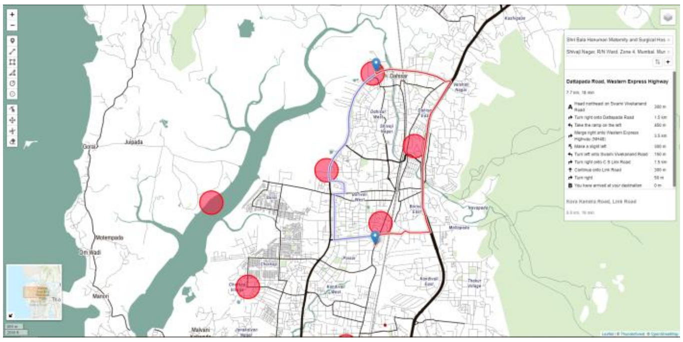
Since we are dealing with Geospatial database, clustering is required to be done on actual geographical coordinates. In such a database, algorithms using Euclidean distance measure usually fail. Consider a single point which is far away from an actual cluster. This point serves no use in pattern identification, but still the algorithm puts this point as an individual cluster. This paper attempts to cluster the data using DBSCAN algorithm. 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'.

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 [16].



Shri Bala Hanuman Maternity and Surgical Hos
Shivaji Nagar, R/N Ward, Zone 4, Mumbai, Mun

Dattapada Road, Western Express Highway
7.7 km, 18 min

A	Head northeast on Swami Vivekanand Road	300 m
↱	Turn right onto Dattapada Road	1.5 km
↰	Take the ramp on the left	450 m
↱	Merge right onto Western Express Highway (NH48)	3.5 km
↖	Make a slight left	300 m
↰	Turn left onto Swami Vivekanand Road	150 m
↱	Turn right onto C S Link Road	1.5 km
↑	Continue onto Link Road	300 m
↱	Turn right	50 m
B	You have arrived at your destination	0 m

Kora Kendra Road, Link Road
8.8 km, 19 min

Figure 6. Proposed system sample

The above figure gives an idea of the proposed system UI. The components of the figure involve a Leaflet map centred at Mumbai, Red coloured clusters which depict crime prone areas, and 2 routes from location to destination, a primary and a secondary, and a directions UI on the right by Leaflet routing machine.

6. Future Scope

Crime data is very sophisticated. It cannot be in wrong hands. There will always be scope to increase system security, admin access, reliability, stronger servers etc. To integrate a public input system where users can report some illegal activities through the app. Police data can be integrated with the system directly for improving the results. Various other technologies such as classification by classifiers, neural networks, SVMs can be used for similar applications to improve the result.

7. Conclusion

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.

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. Aljanabi, K.B. (2011). A Proposed Framework for Analyzing Crime Data Set Using Decision Tree and Simple K-Means Mining Algorithms.
4. Malathi, A., & Baboo, S.S. (2011). Algorithmic Crime Prediction Model Based on the Analysis of Crime Clusters.
5. Nath, S.V. (2006). Crime Pattern Detection Using Data Mining. *2006 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology Workshops*, 41-44.
6. Malathi., A., Baboo, D.S., & Anbarasi, A. An intelligent Analysis of a City Crime Data Using Data Mining.
7. Baidari, I., & Sajjan, S. (2016). Location Based Crime Detection Using Data Mining.

8. Kumar, P., & Ashok, M.S. (2013). A Survey of Positioning Algorithms on Mobile Devices in Location Based Services.
9. Iqbal, R., Azrifah, M., Murad, A., & Mohd, N.B. (2013). Emerging and prospering trends in crime analysis and investigation systems: a literature review.
10. Thota, L.S., Alalyan, M., Khalid, A.A., Fathima, F., Changalasetty, S.B., & Shiblee, M. (2017). Cluster based zoning of crime info. *2017 2nd International Conference on Anti-Cyber Crimes (ICACC)*, 87-92.
11. Wadhai, C.G., Kakade, T.P., Bokde, K.A., & Tumsare, D.S. (2018). Crime Analysis Using K-Means Clustering.
12. Rajkumar, S., Pandi.M, S., Jagan.J, S., & Varnikasree, P. (2019). CRIME ANALYSIS AND PREDICTION USING DATA MINING TECHNIQUES.
13. Syed, A., & Mohammad, A. (2017). Improving the Performance for Crime Pattern Analysis Using Chhaya Yadav.
14. 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).
15. Pedregosa, F., Varoquaux, Ga"el, Gramfort, A., Michel, V., Thirion, B., Grisel, O., … others. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12(Oct), 2825–2830.
16. Sumanta Das, Malini Roy Choudhury," A Geo-Statistical Approach for Crime hot spot Prediction", Department of Civil Engineering.,2016.
17. Tayal, D.K., Jain, A., Arora, S. et al. AI & Soc (2015) 30: 117. https://doi.org/10.1007/s00146-014-0539-6 Yogesh, Ashwani kumar Dubey, "Fruit Defect Detection Based on Speeded Up Robust Feature Technique", 5th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO), 2016.
18. Cesario, Cesario E, Catlett C, Talia D. Forecasting Crimes Using Autoregressive Models. InDependable, Autonomic and Secure Computing, 2016 IEEE 14th Intl C 2016 Aug 8 (pp. 795-802). IEEE.