International Journal of Engineering & Technology, 7 (2.6) (2018) 253-258

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research Paper

Crime analysis in India using data mining techniques

Deepika K.K [1]*, Smitha Vinod [2]

[1] *PG Research Scholar, Department of Computer Science, Christ (Deemed to be University), India*
[2] *Associate Professor, Department of Computer Science, Christ (Deemed to be University), India*
* *Corresponding author E-mail:* kkdeepika047@gmail.com

Abstract

An approach for crime detection in India using Data mining techniques is proposed in this paper. The approach consists of the following steps - Data pre-processing, clustering, classification and visualization. Data mining techniques are often applied to Criminology as it provides good results. Criminology is a field which studies about various crime characteristics. Analyzing crime data means exploring crime data. Crime is identified using k-means clustering and the clusters are formed based on the similarity of the crime attributes. The Random Forest algorithm and Neural networks are applied on the data for classification. Visualization is achieved using the Google marker clustering and the crime spots are marked on the India map. The accuracy is verified using WEKA tool. This approach will benefit the Crime department of India in analyzing crime with better prediction. The paper focuses on the crime analysis of various Indian states and union territories during 2001 to 2012.

Keywords: *Clustering, Classification, Visualization, K-means, Random Forest, Neural Networks*

1. Introduction

As we all know crime is an offense that is prohibited by law. Crimes can be categorised into various types based on the type of crime. Crimes are classified into (1) Property crime – it includes burglary, theft of vehicles and other types of theft, (2) Organised crime - includes drug trafficking, money laundering, murder, fraud, human trafficking and poaching, and (3) Corruption. National Crime Records Bureau (NCRB) published a report about crime rates during 1953 to 2006. According to the report burglary and robbery has reduced over a period of 53 years by 79.84% and 28.85% respectively. Crimes like murder and kidnapping has hiked by 7.39% and 47.80% respectively. Crime analysis which is a part of Criminology plays a very important role in crime detection. Crime analysis can be defined as a task which includes exploration and identification of crimes and their relationships with criminals [1]. Some of the popular crimes in India are – (1) Nithari serial murders during the year 2005-2006 in Uttar Pradesh, where several dead bodies of children were found in the sewer, (2) Terrorist attacks in Mumbai city in the year 2008 (November 2008) killed around 166 people and injured about 308, (3) Rape case (Nirbhaya) in the year 2012 in which a paramedical student was brutally raped and she died due to the severe injuries while undergoing treatment [2]. All these crimes have played with the emotions of many people and has stayed in the news headlines for several days. Some of the cases are not yet solved thus resulting in injustice to the victim and victim's family. Data mining helps in solving the crimes faster and this technique gives good results when applied on crime dataset, the information obtained from the data mining techniques can help the police department.

Crime is committed by people of all age groups. Earlier only males were reported committing crimes whereas now reports prove that both men and women commit crimes. There is a change in the trends of crimes and it is very challenging to find the new trends and patterns in crime [3]. The knowledge which is acquired form the data mining techniques will help in reducing crimes as it helps in finding the culprits faster and also the areas that are most affected by crime.

2. Background and Related work

Since crime is a growing concern in every part of the world it is very essential to find techniques to reduce it and also enable the police officials to easily catch the culprits. There are many approaches in solving crimes faster and a lot of researches are going on to find the best technique in Data mining. The authors of this paper developed a new tool to track the culprits. Two algorithms Data Association and Back Propagation NN-Classifier are used to analyse the data stored in the database. In order to extract criminal relations from an incidents summary and to create a group of suspects two approaches are used; With the help of BPN Classifier and Data Association algorithm the network is partitioned into subgroups and the interaction pattern is studied. The results prove that BPN-Classifier is very accurate in identifying the crime patterns and also for future predictions [1].

A CDCI (Crime Detection and Criminal Identification) technique was used to fasten the process of detecting the crimes in our Indian cities. In this technique the criminals were identified based on features like suspects name, sex, origin, facial features, crime reason, location, weapon used, etc. It had six main modules – data extraction, pre-processing, clustering, map representation, classification and WEKA tool. K-means algorithm was used for crime detection and it generated two clusters of crime. The KNN classification was used for criminal identification. The combination of k- means and KNN helped in improving the filtration for large databases [2].

The authors focused on the day-to-day factors rather than the causes for crime occurrences like the culprit's background or political enmity. The proposed system can predict the regions with high crime occurrences and also visualize those regions. The system will help the investigating officials to resolve crimes faster. The steps followed in this approach are data collection, classification, pattern prediction and visualization. Bayes theorem is used for classification and by using this algorithm the news articles were trained and the model was built. Apriori algorithm helps in finding the frequent patterns of a particular region. The system developed predicts crime regions in India on a particular day [3].

The paper concentrated on analysing the approaches between Computer Science and Police department as one of the main application of data mining. Pattern detection technique has been implemented and suggestion for future prediction is also included. K-means algorithm is used for clustering and this will help in identifying the patterns of crime and hence, will help in solving crimes faster. In order to increase the accuracy of prediction semi-supervised technique is used. The crimes are represented using Geo- spatial spots. Based on the selection of time range, type of crime and geographical region the results are shown graphically [4].

The authors used algorithms like Naïve Bayesian, K-nearest Neighbor and Neural Networks (Multilayer-Perceptron) and proved that it is better than Decision tree and Support Vector Machine. Two different feature selection methods are tested on the dataset. Comparison of algorithms are carried out on the basis of Area Under Curve (AUC). The Chi-square feature selection technique is used for improving the performance of data mining results. KNN gives better results by using Chi-square feature selection technique. The dataset chosen is categorised into two different types [5].

One of the widely used technique for studying crime characteristics is Hotspot Mapping. The distribution of spatial crime depends on socio-economic and other crime factors. A new crime hotspot tool was developed –Hotspot Optimization Tool (HOT) and the main module of HOT is the Geospatial Discriminative Patterns (GDP). The GDP can find the differences between two classes in a dataset containing spatial information. HOT can accurately map the crime hotspots and is very effective in searching and utilizing patterns in a geospatial space. Grid thematic mapping is used to represent spatial distributions [6].

In this paper the authors concentrated on building a forecasting model in cooperation with the police division of the US city in the Northeast. The approach first extracts the dataset from the original crime record and the dataset contains details like the crime location, time and other crime related factors. In this approach a classification technique is used for crime forecasting. The best classification method is chosen after analysing a variety of methods [7].

ata mining methods like clustering and classification is used by the authors of the paper. A system is developed to analyse the crime information and this will help in automating the investigation in India. The tool developed is very useful in identifying the criminals and hence, helps in speeding up the investigation [8].

In this paper the authors focused on the features that lead to increased crime rate. Both supervised and unsupervised learning techniques are used and based on the ranking of the features the predictions are made. The Random Forest classifier gives better accuracy [9].

3. Methodology

This section consists of the following subsections: 3.1 Dataset collection, 3.2 Proposed approach.

3.1 Dataset collection

In this step the dataset is collected from the kaggle website (https://www.kaggle.com/rajanand/crime-in-india/). The dataset named Crime in India by Rajanand llangovan is chosen for this research (Fig 1). The State-wise data from 2001 to 2012 is considered. The dataset consists of 34 attributes where the 34th attribute is the class variable. There are 8597 instances in the dataset. The dataset consists of 7 Union-territories and 28 states with the corresponding districts. The crime in each district is recorded from 2001 to 2012. In the dataset different types of crimes (attributes) are considered like murder, rape, kidnapping, dacoity, robbery, burglary, cheating, dowry deaths, arson, etc.

The total IPC (Indian Penal Code) crime for each district is given in the dataset. Since the dataset does not contain a class variable it was calculated manually using the total_IPC_crime attribute. If the total_IPC_crime is greater than the average crime rate, then the region is considered as critical else it is considered as non-critical. The dataset is very useful in spotting the crime prone regions on the India map. The dataset is implemented using WEKA. The dataset correctness is verified using WEKA tool. Weka tool has inbuilt feature selection technique and it is very easy to analyse the data using this tool.

3.2 Proposed approach

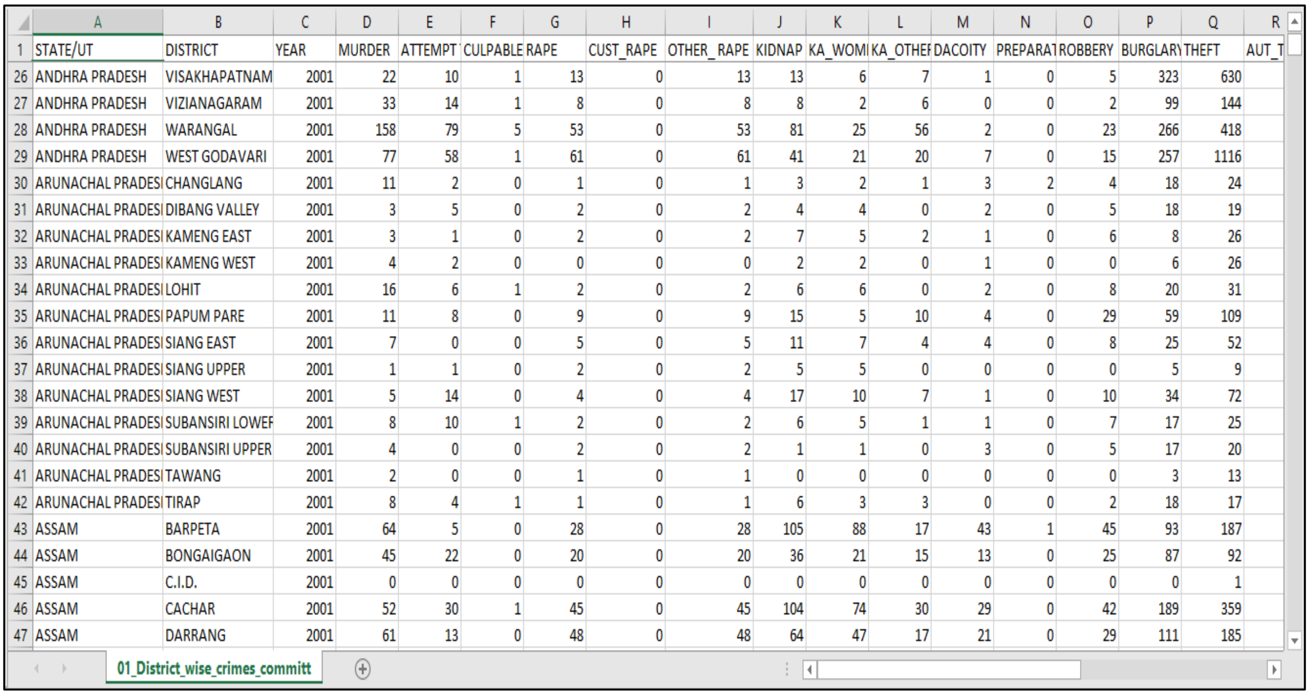
In this section the workflow of the proposed model is depicted as shown in Fig 2. The first step in the model is Data pre-processing (DP) which includes filling the missing values, data cleaning and transformation of data. The k-means algorithm is used to replace the missing values and it is replaced with the mean/mode value of the corresponding attributes instance. The k-means algorithm classifies the crime instances into clusters with alike attributes by performing the required number of iterations. The k-means clustering is then followed by the classification and this process is divided into two steps firstly, building a model and then using the model for classification. Classification helps in finding a set of models which can be used for future prediction of unknown class labels. By using the training dataset, the predictive accuracy of the model is measured. The correctness depends on how many instances are correctly classified and if the accuracy is good the model can be used for future prediction. In this approach two classification algorithms have been used to check which gives better results for the chosen dataset and its kind. The algorithms used are Random Forest and Neural Networks (MutilayerPerceptron).

The Random Forest algorithm is a supervised learning technique which creates a forest with the number of trees. It can be used for regression, classification and other tasks. The Random forest works by creating multiple decision trees during training. Using Random forest, the variables can be ranked on the basis of their priority.

Neural Networks which is a nonlinear model helps in modelling real world complex relationships. The algorithm is capable of estimating the posterior probabilities which helps in setting up the classification rules and also in conducting the statistical analysis [10].

Neural Networks generates algorithms capable of learning and recognizing patterns. It basically has three layers: input layer, hidden layer and output layer. The neurons represent the relationship or connection between the three layers [11].

Google map marker clustering (GMAPI) is used in this research and it is very helpful in representing the crime prone regions of a country. GMAPI requires attributes like a locations latitude and longitude.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	STATE/UT	DISTRICT	YEAR	MURDER	ATTEMPT	CULPABLE	RAPE	CUST_RAPE	OTHER_ RAPE	KIDNAP	KA_WOMI	KA_OTHEF	DACOITY	PREPARAT	ROBBERY	BURGLARY	THEFT	AUT_T
26	ANDHRA PRADESH	VISAKHAPATNAM	2001	22	10	1	13	0	13	13	6	7	1	0	5	323	630	
27	ANDHRA PRADESH	VIZIANAGARAM	2001	33	14	1	8	0	8	8	2	6	0	0	2	99	144	
28	ANDHRA PRADESH	WARANGAL	2001	158	79	5	53	0	53	81	25	56	2	0	23	266	418	
29	ANDHRA PRADESH	WEST GODAVARI	2001	77	58	1	61	0	61	41	21	20	7	0	15	257	1116	
30	ARUNACHAL PRADESI	CHANGLANG	2001	11	2	0	1	0	1	3	2	1	3	2	4	18	24	
31	ARUNACHAL PRADESI	DIBANG VALLEY	2001	3	5	0	2	0	2	4	4	0	2	0	5	18	19	
32	ARUNACHAL PRADESI	KAMENG EAST	2001	3	1	0	2	0	2	7	5	2	1	0	6	8	26	
33	ARUNACHAL PRADESI	KAMENG WEST	2001	4	2	0	0	0	0	2	2	0	1	0	0	6	26	
34	ARUNACHAL PRADESI	LOHIT	2001	16	6	1	2	0	2	6	6	0	2	0	8	20	31	
35	ARUNACHAL PRADESI	PAPUM PARE	2001	11	8	0	9	0	9	15	5	10	4	0	29	59	109	
36	ARUNACHAL PRADESI	SIANG EAST	2001	7	0	0	5	0	5	11	7	4	4	0	8	25	52	
37	ARUNACHAL PRADESI	SIANG UPPER	2001	1	1	0	2	0	2	5	5	0	0	0	0	5	9	
38	ARUNACHAL PRADESI	SIANG WEST	2001	5	14	0	4	0	4	17	10	7	1	0	10	34	72	
39	ARUNACHAL PRADESI	SUBANSIRI LOWEF	2001	8	10	1	2	0	2	6	5	1	1	0	7	17	25	
40	ARUNACHAL PRADESI	SUBANSIRI UPPER	2001	4	0	0	2	0	2	1	1	0	3	0	5	17	20	
41	ARUNACHAL PRADESI	TAWANG	2001	2	0	0	1	0	1	0	0	0	0	0	0	3	13	
42	ARUNACHAL PRADESI	TIRAP	2001	8	4	1	1	0	1	6	3	3	0	0	2	18	17	
43	ASSAM	BARPETA	2001	64	5	0	28	0	28	105	88	17	43	1	45	93	187	
44	ASSAM	BONGAIGAON	2001	45	22	0	20	0	20	36	21	15	13	0	25	87	92	
45	ASSAM	C.I.D.	2001	0	0	0	0	0	0	0	0	0	0	0	0	0	1	
46	ASSAM	CACHAR	2001	52	30	1	45	0	45	104	74	30	29	0	42	189	359	
47	ASSAM	DARRANG	2001	61	13	0	48	0	48	64	47	17	21	0	29	111	185	

Fig 1: Dataset used for Crime analysis

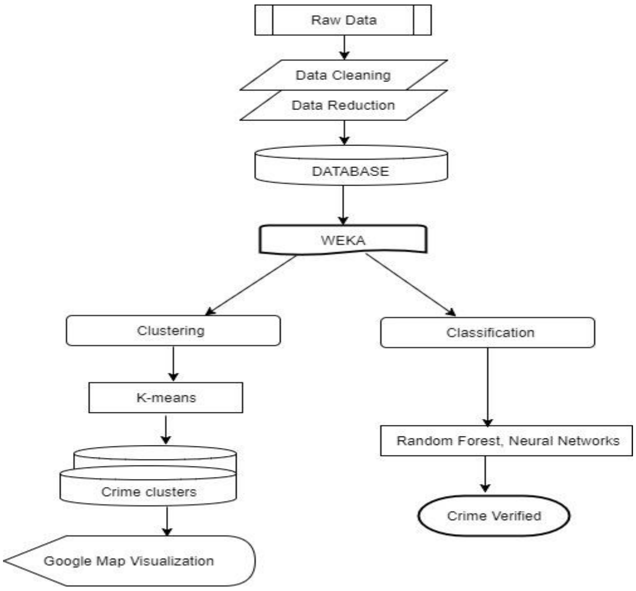


Fig 2: Work Flow of Proposed model

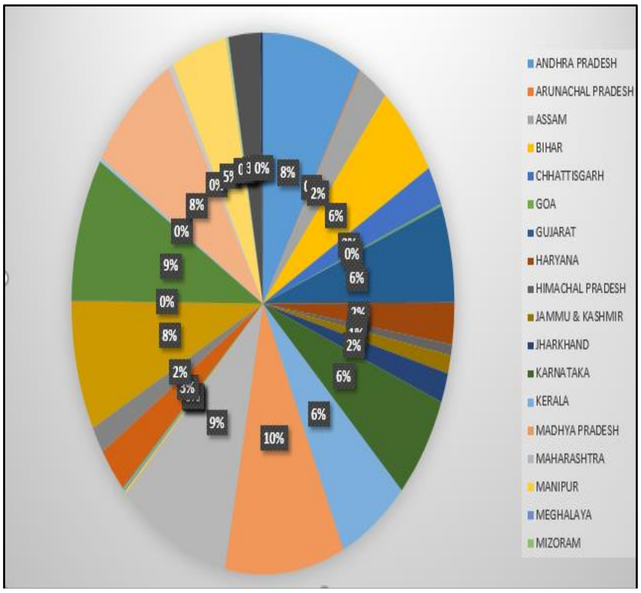


Fig 3: Crime rates in % of Indian states and union territories

4. Experimentation and results

Since crime is increasing at an alarming rate globally it is important to control it. In order to reduce crime rates we need to study the crime rates of various places of a country. In this research all the states and union territories crime rate (Fig 3) is studied in detail for different types of crimes. Since unsupervised and supervised learning techniques are used it helps in improving the filtration of large crime databases. Thus, by following the proposed approach the crime rate can be reduced in time and effort. Hence, this section discusses in detail about the experimentation and the results that are obtained using WEKA tool.

4.1 Selection of Indian states

In this approach the dataset consists of 28 states and 7 union territories with the corresponding districts. The Indian states and union territories chosen for the clustering are Andhra Pradesh, Bihar, Gujarat, Karnataka, Kerala, Madhya Pradesh, Rajasthan, Tamil Nadu, Uttar Pradesh and Delhi UT. The average IPC crimes for each of the states (based on the districts) during 2001 – 2012 are analysed with a line graph. The results are generated using the attributes “States/UT”, versus attribute “average_IPC_crime” (Fig 4).

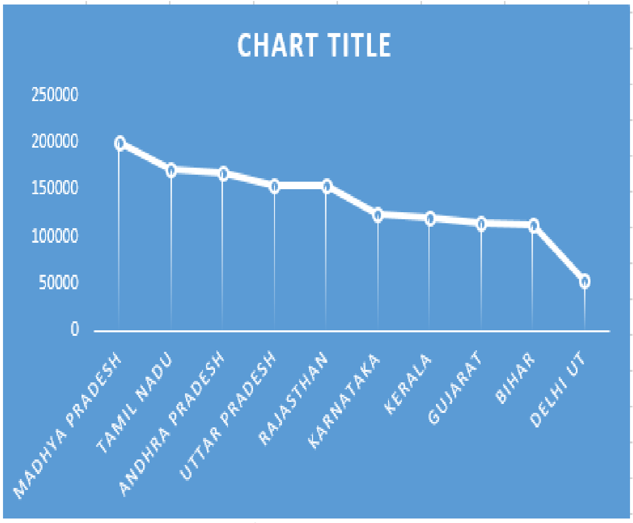


Fig 4: Line graph of Indian states versus the IPC crime

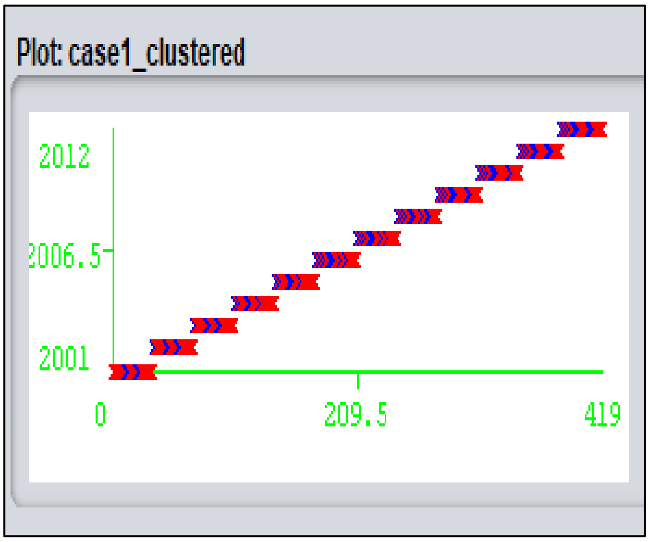


Fig 5: Cluster for Case 1

4.2 K-means implementation

In this model the existing patterns and the relations are searched in the dataset by using k-means and Google Map Marker clustering. This technique helps in providing an overview of the dataset and hence, helps in searching, handling and retrieving of the required or desired information. From the dataset 10 states are selected for formulating the clusters and the selection is done on the basis of the average IPC value of each state during 2001-2012. Since the locations are marked on the map it helps in analysing the states. The information is quite helpful for investigating agencies and the police officials. Clustering is achieved using the WEKA tool for the cases 1 to 4 - crime in Indian states, the attributes used for the k-means are “State”, “year” and “average_IPC_crimes” and “total_IPC_crime”.

Case 1 crime detection in India during 2001 to 2012: K-means helps in grouping objects (crimes in Indian states during 2001 – 2012) into clusters and here we are denoting each object using A, B, C, … L. Clusters are formed using the two crime attributes “year” and “total_IPC_crime” (Fig 5). The number of clusters by default is 2. Totally 9 iterations are performed on the dataset for case 1.

Case 2 crime detection in Uttar Pradesh and Delhi during 2001 to 2012: To analyse the number of crimes during 2001 to 2012 in Uttar Pradesh and Delhi clusters are generated. The attributes “year” and “total_IPC_crime” are used to generate the clusters and it is independent of the crime type. The clusters are generated in the same way for other states and union territories.

Case 3 crime detection on Murder and Rape in India during 2001 to 2012: Here clusters are created to find the number of crimes of a particular type (rape or murder) during the 12 years (2001 to 2012) in various states and union territories of India. The attributes that are considered are “year”, “crime_type” and the attribute “state” is not considered for this case. Similarly, the clusters are generated for other crime types in India.

Case 4 crime identification / detection of type rape in Uttar Pradesh during 2001 to 2012: The clusters are created for this case using the attributes “crime_type” and “district”. This helps in knowing which region has the highest crime rate in a state (Uttar Pradesh). In the same way clusters are generated for other states and union territories.

4.3 Google map marker clustering

The k-means results are enhanced using the Google map marker clustering (GMAPI). Google map marker clustering helps in representing the most crime prone regions in India. The attributes chosen for this are “state”, “location_longitude”, “location latitude” and “average_IPC_crime”. The Google map marker clustering uses latitude and longitude of locations to plot the crime prone areas. For instance, “location_longitude” =28.64 N, “location_latitude” =77.22 E for state = “Delhi”.

The crime rates in a specific year can be located directly using the marker cluster. The states are ranked based on their crime rate, the state or union territory with the highest crime rate (average_IPC_crime) is ranked 1 and the state or union territory with lowest crime is numbered 35. The region with highest crime rate is shown in the darkest colour and the region with the lowest crime rate in shown in lighter colour as in Fig 6. Google marker clustering helps in identifying the hotspots of various crimes. Since the affected areas are spotted it is easy to solve crimes and predict if that area is prone to more or less crimes in near future.

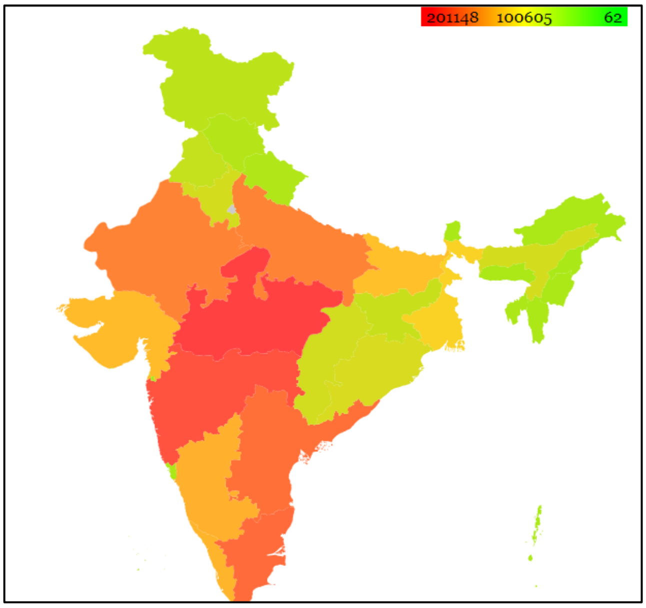
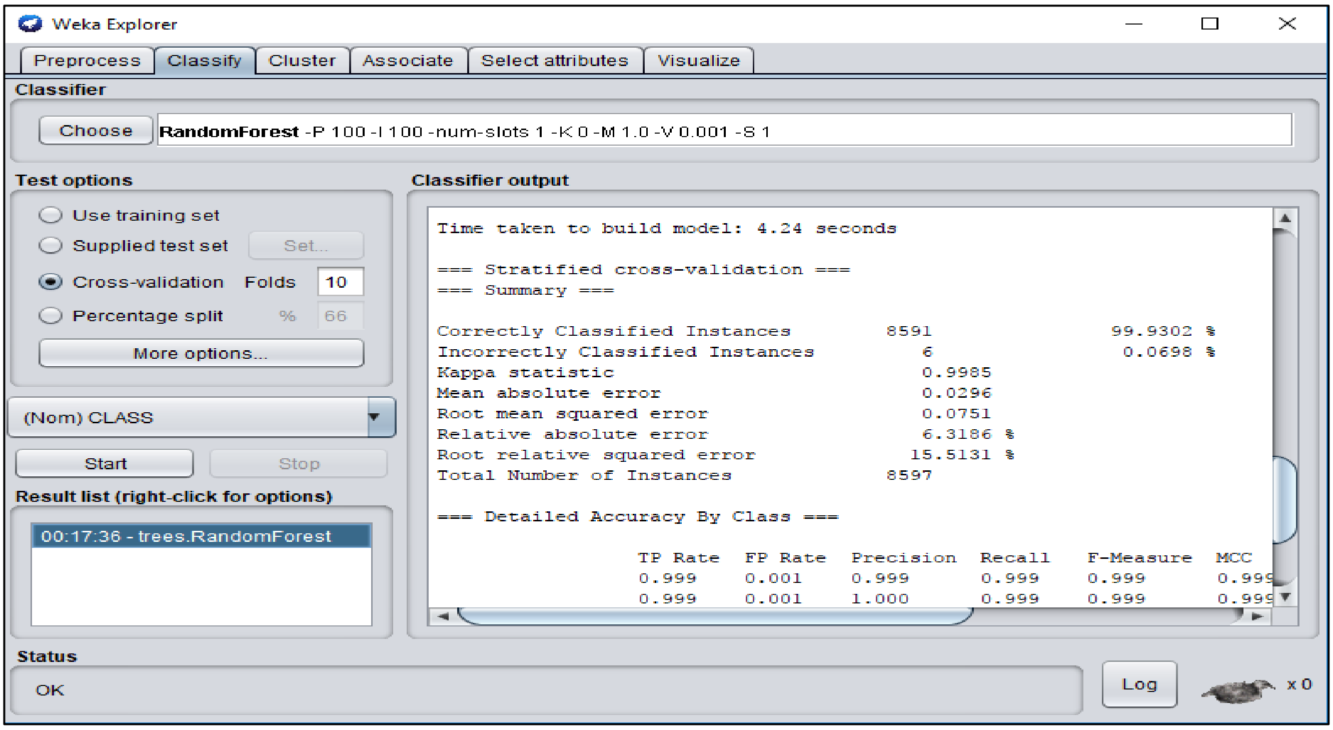


Fig 6: Density of crimes in various states and union territories

The Google map marker clustering (GMAPI) produced the following results (Fig 7) on our dataset. The crime prone areas are marked with a number based on the intensity of crime in that area.



Classifier

RandomForest -P 100 -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1

Test options

○ Use training set
○ Supplied test set
● Cross-validation Folds 10
○ Percentage split % 66

(Nom) CLASS

Result list (right-click for options)

00:17:36 - trees.RandomForest

Classifier output

```
Time taken to build model: 4.24 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        8591               99.9302 %
Incorrectly Classified Instances         6                0.0698 %
Kappa statistic                          0.9985
Mean absolute error                      0.0296
Root mean squared error                  0.0751
Relative absolute error                  6.3186 %
Root relative squared error             15.5131 %
Total Number of Instances             8597

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC
                 0.999    0.001    0.999      0.999    0.999      0.999
                 0.999    0.001    1.000      0.999    0.999      0.999
```

Status

OK

Fig 8: WEKA tool

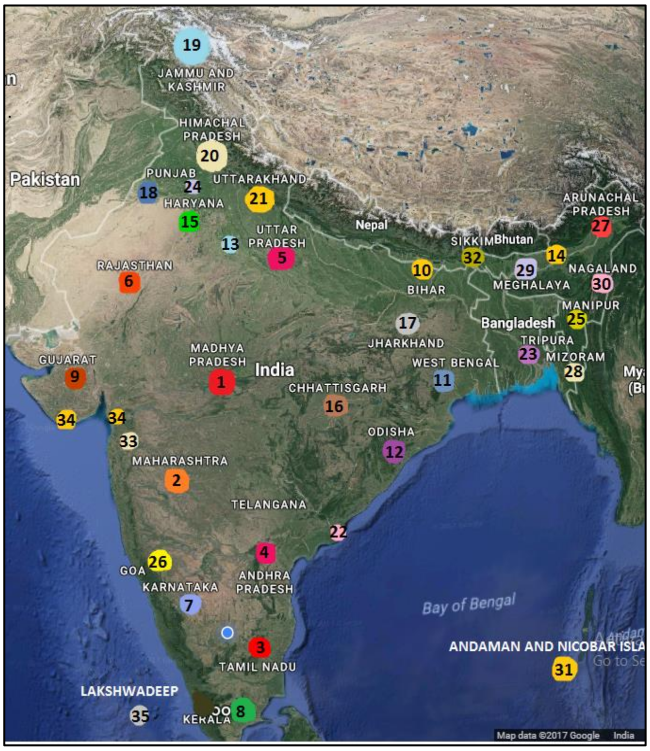


Fig 7: Google map marker cluster

4.4 Crime analysis using WEKA

WEKA is a useful open source free tool. It is used for analysing the real – world datasets. WEKA tool helps in verifying the data mining algorithms. The results of k-means are verified with WEKA. The Random Forest algorithm and the Neural Networks (classification) selected are applied on the dataset and the accuracy is verified. The number of correctly classified instances helps in knowing how many instances are classified correctly and the number of incorrectly classified instances can be reduced by removing the attributes which affect the accuracy. The Random Forest is applied on the dataset and the test option selected is Cross- validation (Folds 10). The number of correctly classified instances are 8591 and wrongly classified instances are 6. The classifier verifies an accuracy of 99.9302 %. The Neural networks algorithm verifies an accuracy of 90.02%.

5.Conclusion

The crime rates in India is increasing day by day due to many factors such as increase in poverty, unemployment, corruption, etc. The 10 Indian states or union territories selected are chosen on the basis of their crime rate. This approach is very useful in studying if the crime rate is increasing or decreasing in a particular region. If the crime has increased necessary measures can be taken by the officials to study why the crime has increased and also how to reduce the crime rate in that region. In this research the crime rates during 2001 to 2012 are analysed and this has helped in ranking the states and union territories on the basis of their average IPC crime rate. Using WEKA, the accuracy of the proposed model is measured and verified. A good accuracy of 99.93 % is obtained and this verifies the correctness of the instances.

The proposed model is very useful for both the investigating agencies and the police officials in taking necessary steps to reduce crime. The model can be applied to any countries dataset. By spotting the crime prone areas the general public can be given an alert about the crimes in different parts of a country.

Future enhancement of this research work focuses on training bots to predict the crime prone areas by using machine learning techniques. Since, machine learning is similar to data mining advanced concepts of machine learning can be used for better prediction. The data privacy, reliability, accuracy can be improved for enhanced prediction.

References

[1] G. Jiji-S. Anantharadha, "Automatic Tracking of Criminals using Data Mining Techniques", *Journal of The Institution of Engineers (India)*: Series B, 2012, https://doi.org/10.1007/s40031-013-0036-1

[2] Devendra Tayal, Arti Jain, Surbhi Arora, Surbhi Agarwal, Tushar Gupta, Nikhil Tyagi, "Crime detection and Criminal identification in India using data mining technique", *Ai & Society*, 2014, https://doi.org/10.1007/s00146-014-0539-6

[3] Shiju Sathyadevan, Devan S, Surya S, "Crime analysis and prediction using data mining", *First International Conference on Networks & Soft Computing (ICNSC2014)*, 2014, DOI: 10.1109/CNSC.2014.6906719

[4] Shyam Nath, "Crime Pattern Detection Using Data Mining", *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology Workshops*, 2006, DOI: 10.1109/WI-IATW.2006.55

[5] Somayeh Shojaee, Aida Mustafa, Fatimah Sidi, Marzanah Jabar, "A Study on Classification Learning Algorithms to Predict Crime Status", *International Journal of Digital Content Technology and its Applications(JDCTA)*, Volume 7, Number 9, 1-3, 2013, DOI: 10.4156/jdcta.vol7.issue9.43.

[6] Dawei Wang, Wei Ding, Henry Lo, Tomasz Stepinski, Josue Salazar, Melissa Morabito, "Crime hotspot mapping using the crime related factors—a spatial data mining approach", *Applied Intelligence*, 2012, https://doi.org/10.1007/s10489-012-0400-x

[7] Chung-Hsien Yu, Max Ward, Melissa Morabito, Wei Ding, "Crime Forecasting Using Data Mining Techniques", *2011 IEEE 11th International Conference on Data Mining Workshops*, 2011, DOI: 10.1109/ICDMW.2011.56

[8] Arunima Kumar, Raju Gopal, "Data mining based crime investigation systems: Taxonomy and relevance", *2015 Global Conference on Communication Technologies (GCCT)* – 2015, DOI: 10.1109/GCCT.2015.7342782

[9] Prajakta Yerpude and Vaishnavi Gudur, "Predictive Modelling of Crime Dataset Using Data Mining", *International Journal of Data Mining & Knowledge Management Process (IJDKP)*, Vol.7, No.4, July 2017, DOI: 10.5121/ijdkp.2017.7404

[10] Mohammad Keyvanpour, Mostafa Javideh, Mohammad Ebrahimi, "Detecting and investigating crime by means of data mining: a general crime matching framework", *Procedia Computer Science*,2011, https://doi.org/10.1016/j.procs.2010.12.143

[11] Ubon Thongsatapornwatana, "A survey of data mining techniques for analyzing crime patterns", *2016 Second Asian Conference on Defence Technology (ACDT)*, 2016, DOI:

[12] 10.1109/ACDT.2016.7437655