Analysis and Prediction of Crime in India

Tanya Gautam(MT18048), Ekta Tank(MT18147), Sujay Raj(MT18108)

Abstract

Crime Analysis is crucial for the governments for numerous aspects and for enforcing policies in order to re-establish law and order. Crime rate is one way of analyzing the crime status. We present an alternative measure for analyzing crime in the states and Union Territories of India based on their frequencies as well as severity during different years. The term severity has here been defined in terms of the amount of punishment prescribed in the IPC for a crime. The procedures of Clustering followed by Classification are applied to the crime dataset recorded between 2001 to 2012. Clustering divides the states and Union Territories into three main categories: 'Safe', 'Moderate' and 'Relatively Dangerous'. Classification is used to later determine the crime category of a given state or Union Territory based on the existing available data.

1 Introduction

Crime, in literal terms, is an action violating the given set of rules. A crime is punishable on the account of its severity and its extent. It is defined specifically on the premise of civilization and bound to act upon it.[1]

Crime analysis is a law enforcement function that involves systematic analysis for identifying and analyzing patterns and trends in crime and disorder. Information on patterns can help law enforcement agencies deploy resources in a more effective manner, and assist detectives in identifying and apprehending suspects. Crime analysis also plays a role in devising solutions to crime problems and formulating crime prevention strategies. Quantitative social science data analysis methods are part of the crime analysis process.[2]

Crime in India is reported by the National Crime Records Bureau (NCRB). The NCRB collects and evaluates the data and on the basis of which, comparison is made subsequently.

The main purpose of crime analysis in our project is:

i. Extraction of crime patterns by crime analysis

ii. Prediction of crime status of a state and Union Territory based on the spatial distribution of existing using various data mining techniques.

The project categorizes each state and Union Territory into one of the three categories defined as:

i. Relatively Dangerous - The states and Union Territories with the most frequent occurrences of the most severe crimes.

ii. Safe - The states and Union Territories with the least frequency of the most severe crimes, although the frequencies of less severe crimes may be more.

iii. Moderate - The remaining states and Union Territories are categorized as moderate.

2 Motivation behind the project

One way of analyzing the crime status of different states and Union Territories in India is via calculating the crime rates. The calculation of crime rate is done by dividing the total number of crimes reported by the police by the total population of that state and Union Territory, thus every crime is treated the same. We have presented an alternative analysis of crime which also takes into account the severity of the crimes performed.

3 Literature Review

J. Agarwal, R. Nagpal and R. Sehgal in [3] have analyzed crime and considered homicide crime taking into account the corresponding year

and that the trend is descending from 1990 to 2011. They have used the k-means clustering technique for extracting useful information from the crime dataset using RapidMiner tool because it is solid and complete package with flexible support options.

Rasoul Kiani et al in [4] have used k-means clustering followed by classification for classifying clustered crimes based on occurrence frequency during different years. They have applied a theoretical model based on data mining techniques such as clustering and classification to real crime dataset recorded by police in England and Wales within 1990 to 2011. We assigned weights to the features in order to improve the quality of the model and remove low value of them.

4 Methodology

4.1 Collection of dataset

The dataset has been obtained from the open source datasets available on kaggle.com.
Link: https://www.kaggle.com/rajanand/crime-in-india
The dataset consists of the state-wise and district-wise database of different IPC crimes based on the frequencies of occurrences of those crimes between the years 2001 and 2012.

4.2 Pre-processing phase

4.2.1 Data Cleaning : The data of all districts of a state and Union Territory was merged into a single record since we are not doing the district-wise analysis.

4.2.2 Feature Selection: Attributes like 'Custodial Rape', 'Other Rape' that were specific instances of a single crime were merged into a single attribute.

4.2.3 Division of all crime frequencies by the total number of IPC crimes: To obtain a quantity similar to the crime rate for each particular crime, we divided, for each state and Union Territory, the crime frequencies by the total number of crimes that occurred in that state and Union Territory. The assumption here is that more the population more is the total number of crimes since we could not get the exact population of every state and Union Territory for every year between 2001 to 2012.

4.2.4 Min-Max Scaling: Every attribute was min-max scaled to get values between the range 0 to 1.

4.2.5 Biasing: Applying our self-defined weights to the attributes: In order to make the more severe crimes the more distinguishing factors between the states, we applied four different weights viz. 4,3,2 and 1 to the attributes.

The values of these weights were determined as follows:

i. Punishment for the crime according to IPC is less than 5 years: Weight=1.

ii. Punishment for the crime according to IPC is between 5 years and 10 years: Weight=2.

iii. Punishment for the crime according to IPC is above 10 years but not life sentence: Weight=3.

iv. Punishment for the crime according to IPC is life sentence: Weight=4.

4.3 Clustering Phase

K-Means clustering with three centroids was used.

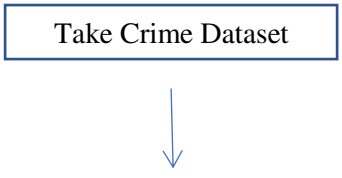
The clusters obtained were labeled as 'Relatively dangerous', 'Moderate' and 'Safe'. We studied the cluster centroids in order to assign the labels. The centroid with the highest frequencies of the most severe crimes was labeled as 'Relatively dangerous'. Similarly, the other two labels were also assigned.

Cluster Validation was done using the Silhouette Coefficient.

4.4 Classification Phase

KNN classifier with K=1 was used. The data of all states for all years, each combination of the two representing a unique data point was fed as the training input to the classifier and 10 fold cross-validation was performed.

Figure 1. shows the system design and process of project



Year	Silhouette coefficient
2001	0.2012
2002	0.2487
2003	0.3036
2004	0.2240
2005	0.2954
2006	0.2317
2007	0.2303
2008	0.2081
2009	0.1809
2010	0.19481
2011	0.19481
2012	0.2088

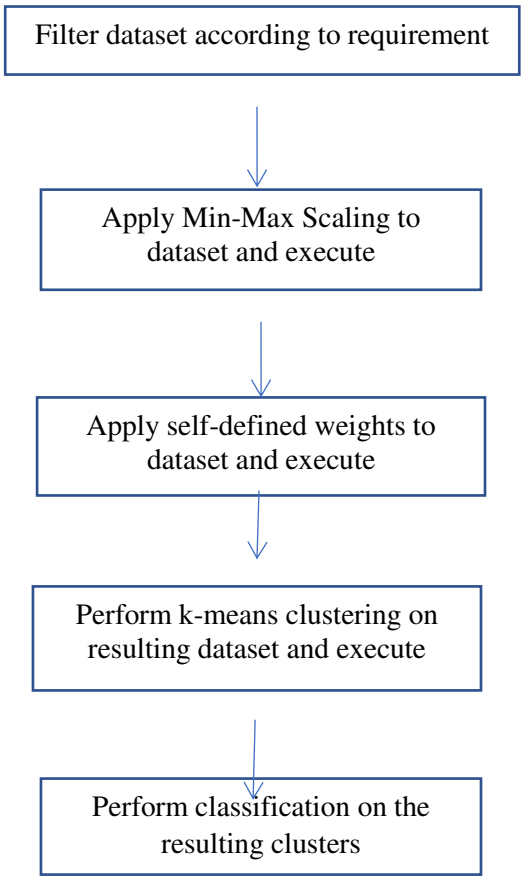


Figure 1: Flowchart of Crime Analysis

5.1 Clustering Validation:

Table 1. The accuracies reported by the Silhouette coefficient for each year.

The overall average Silhouette co-efficient of cluster validation is 0.2269.

5.2 Classification Validation:

The classification accuracy reported by 10-Fold validation is 88.23%.

6 Observations:

- Green : Safe State
- Yellow: Moderate State
- Red: Relatively Dangerous State

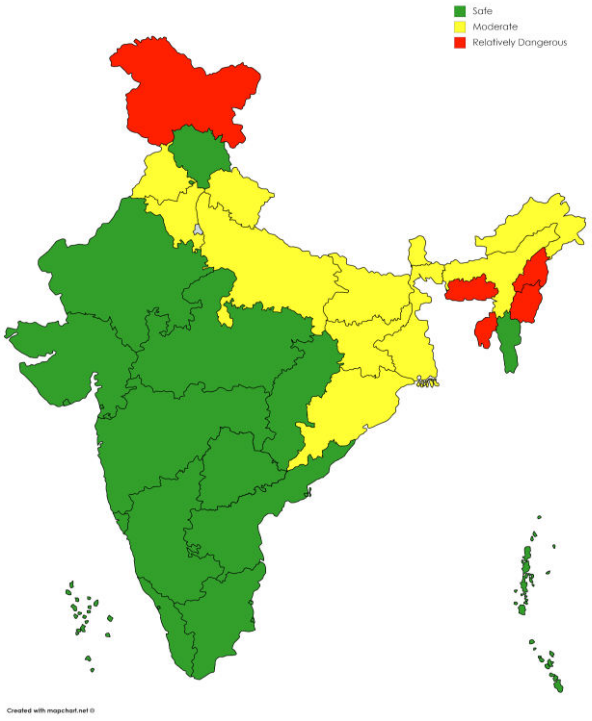


Figure 2. Cluster labeling for 2001

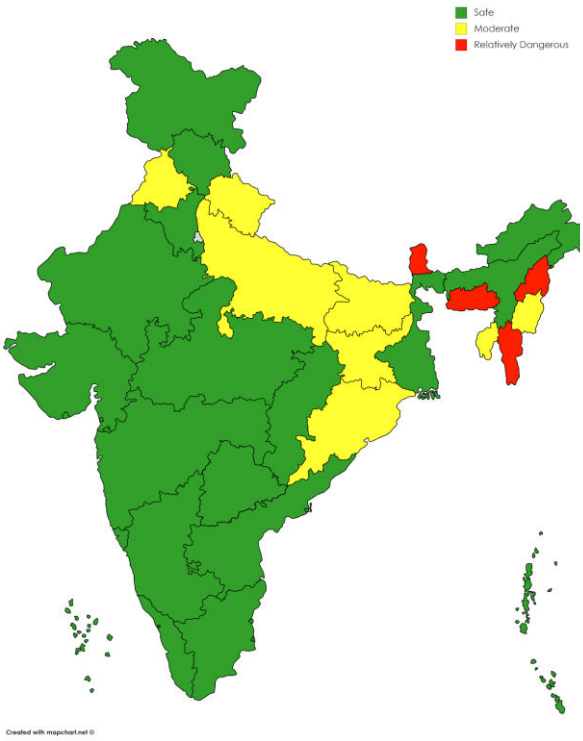


Figure 3. Cluster labeling for 2012

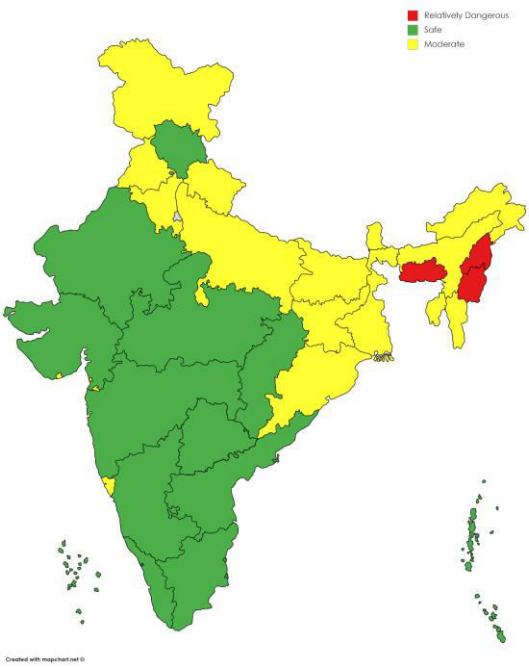


Figure 4. The average crime scene between the years 2001 to 2012

States like Nagaland, Mizoram, and Meghalaya have consistently been labeled as 'Dangerous' since they have very high rates of murders, robberies, and dacoities respectively. In contrast, according to the crime rate statistics of India, Nagaland State stands 30th in all crime records in 2013[5].

The states like Lakshadweep, Madhya Pradesh, Maharashtra, Gujarat, Himachal Pradesh, and others have consistently been categorized as Safe.

The state of Sikkim used to be Safe for the first few years but then gradually moved to the Moderate and then to the Dangerous category finally. This shows an alarming situation for the government of Sikkim and proper measures should be taken.

6.1 State-wise Reports:

- Andhra Pradesh: The state of Andhra Pradesh has stayed in the Safe category throughout from 2001 to 2012.
- Arunachal Pradesh: The state of Arunachal Pradesh sometimes appears in the Moderate category and sometimes in the Safe category.
- Assam: The state of Assam has mostly been labeled as Moderate and it has relatively very high occurrences of rapes compared to the other states.
- Bihar: The crime scene in Bihar is similar to that of Assam with thefts being one of the major crimes in the state.
- Chhattisgarh: The state of Chhattisgarh has been classified in the safe category throughout the analysis with relatively very low frequencies of murders and rapes compared to most of the other states.
- Goa: The state of Goa has been mostly classified as Safe with a transition to Dangerous for the years 2010-11 due to relatively high numbers of robberies and thefts.
- Gujarat: The state of Gujarat has throughout been classified as Safe during the analysis period.
- Haryana: The state of Haryana was clustered in the Moderate category for the earlier years till 2005, then moved to the Safe category. It has medium occurrences of most of the crimes.
- Himachal Pradesh: The state of Himachal Pradesh has been labeled as Safe throughout

the analysis with the state having very low numbers of robberies compared to the rest of the nation.

- Jammu and Kashmir: Dangerous in the earlier years, Jammu and Kashmir transitioned to the Safe category, however, there has been an increase in the number of rapes.
- Jharkhand - Mostly labeled as moderate.
- Karnataka, Kerala, Madhya Pradesh, Maharashtra - These states have been labeled as Safe throughout the analysis.
- Manipur - The state of Manipur has been mostly Dangerous or Moderate because of very high rate of rapes compared to the rest of the nation.
- Meghalaya - The state of Meghalaya has been mostly Dangerous because of very high rate of dacoities.
- Mizoram - The state of Mizoram has transitioned from the Safe category to Moderate and then subsequently to the Dangerous category.
- Nagaland - The state of Nagaland has been consistently Dangerous because of the highest rate of murders compared to the rest of the nation.
- Odisha, Punjab - The states of Odisha and Punjab have been mostly labeled as moderate.
- Rajasthan - The state of Rajasthan has been labeled as safe.
- Sikkim - The state of Sikkim has also slowly moved from Safe category to Dangerous like Mizoram.
- Tamil Nadu - The state of Tamil Nadu has been labeled as a safe state throughout the analysis.
- Tripura - The state of Tripura has been mostly Moderate, with sometimes being Dangerous due to the higher rate of kidnappings and abductions compared to the rest of the country.
- Uttar Pradesh, Uttarakhand: The state of Uttar Pradesh has mostly been labeled as Moderate.
- West Bengal - Mostly labeled as a moderate state.
- Andaman and Nicobar Islands, Chandigarh, Lakshadweep, Pudducherry - Always labeled as Safe states.
- Dadra and Nagar Haveli, Daman and Diu - From being in the Safe category, these Union Territories drifted towards the Dangerous category for the years 2010 and 2011 due to high rates of kidnappings and dacoities in Daman and Diu, and kidnappings and riots in Dadra and Nagar Haveli.

7 Results and Applications:

We have successfully performed the analysis and prediction of crime in India using the data mining techniques of Clustering and Classification. The results obtained from this project offer insights into the dominant nature of crimes in the different states and Union Territories. The findings of this project can be used to review the law and order enforcement in a particular state.

References :

1. https://community.data.gov.in/crime-in-india
2. https://en.wikipedia.org/wiki/Crime_analysis
3. J. Agarwal, R. Nagpal, and R. Sehgal, ―Crime analysis using k-means clustering,‖ International Journal of Computer Applications, Vol. 83 – No4, December 2013.
4. Kiani Rasoul, Mahdavi Siamak, and Keshavarzi Amin, ―Analysis and Prediction of Crimes by Clustering and Classification,‖International Journal of Advanced Research in Artificial Intelligence, Vol. 4 – No.8, 2015.
5. http://www.neighbourhoodinfo.co.in/crime/Nagaland