CRIME PREDICTIVE MODEL USING BIG DATA ANALYTICS

RajaniKanth Aluvalu [1], Tirthraj Chauhan [2]

[1] Department of CSE, Vardhaman College of Engineering, Hyderabad, India
[2] Department of Computer Engineering, Darshan Institute of Engineering, Rajkot, Gujarat, India

Corresponding author: RajaniKanth Aluvalu, rajanik.rkcet@gmail.com

ABSTRACT: With the increasing crime rate, the technological advancement is also increasing which can be considered as a reason for increasing the crime rate. The crime related data is in any of the formats i.e., structured, semi-structured or unstructured data. In early time the data recorded was less and mostly in the structured format so it was easy to analyze that data. If it is possible to analyze the structured, semi-structured and unstructured data (collectively known as Big Data) then it would be beneficial to security authority to use the past data for the prediction purpose. Here we are using R studio for analyzing the Big Data. First, we obtained the data and a plot that data on the map. Then by applying the clustering algorithm on the data we plotted and finally we plot the clustered data on the basis of which prediction can be made. This can be considered as the way that how a Big Data Analytics can be used in developing crime predictive model.
KEY WORDS: Crime prediction, Unstructured data, Kmeans clustering, R language, KDE.

1. INTRODUCTION

Human society is facing the issue of crime nowadays. When there is an intersection between the personal or professional space of a person or group of person the crime may occur [SB15]. The person wanted to commit the crime, initially research the place and person along with the study of it accordingly to implement crime. Crime can occur at secluded places with the higher rate [SB15].
For the prevention of crime it is required to allocate the security resources in a proper way and for allocating resources the crime prediction is required. Crime prediction can be done with the help of crime mapping [GAB07]. This is how crime mapping is responsible for reducing the crime from the society by recognizing the Hotspots (the area having higher crime rate) [MES15].
The past data related to criminal activities plays a vital role in mapping crime and prediction of places where crime can occur [MES15]. Data generation nowadays is vast due to increased crime rate which cannot be handled by traditional data analysis techniques. This vast generated data is Big Data which can be easily treated with the help of Big Data Analytics [NS13]. Digital data may be structured, semi-structured or unstructured. Mostly the digital data which are analyzed till now was a structured kind of data for predicting crime [NS13]. Structured data can be considered as the data arranged in tabular format with the help of suitable rows and columns. After applying some data mining techniques like clustering, classification, and other techniques the places having higher chances of crime to be occur were identified and police capabilities can be allocated there. Nowadays the use of the internet is increasing rapidly. The use of the internet is also responsible for providing communication between criminals for completing their targeted mission. So the data generation is in huge amount which is mostly in semi-structured or unstructured data format and can be analyzed using clustering for Big Data [JV14]. To analyze such huge amount of data either in semi-structured or unstructured format traditional data mining techniques are not that much capable. For that purpose data, Big Data Analytics is used [BS14].
R tool is used to distribute the data geographically. This tool is capable of generating a geospatial representation of data geographically distributed data. Different packages are available with this tool which needs to be installed in order to perform the data distribution. Data analysis, as well as different visualization patterns of distributed data, can be obtained from this tool.

2. RELATED WORK

Crime can be considered as an "act against the law which harms the innocent peoples and results in acquiring punishments from the legal authorities like law enforcement or judiciary authority of government". Different types of crime are mainly traffic violations, fraud, sex crime, arson, drug offenses, violent crimes, murders, robbery, damage, theft and cyber-crime [SB15]. It can be observed that the past data which were relevant to criminal activities are helpful for predicting the crime hotspots.
Crime data analysis can be done using data mining techniques with the tools like weka tool, rapid minor tool, R tool, KNIME, ORANGE and Tanagra etc. Mostly the crime data analysis is done using k-means

clustering technique of data mining. Due to development in technology the criminals are using their technological equipment for doing the crime. That digital data is being used to analyze the crime [MES15]. The analyzed crime will be useful in predicting the hotspots. Again the data used for analyzing and for prediction purpose using data mining is structured data, when there is unstructured or semi-structured data, data mining techniques are somewhat time-consuming at that moment [NS13]. Obtained criminal data was taken, preparing that data for rapid minor tool and perform k-means clustering on that data to obtain the clusters. After obtaining clusters, analyzing that clusters to predict the crime.

Table 1. The comparative study of all the 3 types of digital data

	Structured Data	Semi-Structured Data	Unstructured Data
Characteristics	- data is stored in the form of rows and columns - conforms to a data model - attributes in a group are the same	- does not conform to any data model but contains tags and elements (metadata) - attributes in a group may not be the same - similar entities are grouped	- not in any particular format or sequence - does not conform to any data model - not easily usable by a program - does not follow any rules or semantics
Sources	- Databases - Spreadsheets - SQL - OLTP systems, etc.	- E-mail - XML - Zipped files - Mark-up languages, etc.	- Web pages - PowerPoint presentations - Videos, Images - Reports - Surveys, etc.
Challenges faced	- limited storage - contains only homogeneous data	- storage cost - limited tools available - no ready tool available for querying. -data heterogeneity.	- indexing and searching - security (varied sources of data) - retrieve information - lack of technical expertise

Other data mining techniques can also be applicable to analyze the crime data and prediction can be done to identify the hotspots. Other technique includes mainly classification, a K-means clustering algorithm, Expectation maximizing algorithm etc. After applying a K-means clustering algorithm it might provide improved results then what we obtain after only applying k-means clustering [NS13]. K-Means algorithm can be implemented in Big Data Analytics [JV14]. These are some traditional and time-consuming techniques to map the crime as it requires more over the structured data.

To distribute the data relevant to crime geographically is also a tedious task but now it can be implemented using the tools like R tool. With some geospatial packages that need to be installed and running with the R tool will greatly influence the data to be distributed over geographic areas. Clustering of that criminal data can be done using appropriate technology. [HMO15] In Big Data Analytics GA (Genetic Algorithm) based clustering can also be implemented for analyzing or implementing clustering. The drawback of using Artificial Neural Networks is to learn how to implement that [CD14].

Table 2. Comparative Analysis of different mining approaches used

S. no	Approach	Concept	Accuracy	Performance	Drawback
1	Support vector machine	Find a linear hyperplane (decision boundary) that will separate the data.	Good accuracy	Robust to noise and reduced overfitting	Runs at slower speed
2	Multivariate Time Series Clustering	Based on Minkowski Model	Prediction is good	With different dimensions gives good performance	Dimensions are required to have the same weightage.
3	Bayesian Network	Uses Bayes Theorem	Depends mainly on Geographical factor which in return gives higher accuracy	Good performance	Fully dependant on selection of parameter
4	Artificial Neural Network	Processing using processing elements termed as neurons	Accuracy of prediction is high	Fast evaluation	Time is required for training
5	Fuzzy Time Series	Use of fuzzy logic is the main concept behind Fuzzy time series	Better prediction even if some of the data are not available	Performance is better in time or state space	Results are affected by the different factors

3. PROPOSED WORK

Using R tool, Big Data Analytics and Artificial Neural Network we are going to perform the crime mapping. It contains mainly three phase – Distribution of data geographically and creating clusters, Cluster analysis of created clusters and prediction of crime.

Distribution of data geographically is the first phase where the available data is distributed over geographical areas. Here the available data is related to crime. This can be implemented using the R tool with the geospatial packages. With that, the clusters are created after allocation of centroids. The KDE (Kernel Density Estimation) technique will be used for estimating or creating clusters on the basis of mapped data. As this created clusters are being utilized by cluster analysis phase.

Hadoop platform is used for cluster analysis purpose which is the second phase. Clusters created in primary phase are used as input this phase and the suitable clustering algorithm is to apply over here for the analysis purpose. Hadoop can perform parallel processing on different clusters the processing will be fast as compared to traditional processing capabilities. This will result in less time consumption

and gives the output earlier than the normal data mining cluster analysis process. The GAMMA Test is used for cluster analysis of this phase.

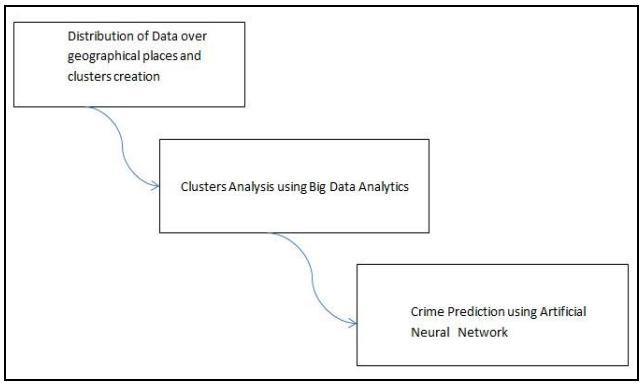

Figure 1. Crime Prediction process Model

As shown in figure 1, Analyzed data of cluster analysis i.e., identified cluster from the Hadoop is utilized by the Artificial Neural Network as an input for crime forecasting purpose is a third and final phase. It also uses the regression tree prediction specification and classification. The output of Artificial Neural Network is the pattern that predicts the crime rate at different places or the places where the chances of crime occurrence are high.

4. EXPERIMENTAL EVOLUTION

4.1. Obtaining and Plotting Data on Map

For the experimental purpose, the data is obtained in from the data.gov.in site for the year 2013 for the country India [***16]. The different reasons for crime are there in the data. The facts and figure of the related crime are also there along with the different city and states of India. Primarily the obtained data is being plotted on the map using the R studio and the R version 3.3.0. The Figure 2 shows the plot which is plotted using the ggmap and ggplot packages.

4.2. Clustering the plotted data

The plotted data is clustered using the k-means clustering algorithm. For this purpose, we have to import the cluster package in the R studio.

K-Means Clustering algorithm:

- Input: Number of clusters
- Randomly select k objects from dataset D which belongs to N objects as the starting centers of clusters
- Reassigning of all of the objects with consideration of cluster center and the similarity or say minimum distance
- Recalculate the mean of the cluster, calculation of object's mean value for all the clusters.
- Output: A set of K clusters

Above is the K-Means algorithm steps are given but in R studio it is quite easy to implement K-means clustering. Simply assigning the data matrix along with the numbers of clustered to be created as arguments in the K-Means() function [CD14]. The clusters are created after executing the command. The clustering is based on the data we provided. Here clustering is based on the distance between the cities and the numbers of crime.

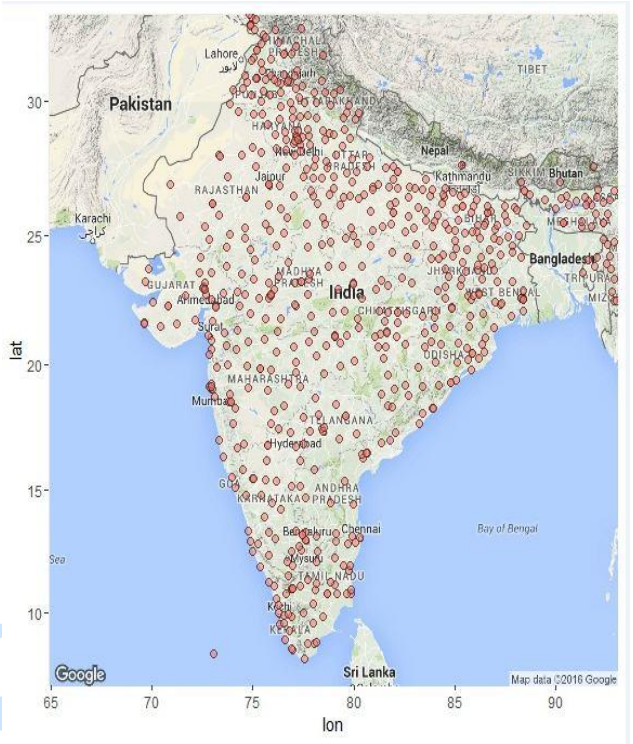

Figure 2. Plotting of crime data on Map of India

4.3. Re-plotting the Clustered Data

After the clusters creation, the clustered data is plotted on the map which is interactive in nature. For creating interactive maps we are using leaflet package in R studio and for the facilities like downloading of crime data of specific clusters as well as for displaying the crime facts and figures on the plotted mark after clicking it we are using the shiny framework. For creating such interactive maps, we created an application in the shiny framework which is based on the client-server architecture. The plotting of clustered data is shown in Figure 3.

CONCLUSION

The proposed work focuses on crime prediction by crime mapping with recorded data using the latest technology. The model helps in reducing crime for the security authorities. The implemented work show that how Big Data Analytics with R studio can help in developing crime predictive model. The model also helps the authorities in the investigation of crimes. Using Bigdata analytics with clustering approach reduces the investigation time and helps in retrieving the hidden information through correlation and categorization.

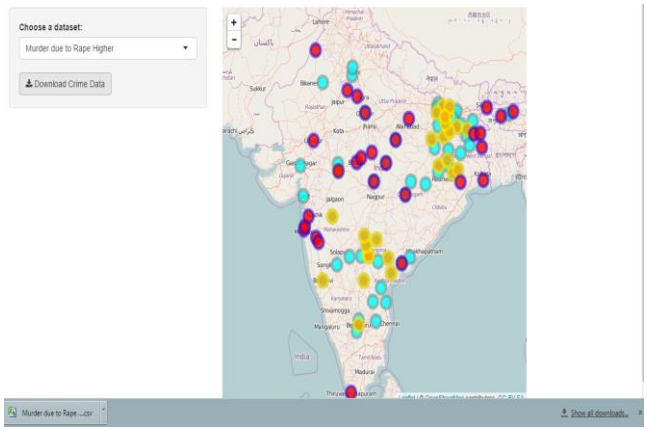

Figure 3. Re-plotting the Clustered data on interactive maps

FUTURE WORK

The performed experiment is limited to structured data. Nowadays there is a huge generation of Semi-Structured and Unstructured data especially in the field of crime. The Crime related data mainly have the Semi-Structured data so in future using the tools for semi-structured and unstructured data the analytics can be performed. The application created for the interactive maps are purely based on the clusters we created. In future, it is possible that a developer may develop the application that accepts the csv or xls data directly and demands the numbers of clusters with the parameter based on which the clusters are to be created. The plotting will be available with just importing the data in an application of shiny framework and leaflet maps.

REFERENCES

[BS14] A. Bharthi, R. Shilpa - *A Survey On Crime Data Analysis of Data Mining using Clustering Techniques*, International Journal of Advance Research in Computer Science and Management Studies, Volume 2, Issue 8, August 2014.

[CD14] Setu Kumar Chaturvedi, Nikhil Dubey - *A Survey Papaer on Crime Prediction Technique Using Data Mining*, Int. Journal Of Engineering Reasearch and Applications, Vol. 4, Issue 3 (version 1), March 2014.

[GAB07] Vikas Grover, Richard Adderley, Max Bramer - *Review of Current Crime Prediction Techniques*, in Applications and Innovations in Intelligent Systems XIV, pp. 233-237, 2007.

[HMO15] Nivranshu Hans, Sana Mahajan, S. N. Omkar - *Big Data Clustering Using Genetic Algorithm On Hadoop Map reduce*, International Journal Of Scientific & Technology Research Volume 4, Issue 4, April 2015.

[JS15] Shalini Jain, Satendra Sonare - *Big Data Analysis Using HDFS, C-MEANS and Map reduce*, International Journal Of Advanced Reasearch in Computer Science and Software Engineering, Volume 5, Issue 4, 2015.

[JV14] Mugdha Jain, Chakradhar Varma - *Adapting K-means for Clustering in Big Data*, International Journal of Computer Application, Volume 101-No.1, September 2014.

[MES15] Lenin Mookiah, William Eberle, Ambareen Siraj - *Survey of Crime Analysis and Prediction*, Proceedings of the 28th International Florida Artificial Intelligence Research Society Conference, 2015.

[NS13] Renuka Nagpal, Rajni Sehgal - *Crime Analysis using K-Means Clustering*, International Journal of Computer Applications (0975 – 8887) Volume 83 – No 4, December 2013.

[SB15] Saoumya, Anurag Singh Baghel - *A Predictive Model For Mapping Crime Using Big Data Analytics*, IJRET, eISSN: 2319-1163, 2015.

[SM12] K. K. Sindhu, B. B. Meshram - *A Digital Forensic Tool for Cyber-Crime Data Mining*, Engineering Science and Technology: An International Journal (ESTIJ), Vol.2, No.1, 2012.

[SS15] Keshav Sanse, Meena Sharma - *Clustering methods for Big data analysis*, IJARCET, Volume 4, Issue 3, March 2015.

[SMW15] Ms. Sonali. B. Maind, Ms. Priyanka Wankar - *Research Paper on Basics Of Artificial Neural Network*, International Journal on Recent and Innovation Trends in Computing and Communication, Volume:2, Issue :1, 2015.

[S+10] Konstantin Shvachko, Hairong Kuang, Sanjay Radia, Robert Chansler - *The Hadoop Distributed File System*, Proceedings of the 2010 IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST), 2010.

[***16] https://data.gov.in/catalog/district-wise-crimes-under-various-sections-indian-penal-code-ipc-crimes, accessed 2016.