DOI: http://dx.doi.org/10.26483/ijarcs.v8i9.4862

ISSN No. 0976-5697

Volume 8, No. 9, November-December 2017

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

MINING OF CRIMINAL MINDSET AND GEOSPATIAL CRIME PATTERN DETECTION

Anchal Oberai
Department of IT
Bharati Vidyapeeth's College of Engineering
Delhi, India

Malika Jain
Department of IT
Bhagwan Parushram Institute of Tech.
Delhi, India

Gautam Khosla
Department of IT
Bhagwan Parushram Institute of Tech.
Delhi, India

Priyanka Dewan
Department of IT
Bhagwan Parushram Institute of Tech.
Delhi, India

Monika Arora
Department of IT
Bhagwan ParushramInstitute of Tech.
Delhi, India

Abstract:The current scenario is witnessing an exponential increase in the crime rate in many countries. Therefore, to combat this upsurge, data mining has been a vital concept adopted by the law enforcers for prediction and analysis. Data mining is the practice of examining large pre-existing databases in order to generate new and useful information. This paper introduces a methodology of mining the psychological behavior of criminals based on the type of crime committed, the motive behind it and by analyzing the relationship between the criminal and the victim. It also includes Geospatial (state-wise) crime pattern detection. The analysis has been done on real crime data by using K-means clustering algorithm.

Keywords: Data Mining, K-Means clustering, Crime, Criminal, Mindset

1. INTRODUCTION

The desire to have sybaritic lifestyle and irresistible ambitions has lured the humans to indulge into criminal activities. The law enforcers have to face various challenges for crime control and to ensure maintenance of public order. For this, they need new technologies and methodologies to study, analyze and predict the crime patterns and the mindset of criminals. Data mining, also known as data or knowledge discovery is a process of analyzing data from different perspectives and summarizing it into useful information and has proved to be an instrumental approach in field of criminology.

In this paper, a database has been created and maintained from the unstructured criminal data available online on the official website of Legal Information Institute of India[1]. This data has been manually transformed into structured data and then further converted into RDD (Resilient Distributed Dataset) to predict the psyche of the criminals. The relationship between aging and crime activity has been observed since the beginning of criminology. Hirschi el al(1983) [2] claimed that the age-crime relationship is invariant or universal across groups, societies and times, a claim that has been reiterated as grounds for focusing on biological or evolutionary explanations for age-crime relationships. Hence, we have performed k-means clustering on the age groups of the criminals. Clustering is an unsupervised learning problem whereby we aim to group subsets of entities with one another based on some notion of similarity. K-Means algorithm has been used to cluster the data points into pre-defined number of clusters. Each cluster has been categorized according to the type of crime, the motives of the criminals behind it, the criminal-victim relationship and states. This paper particularly highlights on two issues. Firstly, crime committed by criminals of various age groups using K-means algorithm. And the other is, classification of the criminals according to various attributes like type of crime, motives, criminal-victim relationship and states.

In this paper, section 2 gives the existing contributions in the field of criminology. Section 3 elaborates K-Means clustering technique and section 4 explains the methodology adopted for the analysis. Section 4 provides experimental results with real crime data sets that confirm the virtue of our approach. Section 5 draws conclusions. and explains the future expansion of this study.

2. RELATED WORK

Data Mining of criminology can be used for suppressing the crime rate in the country.

Uttam Mande et al [3] introduces binary clustering technique for clustering of crime data. The dataset is generated using binary format of 1's and 0's, and then clusters of types of crimes are created using the weighing scheme.

Jyoti Aggarwal et al [4] has analyzed homicide by applying K-means clustering using the Rapid Miner tool.

Kadhim B. Swadi Al-Janabi [5] has used WEKA tool to analyze the crime dataset. It is a second generation tool for analysis with less dimensions and accuracy..

Tony H. Grubesic el at[6] has used hierarchical clustering to help analysts examine the concentration of crime events in geographical areas. He has explained the problems associated with it is that there is no guarantee that, one ends up with p (required) groups. If we are intent on identifying entities or areas that are strongly related in some predefined sense, then K-Means clustering is potentially useful.

As an extension to the existing contribution in the field of criminology, we have used K-Means clustering algorithm. It creates pre-defined number of clusters of different age-groups without converting the dataset into binary format . We have used Apache Spark (support Hadoop) tool which has more dimensions for analysis and data processing is faster.

3. CLUSTERING APPROACH: K-MEANS ALGORITHM

K-Means is an unsupervised learning algorithm used to classify a given data set through a certain number of clusters (assume k clusters) fixed apriori. It focuses to define k centroids, one for each cluster. These centroids are placed in a crafty way as different location causes different result. So, the better way is to place them as far away from each other as possible. Then each point belonging to a given data set is taken and is associated to the nearest centroid. k new centroids are re-calculated as barycenter of the clusters resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data set as far away from each other as possible. Then each point belonging to a given data set is taken and is associated to the nearest centroid. k new centroids are re-calculated as barycenter of the clusters resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new centroid. As a result, we may notice that the k centroids change their location step by step until no more changes are done or in other words centroids do not move any more.

4. PROPOSED SYSTEM ARCHITECTURE

Here, the crime analysis has been done by applying K-Means clustering algorithm using Spark with Scala. The procedure is as follows :

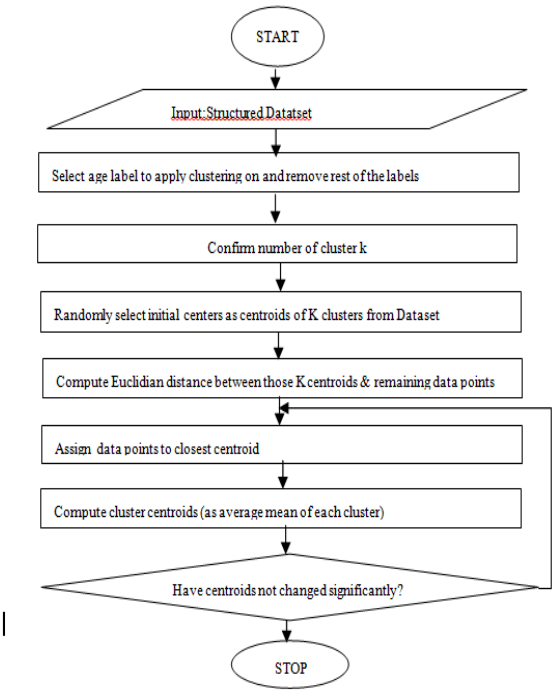


Fig. 1. Flowchart of K-Means Cluster Algorithm

Step 1: Transform unstructured criminal data into structured dataset manually and store in CSV format. Replace missing values are by a '?'.
Step 2: Open IntelliJ IDEA and add Scala plugin and Spark dependency.
Step 3: Convert the CSV file into Resilient Distributed Dataset(RDD).
Step 4: Convert age attribute (numeric) of the dataset into vectors as k-mean clustering is applicable only on vectors.
Step 5: Apply K-Mean clustering on age attribute.
Step 6: Classify the crimes ,states, relationship with victim and criminal motives into various age-groups.
Step 7: Analysis is done on resulting clusters.

5. EXPERIMENTAL SETUP AND RESULTS

K-Means Clustering Analysis

K-Means operates on the featured vectors in Spark. Therefore, the domain values of the age attribute in the RDD data are first converted into vectors. These vectors are then used for clustering. For each resulting cluster, the centroids, the distance of the data-point to its nearest cluster's centroid and the Euclidian distances have been calculated. For each value of k, the average distance to the centroid has been displayed. A point past which increasing k stops reducing the score much is considered to be the most appropriate value for k. Each cluster represents a specific age-group. Criminals in these age-groups have been further classified into three sets:

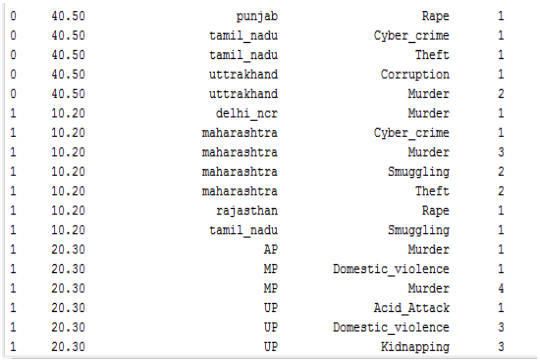
- on the basis of the type of crime and motive behind it.
- on the basis of the states and the type of crime(for geospatial crime pattern)
- on the basis of the type of crime and the relationship with the victim.

Following are the images showing clusters of age-group formed by using Apache Spark, further classified

0	30.40	Smuggling	Greed_Money	1
0	30.40	Theft	Greed_Money	14
0	40.50	Theft	Greed_Money	3
1	20.30	Acid_Attack	Hatred_Jealousy	2
1	20.30	Acid_Attack	Revenge_SettlingScores	1
1	20.30	Corruption	Greed_Money	1
1	20.30	Cyber_crime	Greed_Money	1
1	20.30	Domestic_violence	Greed_Money	4
1	20.30	Dowry	Greed_Money	2
1	20.30	Foeticide	Greed_Money	1
1	20.30	Kidnapping	Greed_Money	4
1	20.30	Kidnapping	Lust_Love	3
1	20.30	Murder	?	12
1	20.30	Murder	Greed_Money	4
1	20.30	Murder	Hatred_Jealousy	1
1	20.30	Murder	Lust_Love	2
1	20.30	Murder	Revenge_SettlingScores	35
1	20.30	Rape	Hatred_Jealousy	2
1	20.30	Rape	Lust_Love	28
1	20.30	Rape	Revenge_SettlingScores	1
1	20.30	Smuggling	Fraud_IllegalGain	2
1	20.30	Theft	Fraud_IllegalGain	3
1	20.30	Theft	Greed_Money	22
1	20.30	Theft	Revenge_SettlingScores	1
2	50.60	Corruption	Greed_Money	1
2	50.60	Dowry	Greed_Money	1

Fig. 2. On basis of type of crime and motive behind it

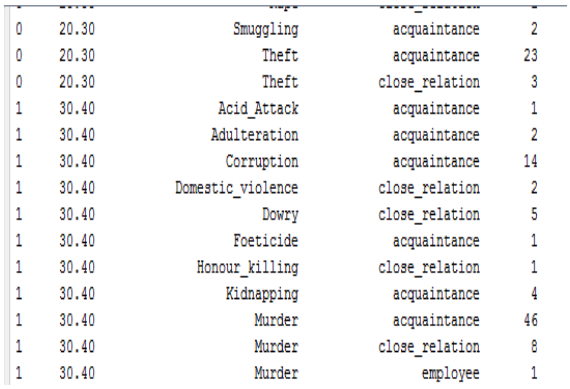
In Fig. 2.,the first column indicates cluster number, the second column indicates the age-group, the third column indicates the type of crime , the fourth column indicates the motive behind the crime and last column indicates the total number of criminals.



0	40.50	punjab	Rape	1
0	40.50	tamil_nadu	Cyber_crime	1
0	40.50	tamil_nadu	Theft	1
0	40.50	uttrakhand	Corruption	1
0	40.50	uttrakhand	Murder	2
1	10.20	delhi_ncr	Murder	1
1	10.20	maharashtra	Cyber_crime	1
1	10.20	maharashtra	Murder	3
1	10.20	maharashtra	Smuggling	2
1	10.20	maharashtra	Theft	2
1	10.20	rajasthan	Rape	1
1	10.20	tamil_nadu	Smuggling	1
1	20.30	AP	Murder	1
1	20.30	MP	Domestic_violence	1
1	20.30	MP	Murder	4
1	20.30	UP	Acid_Attack	1
1	20.30	UP	Domestic_violence	3
1	20.30	UP	Kidnapping	3

Fig. 3. On basis of state and type of crime

In Fig. 3.,the first column indicates cluster number, the second column indicates the age-group, the third column indicates states , the fourth column indicates the type of crimes and last column indicates the total number of criminals.



0	20.30	Smuggling	acquaintance	2
0	20.30	Theft	acquaintance	23
0	20.30	Theft	close_relation	3
1	30.40	Acid_Attack	acquaintance	1
1	30.40	Adulteration	acquaintance	2
1	30.40	Corruption	acquaintance	14
1	30.40	Domestic_violence	close_relation	2
1	30.40	Dowry	close_relation	5
1	30.40	Foeticide	acquaintance	1
1	30.40	Honour_killing	close_relation	1
1	30.40	Kidnapping	acquaintance	4
1	30.40	Murder	acquaintance	46
1	30.40	Murder	close_relation	8
1	30.40	Murder	employee	1

Fig. 4. On basis of type of crime and criminal-victim relationship

In Fig.4.,the first column indicates cluster number, the second column indicates the age-group, the third column indicates the type of crimes , the fourth column indicates the relationship between criminal and victim and last column indicates the total number of criminals.

Given below are the following graphs representing the analysis of age-group 20-30 from the above outputs:

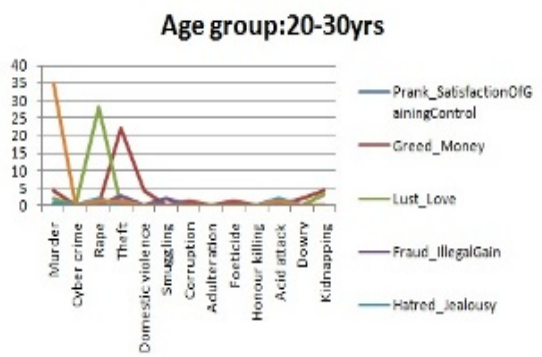


Fig. 5. Graph between type of crimes(x-axis) and number of criminals(y-axis) for various motives

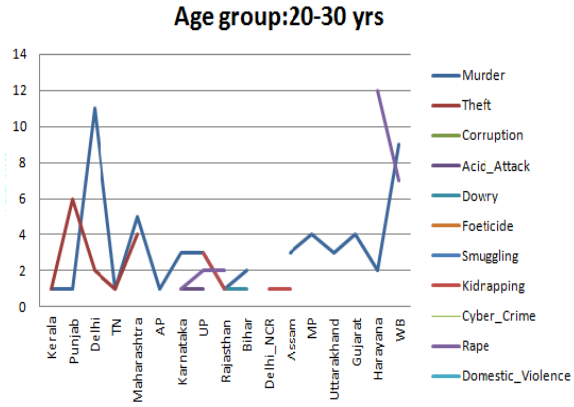


Fig. 6. Graph between states(x-axis) and number of criminals(y-axis) for various Crimes

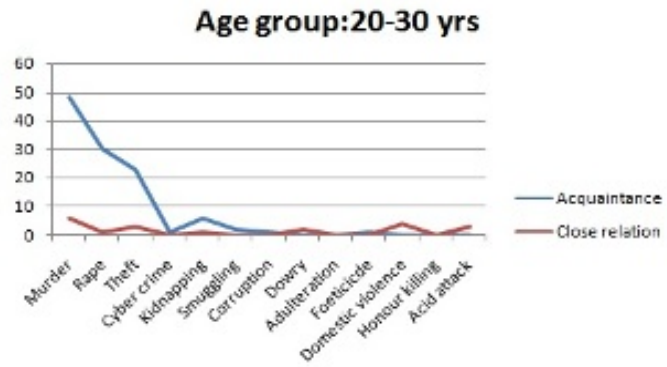


Fig. 7. Graphbetween type of crimes(x-axis) and number of criminals(y-axis) for criminal-victim relationship

From the graphs, we observe that the cluster of an age-group 20-30, the criminals have committed maximum murder with a motive of taking revenge or settling score and mostly of acquaintances. Maximum murders have occurred by criminals belonging to Delhi. So from this, it can be inferred that the criminals of Delhi have committed murder with a mindset of taking revenge or settling score.

Similarly, the analysis has been done for all the type of crimes for the remaining age-groups, including ages from 0 to maximum human lifespan (90).

Dataset Used

The data has been collected from the unstructured criminal data available online on the official website of Legal Information Institute of India . It has been converted into structured dataset in the form of table given below.

Table 1. Dataset of real-criminal record

C_name	C_gender	State	Motive	Crime	Relation	Age-group
Sujit_nair	male	Maharashtra	Greed_Money	Murder	Acquaintance	20-30
Ayush_bhat	male	Maharashtra	Greed_Money	Murder	Acquaintance	10-20

Tools Used

In this paper, following tools have been used.

- Apache Spark support Hadoop [7]:It is a fast and general engine for large scale data processing to implement clustering in Machine Learning Algorithms. It supports big data analysis and shows magnificent accuracy in results.
- Scala language: It has been used in for coding.
- IntelliJ IDEA :Its Community Version has been used as Integrated Development Environment.

6. CONCLUSION AND FUTURE SCOPE

We looked at the use of data mining for identifying psychological behavior of criminals and predicting the geospatial crime patterns. Our contribution here was to analyze the mindset of criminals as a machine learning task using Apache Spark(support Hadoop) among massive databases. It is found that crime rises rapidly in early adolescence, peaks in late adolescence and during 20's, and levels off and declines slowly during older ages. Appropriate counseling session for various age-groups can be conducted considering their mindsets mined in this paper Nevertheless, there are limitations. Firstly, this methodology can only be used to cluster structured data. It doesn't support clustering of unstructured data. So, it is time consuming to convert unstructured data to structured data to apply clustering algorithm. Secondly, only numeric data can be clustered using the methodology mentioned in this research paper. In order to apply this technique on textual data, one-hot coding can be applied to transform text to vector data. As a future extension of this study, we can apply Association Rule Learning to discover the association between different crimes committed by the same criminal. Vladimir Estivill-Castro and Ickjai Lee [8] stated association-rule mining has been a powerful tool for discovering correlations among massive databases. For instance, if a criminal commits rape then he may also murder the same victim. This implies that the commission of rape is likely to be followed by murder of the same victim. Some additional attributes like weapon used, gender of victim and criminal, evidence related to the crime and information given by witnesses can be considered to make the analysis more accurate and effective.

REFERENCES

[1] Legal Information Institute of India(LIIofIndia), www.liiofindia.org/in/cases/cen/INSC/

[2] Ulmer, J.T., Steffensmeier, D.: Trends, Current Issues and Policy Implications.SAGE Publications Ltd. (2014)

[3] Mande, U., Srinivas, Y., Murthy, J.V.R.: Witness Based Criminal Identification Using Data Mining Techniques and New Gaussian Mixture Model. International Journal of Modern Engineering Research 2, 1507-1510 (2012)

[4] Agarwal, J., Nagpal, R.:Crime Analysis using K- Means Clustering, International Journal of Computer Applications(0975-8887), vol. 83 (2013)

[5] Kadhim B.Swadi Al-Janabi: A Proposed Framework for Analyzing Crime Data Set Using Decision Tree and Simple K-Means mining Algorithm. Journal of Kufa for Mathematics and Computer 1, 8-24 (2011)

[6] Grubesic, T.H., Murray, A.T.: Detecting Hot Spots Using Cluster Analysis and GIS. Proceedings from the fifth Annual International Crime Mapping Research Conference 26 (2001)

[7] Apache Spark, spark.apache.org/documentation.html Estivill-Castro, V., Lee, I.:Data Mining Techniques for Autonomous Exploration of Large Volumes of Geo-referenced Crime Data