Crime Prediction using K-means Algorithm

Vineet Jain
UG Student
Department of Computer Science and Engineering
Maharaja Agrasen Institute of Technology, Delhi

Yogesh Sharma
Assistant Professor
Department of Computer Science and Engineering
Maharaja Agrasen Institute of Technology, Delhi

Ayush Bhatia
UG Student
Department of Computer Science and Engineering
Maharaja Agrasen Institute of Technology, Delhi

Vaibhav Arora
UG Student
Department of Computer Science and Engineering
Maharaja Agrasen Institute of Technology, Delhi

Abstract

Crime analysis and prevention is a systematic approach for identifying and analyzing patterns and trends in crime. Our system can predict regions which have high probability for crime occurrence and can visualize crime prone areas. With the increasing advent of computerized systems, crime data analysts can help the Law enforcement officers to speed up the process of solving crimes. About 10% of the criminals commit about 50% of the crimes. Even though we cannot predict who all may be the victims of crime but can predict the place that has probability for its occurrence. K-means algorithm is done by partitioning data into groups based on their means. K-means algorithm has an extension called expectation - maximization algorithm where we partition the data based on their parameters. This easy to implement data mining framework works with the geospatial plot of crime and helps to improve the productivity of the detectives and other law enforcement officers. This system can also be used for the Indian crime departments for reducing the crime and solving the crimes with less time.

Keywords- Crime Prediction, K-Means, Clustering, Data Mining, Crime Prone Areas

I. INTRODUCTION

Criminals are nuisance for the society in all corners of world for a long time now and measures are required to eradicate crimes from our world. Our mission is to offer crime prevention application to keep public safe. Current policing strategies work towards finding the criminals, basically after the crime has occurred. But, with the help of technological advancement, we can use historic crime data to recognize crime patterns and use these patterns to predict crimes beforehand.

We are using clustering algorithms to predict crime prone areas. There are many clustering algorithms to group the relevant data into desired clusters. The large volumes of crime datasets as well as the complexity of relationships between these kinds of data have made criminology an appropriate field for applying data mining techniques. Criminology is an area that focuses the scientific study of crime and criminal behavior and law enforcement and is a process that aims to identify crime characteristics. It is one of the most important fields where the application of data mining techniques can produce important results. Identifying crime characteristics is the first step for developing further analysis. The knowledge gained from data mining approaches is a very useful tool which can help and support police forces. Clustering techniques converts dataset to clusters which are further examined for determining crime prone areas. These clusters visually represent group of crimes overlaid on map of police jurisdiction. Clusters store location of crimes along with other credentials of crime like type and time. These clusters are classified on the basis of their members. Densely populated clusters become crime prone areas whereas clusters with fewer members are ignored. Preventive measures are implemented according to crime type in crime prone areas.

K-means is the simplest and most commonly used clustering algorithm in scientific and industrial software. Due to less computational complexity, it is suitable for clustering large data sets. As such, it has been successfully used in various topics, including market segmentation, computer vision, geostatistics, astronomy and agriculture. It often is used as a preprocessing step for other algorithms, for example to find a starting configuration.

We chose clustering technique over any other supervised technique such as classification, since crimes vary in nature widely and crime database are often filled with unsolved crimes. Therefore, classification technique that will rely on the existing and known solved crimes, will not give good predictive quality for future crimes.

II. MATHEMATICS

Our algorithm aims at minimizing an objective function known as squared error function given by:

$$J(V) = \sum_{i=1}^{c} \sum_{j=1}^{c_i} (\|x_i - v_j\|)^2$$

Where,
'$||x_i - v_j||$' is the Euclidean distance between x_i and v_j.
'c_i' is the number of data points in i^{th} cluster.
'c' is the number of cluster centers.

A. Algorithmic Steps for K-Means Clustering

1) Let $X = \{x_1,x_2,x_3,……..,x_n\}$ be the set of data points and $V = \{v_1,v_2,…….,v_c\}$ be the set of centers.
2) Randomly select *'c'* cluster centers.
3) Calculate the distance between each data point and cluster centers.
4) Assign the data point to the cluster center whose distance from the cluster center is minimum of all the cluster centers..
5) Recalculate the new cluster center using:

$$v_i = (1/c_i) \sum_{j=1}^{c_i} x_i$$

Where, 'c_i' represents the number of data points in i^{th} cluster.
6) Recalculate the distance between each data point and new obtained cluster centers.
7) If no data point was reassigned then stop, otherwise repeat from step 3).

III. METHODOLOGY

A. Clustering

Clustering is an unsupervised task without having a priori knowledge by discovering groups of similar documents. There are two types of categories in clustering algorithms; they are the partitioned algorithm and the hierarchical algorithm. K-Means algorithm and the link clustering they come under these two categories. K-Means and hierarchical clustering have many comparisons. In hierarchical clustering the size of data increases as the computational expansive, since to merge small clusters and D_D similarity matrix by using the certain link functions. By comparing with them K-Means is faster. It updates the centroid clusters with each iteration and reallocates each document by its nearest centroid by this we can say that it is an iterative algorithm.

B. K-Means Algorithm

K-Means clustering investigation plans to partition n perceptions into k bunch during which each perception includes a place with the bunch with the nearest centroid.

C. Process

1) Initially, the number of clusters must be known let it be k
2) The initial step is to choose a set of K instances as centers of the clusters.
3) Next, the algorithm considers each instance and assigns it to the cluster which is closest.
4) The cluster centroids are recalculated either after whole cycle of re-assignment or each instance assignment.
5) This process is iterated.

K means algorithm complexity is O (tkn), where n is instances, c is clusters, and t is iterations and relatively efficient. It often terminates at a local optimum. Its disadvantage is applicable only when mean is defined and need to specify c, the number of clusters, in advance. It unable to handle noisy data and outliers and not suitable to discover clusters with non-convex shapes.

IV. IMPLEMENTATION AND RESULTS

A. Experimental Setup

Step 1: Create a new server on the web hosting sites available
Step 2: Create two databases; one for storing the details of the authorized user and the other for storing details of the crime occurring in a particular location
Step 3: The data can be added to the database using SQL queries
Step 4: Create PHP scripts to add and retrieve data.
The project is implemented by following steps:
Step 1: The PHP file to retrieve data converts the database in the JSON format.
Step 2: This JSON data is parsed from the android so that it can be used.
Step 3: The location added by the user from the android device is in the form of address which is converted in the form of latitudes and longitude that is further added to the online database.
Step 4: The added locations are marked on the Google map.
Step 5: The various crime types used are Robbery, Kidnapping, Murder, Burglary and Rape. Each crime type is denoted using a different color marker.

Step 6: The crime data plotted on the maps is passed to the k –means algorithm.
Step 7: The data set is divided into different clusters by computing the distance of the data from the centroid repeatedly.
Step 8: A different colored circle is drawn for different clusters by taking the centroid of the clusters as the center where the color represents the frequency of the crime
Step 9: This entire process of clustering is also performed on each of the crime types individually.
Step 10: In the end, a red colored circle indicates the location where safety measures must be adopted.

B. Results

Fig. 1 represents the crimes occurred in Delhi by using a different colored marker for each crime. This crime data is obtained from the latitude and longitude of each crime stored in the online database.

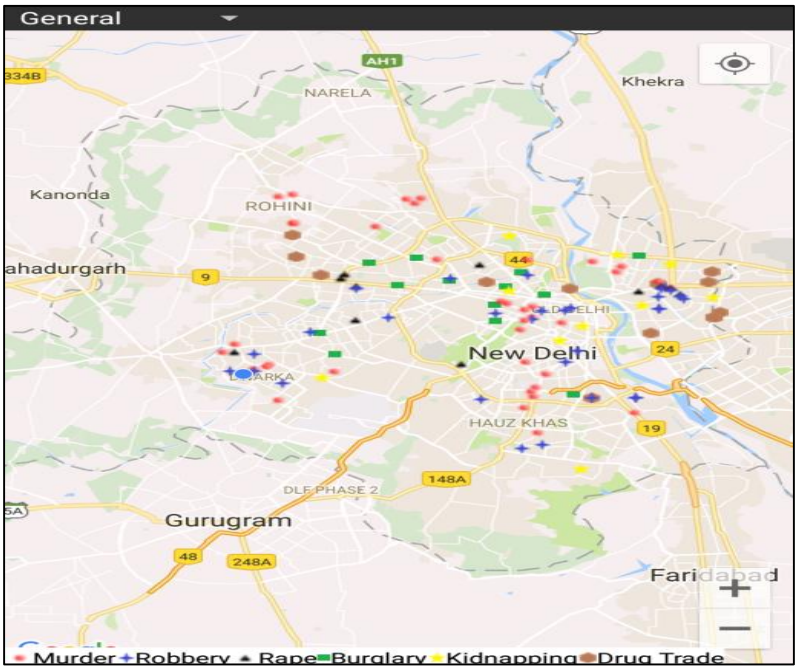


Fig. 2 represents the clusters formed by k means algorithm applied on the data set. The red circle denotes the location of maximum crime occurrence, the blue circle denotes the location of moderate crime occurrence and black circle denotes the location of low crime occurrence.

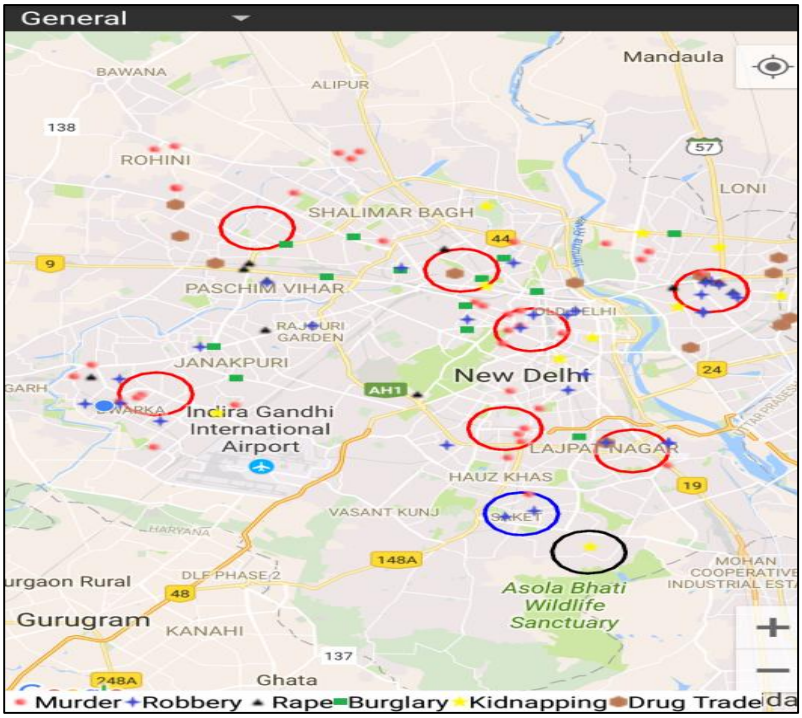
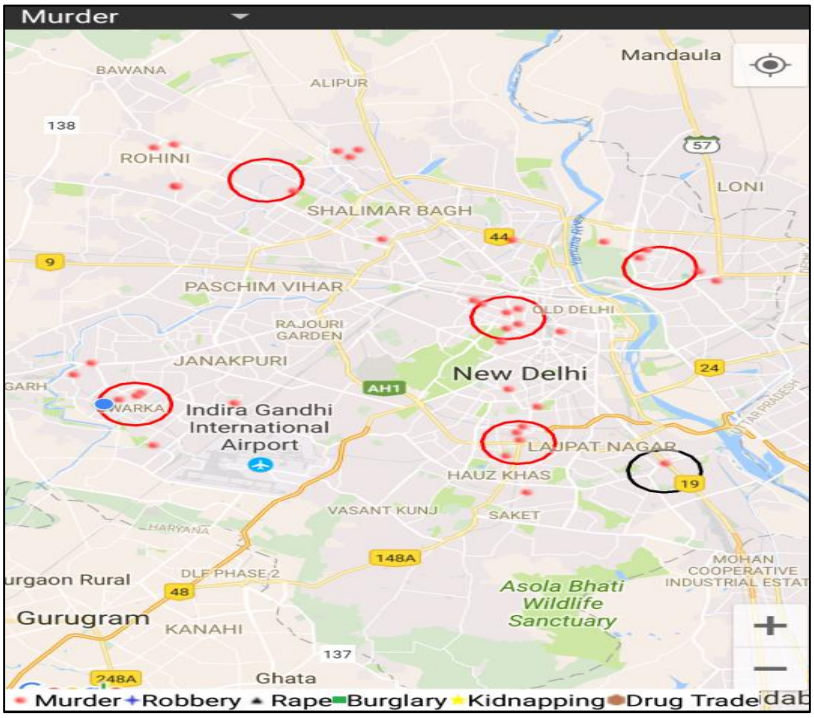


Fig. 3 indicates only the murders occurred in Delhi. Similar results are observed for other crimes such as Kidnapping, robbery, burglary etc.



V. CONCLUSIONS

From the clustered results it is easy to identify crime prone areas and can be used to design precaution methods for future. The classification of data is mainly used to distinguish types of preventive measures to be used for each crime. Different crimes require different treatment and it can be achieved easily using this application.

Experimental results prove that application is effective in terms of analysis speed, identifying common crime patterns and crime prone areas for future prediction. The developed application has promising value in the current complex crime scenario and can be used as an effective tool by Indian police and enforcement of law organizations for crime detection and prevention.

REFERENCES

[1] K. Zakir Hussain, M. Durairaj and G. Rabia Jahani Farzana, "Application of Data Mining Techniques for Analyzing Violent Criminal Behaviour by Simulation Model", International Journal of Computer Science and Information Technology & Society, Vol. 02, No. 01, ISSN: 2249-9555, 2012
[2] A. Malathi, Dr. S. Santhosh Baboo, "Algorithmic Crime Prediction Model Based on the Analysis of Crime Clusters", Global Journal of Computer Science and Technology Vol. 11, No. 11, pp. 139-145, 2011.
[3] Manish Gupta, B. Chandra and M. P. Gupta, "Crime Data Mining for Indian Police Information System", Computer Society of India, Vol. 40, No. 1, pp. 388-397, 2008
[4] Malathi. A, Dr. S. Santhosh Baboo and Anbarasi A., "An intelligent Analysis of a City Crime Data Using Data Mining", International Conference on Information and Electronics Engineering, IPCSIT, Vol. 06, 2011
[5] Kadhim B. Swadi Al-Janabi, "A Proposed Framework for Analyzing Crime Data Set Using Decision Tree and Simple K-Means Mining Algorithms", Journal of Kufa for Mathematics and Computer, Vol. 01, No. 03, pp. 08-24, 2011
[6] Rasoul Kiani, Siamak Mahadavi, Amin Keshavarzi, "Analysis and Prediction of Crimes by Clustering and Classification", International Journal of Advanced Research in Artificial Intelligence, Vol. 04, Issue 8, 2015
[7] Jyoti Agarwal, Renuka Nagpal, Rajni Sehgal, "Crime Analysis using K-Means Clustering", International Journal of Computer Applications(0975-8887), Vol. 83, No. 04, 2013
[8] Shyam Varan Nath, "Crime Pattern Detection Using Data Mining", IEEE Transactions on Knowledge and Data Engineering, Vol. 18, No. 09, pp. 41-44, 2010
[9] Sasha Kapoor, Abhineet Kalra, "Data Mining for Crime Detection", International Journal of Computer Engineering and Applications, Volume VII, Issue III, September 14