Application of Classification Techniques for Prediction and Analysis of Crime in India

Priyanka Das and Asit Kumar Das

Abstract Due to dramatic increase of crime rate, human skills for accessing the massive volume of data is about to diminish. So application of several data mining techniques can be beneficial for achieving insights on the crime patterns which will help the law enforcement prevent the crime with proper crime prevention strategies. This present work collects crime records for kidnapping, murder, rape and dowry death and analyses the crime trend in Indian states and union territories by applying various classification techniques. Analysing the crime would be much easier by the prediction rates shown in this work, and the effectiveness of these techniques is evaluated by accuracy, precision, recall and F-measure. This work also describes a comparative study for different classification algorithms used.

Keywords Crime prediction · Classification · Naïve Bayes · Random Forest Precision · Recall

1 Introduction

Crime is a social nuisance which has been on rise in almost all parts of the world including India. Criminologists analyse the data with varying degrees of success. But with the increasing crime rate, human skills tend to fail when they are provided with huge volume of data sets. Application of data mining techniques can be used to facilitate the task that can extract the hidden knowledge from the massive data sets and provide the crime investigation department a new edge for crime analysis. Collecting crime information from government portals employs data mining

P. Das (✉) · A. K. Das
Department of Computer Science and Technology,
Indian Institute of Engineering Science and Technology,
Shibpur, Howrah 711103, West Bengal, India
e-mail: priyankadas700@gmail.com

A. K. Das
e-mail: akdas@cs.iiests.ac.in

H. S. Behera et al. (eds.), *Computational Intelligence in Data Mining*,
Advances in Intelligent Systems and Computing 711,
https://doi.org/10.1007/978-981-10-8055-5_18

techniques that predicts the future crime trends. Past crime records accumulated from government portals constitute the crime type, time, location, information about the victims, their genders, ages, social status and many more. Thus, crime prediction, a subtask of crime analysis, considers all the past crime records, classifies the crime categories and predicts the future crime. Numerous research works exist in the literature that employs different data mining techniques for crime analysis of different countries and cities. Crime prediction using pattern and association rule mining determines the chances of performing crime by the same criminal [1]. Given a time and place, type of crime occurring in San Fransisco City is predicted in [2]. Again, association rule mining task was incorporated for detecting crime locations in Denver and Los Angeles in [3]. Analysis of crime has been done by mapping, and similarities have been found with the past crime trend when compared to present scenario [4]. This task was an approach to determine places where maximum numbers of crime incidents take place. Application of data mining techniques has proven to be crucial for crime detection and prevention task. A comparative study was conducted for different crime patterns which exhibited better results for linear regression than other classification methods [5]. An architecture was implemented that collects the raw data and categorises the data into crime types, locations and places. Then, existing classification algorithms were used, and the most effective technique was chosen resulting in crime prediction [6]. Among all the available classification techniques, particularly two methods such as Naïve Bayes and backpropagation algorithm were compared for predicting the crime type in distinct states of America. This experiment shows that Bayes classification technique provides better accuracy than the backpropagation method [7]. Apart from the classification techniques, a clustering-based model [8] is used for anticipating the crimes that may help the law enforcement agencies preventing the crimes at a faster pace. An interdisciplinary approach was introduced in [9] that incorporated the knowledge of computer science and criminology to prepare a crime prediction model that focuses on crime factor for each day. Though most of the above-mentioned works have been done on crime data set from USA, none have done an extensive crime analysis for Indian states and in its union territories. The present work demonstrates crime prediction for 28 states (Andhra Pradesh inclusive of Telengana) and 7 union territories of India. It has considered the collection of crime records from 2001–2014 containing information about four different types of crime like kidnapping or abduction, murder, rape and dowry death. The data set for kidnapping comprises information about the number of victims, their gender, age, whereas the data for dowry death contains records of the number of cases pending investigation, cases discharged, cases claimed false, etc. Data set for rape holds information about the victims and types of rape occurred, and for murder cases, a different data set provides list of male and female victims with their ages. The present work has considered all the data sets for 2001–2012 as training data as input to several learning techniques like KNN, Random Forest, Naïve Bayes, AdaBoost and Classification Tree, and the predictive model has been learnt. Then, this predictive model has been used to predict the future crime trends. For kidnapping cases, the purpose of kidnapping has been chosen as the class label, whereas for rape cases, the

types of rape (incest or others), the gender (male/female) for murdered victim and for dowry death, three different aspects have been chosen for prediction.

The rest of the paper is organised as follows: Sect. 2 describes the proposed work in detail. Section 3 shows the results of the proposed method followed by conclusion and future work in Sect. 4.

2 Proposed Framework

This section describes the present work elaborately in the following subsections:

2.1 Crime Data Collection and Preparation

Input data play a crucial role in the field of crime data mining. The crime data for the present work has been collected from National Crime Records Bureau (NCRB) and Open Government Data Platform India which provide documents and applications for research purpose as well as for public use. Collected data contain information about 28 states and 7 union territories of India. The collected data did not contain any missing values so as necessary preprocessing, the raw data have been converted to computer-readable format for further analysis. Most frequent crimes like kidnapping, murder, rape and dowry death have been chosen to deal with. For all the states, each data set for each crime type contains several attributes like name of the state, year, number of victims, number of cases. Though there exist significant methods for attribute selection, the present work has acknowledged the most probable attributes for crime types depending on human perception. Table 1 shows all the details of the integrated attributes from the data sets for the present task.

2.2 Classification Techniques

Classification techniques involve assignment of any object to one of the multiple predefined classes. Here, the predictive modelling is separately used for each crime type for all the states. Five different classification techniques have been used in this work, and they are briefly introduced as follows:

Decision Tree: It is simple yet widely used classifier with three types of nodes. The root and other nodes hold the test conditions for the features, and each leaf node is assigned with a class label.

K-Nearest Neighbour: It thus is an instance based learning where K defines the number of nearest neighbours and a proximity measure is needed for determining the similarity between the instances.

Table 1 Details of the collected data for kidnapping, murder, rape and dowry

Attribute	Description
State/UT	It refers to the names of the states or union territories
Year	Year of crime
Gender	Refers to male or female
Age	It illustrates age-wise male/female victims of crime
Purpose	Demonstrates nine different reasons of kidnapping like begging, adoption, prostitution
Type	It describes types of rapes like incest or other cases occurring in the states
CPIP	Cases pending investigation from previous year
CRY	Cases reported during the year
CWG	Cases withdrawn by the Govt. during investigation
CNI	Cases not investigated or in which investigation was refused
CDF	Cases declared false on account of mistake of fact or of law
CC	Cases in which chargesheets were laid
CCN	Cases in which chargesheets were not laid but final report submitted during the year
CPIY	Cases pending investigation at the end of the year
CPTP	Cases pending trial from the previous year
CST	Cases sent for trial during the year
TCT	Total number of cases for trial
CW	Cases withdrawn
CTC	Cases in which trials were completed
CCon	Cases convicted
CD	Cases acquitted or discharged
CPTY	Cases pending trial at the end of the year

Naïve Bayes: Provided a class label, this classifier assumes the features to be independent and determines their class conditional probability. Let the class label be y, then the conditional independence can be defined as (1).

$$P(x|y) = \sum_{i=1}^{k} P(x_i|y) \tag{1}$$

where feature set $x = (x_1, x_2, \ldots, x_k)$ and k is the number of features.

Random Forest: Random Forest comprises several classification trees, where each tree classifies an object and a voting is done for that particular class. Now, the classification with highest number of votes is selected by the forest. Random Forest is very efficient when dealing with large data sets, and depending on the applications, it often provides better accuracy than other classification techniques.

AdaBoost: Here, adaptive boosting algorithm is used with other classification techniques for achieving optimal performance on the crime data sets.

2.3 Crime Prediction Task

Once the data are collected, they have been divided into two parts, namely training and test data. Data for 2001–2012 have been chosen as training data with all attributes having known class labels. They have been given as input to the learning techniques, and the result of the trained predicting model is applied on the test data with unknown class labels. Data sets for 2013 and 2014 have been used as test data in this task, and trained model classifies the instances present in the test set. Figure 1 shows the basic layout of the crime prediction task, and Table 2 shows that though reason for abduction is the sole task in case of kidnapping, the present work predicts several extreme cases for dowry deaths. It emphasises on the issues regarding the cases convicted, cases that were claimed false and cases pending investigation. Thus, various issues regarding the crime get discovered and crime prediction is done for Indian states. This crime prediction task helps in analysing the future crime trends for crimes like rape, kidnap, murder and dowry deaths.

Table 2 Details of the predicted classes

Crime type	Prediction
Kidnapping	Purpose of kidnapping (adoption, begging, camel racing, etc.)
Rape	Types of rape (incest or others)
Murder	Gender of the victim
Dowry death	Cases convicted, cases pending investigation from previous years, cases declared false on account of mistake of fact or of law

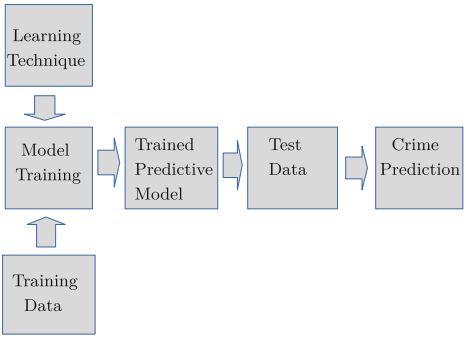


Fig. 1 Layout for crime prediction task

3 Experimental Results

Once the predictive model has been prepared based on the crime data for 2001–2012, the testing and scoring are done based on tenfold cross-validation on test data of 2013–2014. The performance of the classifiers has been measured by evaluation techniques, namely precision, recall and F-measure using (2–4). Confusion matrix denotes the number of classified and misclassified instances in the crime data. True positive (TP) and True negative (TN) represent correctly classified instances, whereas false positive (FP) and false negative (FN) denote the incorrectly classified instances.

$$Precision\,(P) = \frac{TP}{TP + FP} \tag{2}$$

$$Recall\,(R) = \frac{TP}{TP + FN} \tag{3}$$

$$F - measure = \frac{2PR}{P + R} \tag{4}$$

Tables 3 and 4 show the performance for each classifier with correctly classified and misclassified instances for the year 2013–2014. The present method has predicted the purpose of kidnapping; i.e., it classifies the instances as 'adoption', 'begging', 'camel racing', 'prostitution', 'marriage', 'illicit intercourse', 'slavery'. The training data for crime related to murder contain all the instances, and it has been learnt to the predictive model which classifies the gender of the murdered victims. Table 5 shows the performance measures for murder cases. Likewise, Table 6 shows the classification accuracy for predicting the types of rapes such as incest rape cases or others from the separate data set of rape victims. Now for dowry death cases, the present work has focused on predicting three different issues regarding the cases, investigation, etc. It predicts from the instances if the case is pending from the previous year, case convicted and case that is declared false on account of mistake of fact or law. The data set for dowry death cases contains many more attributes but as the motive is to predicting and analysing the crime trend, the present work has focused on the most important attributes. Tables 7, 8 and 9 show the results for different prediction accuracies in dowry death cases.

It has been observed from the results that Random Forest classifier provides the best accuracy of 90% and 95.6% for the year 2013 and 2014, respectively, for predicting the purpose of kidnapping and 95.2% for classifying the gender of murdered victims, whereas Naïve Bayes provides the highest F-measure of 85.9% for predicting the types of rape that occurs mostly in the Indian states. Decision tree based classifier provides the best F-measure of 86.1 for predicting the instances related to cases that have been declared false by law, and Random Forest, Naïve Bayes and AdaBoost all provide same F-measure of 98% for classifying the cases that have been convicted. Here, we have also shown the classification accuracy (CA) and area under the curve (AUC). Figure 2 shows receiver operating characteristic (ROC) curves for

Table 3 Performance measure in (%) for prediction in kidnapping 2013

Classifier	AUC	CA	P	R	F	Correctly classified	Misclassified
KNN	92.8	91.0	91.1	91.0	86.6	90.7	9.3
Tree	88.5	84.9	85.0	84.9	76.1	84.4	15.6
Random Forest	94.3	90.0	90.0	90.0	90.0	91.6	8.4
Naive Bayes	95.0	89.0	88.9	89.0	88.9	90.2	9.8
AdaBoost	90.0	90.4	90.4	90.4	86.1	91.4	8.6

Table 4 Performance measure in (%) for prediction in kidnapping 2014

Classifier	AUC	CA	P	R	F	Correctly classified	Misclassified
KNN	95.1	91.0	91.1	91.0	91.0	95.6	4.4
Tree	93.3	94.7	94.6	94.7	94.5	96.6	3.4
Random Forest	97.6	95.6	94.9	95.0	94.8	96.6	3.4
Naive Bayes	96.5	88.5	91.4	88.5	90.0	97.0	3.0
AdaBoost	82.4	94.7	94.6	94.7	94.5	96.6	3.4

Table 5 Performance measure in (%) for prediction in murder cases

Classifier	AUC	CA	P	R	F	Correctly classified	Misclassified
KNN	95.3	94.3	94.3	94.3	94.3	94.2	5.8
Tree	94.3	92.4	92.4	92.4	92.5	92.3	7.7
Random Forest	98.9	95.2	95.3	95.2	95.3	94.3	5.7
Naive Bayes	97.7	94.3	94.3	94.3	94.3	94.2	5.8
AdaBoost	94.7	93.3	93.2	93.5	93.3	90.9	9.1

Table 6 Performance measure in (%) for prediction in rape cases

Classifier	AUC	CA	P	R	F	Correctly classified	Misclassified
KNN	91.5	77.5	78.3	77.5	75.8	73.2	26.8
Tree	81.2	84.5	85.2	84.5	83.6	80.0	20.0
Random Forest	94.7	83.1	83.6	83.1	82.4	78.4	21.6
Naive Bayes	94.0	85.9	86.1	85.9	85.9	85.7	14.3
AdaBoost	78.3	84.5	84.1	84.8	84.1	81.6	18.4

Table 7 Performance measure in (%) for prediction of the dowry death cases that have been declared false

Classifier	AUC	CA	P	R	F	Correctly classified	Misclassified
KNN	85.0	83.3	84.2	83.3	84.1	89.5	10.5
Tree	82.5	86.1	87.5	86.1	86.4	94.4	5.6
Random Forest	82.5	83.3	84.1	83.3	81.2	89.5	10.5
Naive Bayes	90.0	83.3	84.1	83.3	81.2	89.5	10.5
AdaBoost	80.0	86.1	83.9	86.4	86.1	90.0	10.0

Table 8 Performance measure in (%) for prediction of the dowry death cases that have been convicted

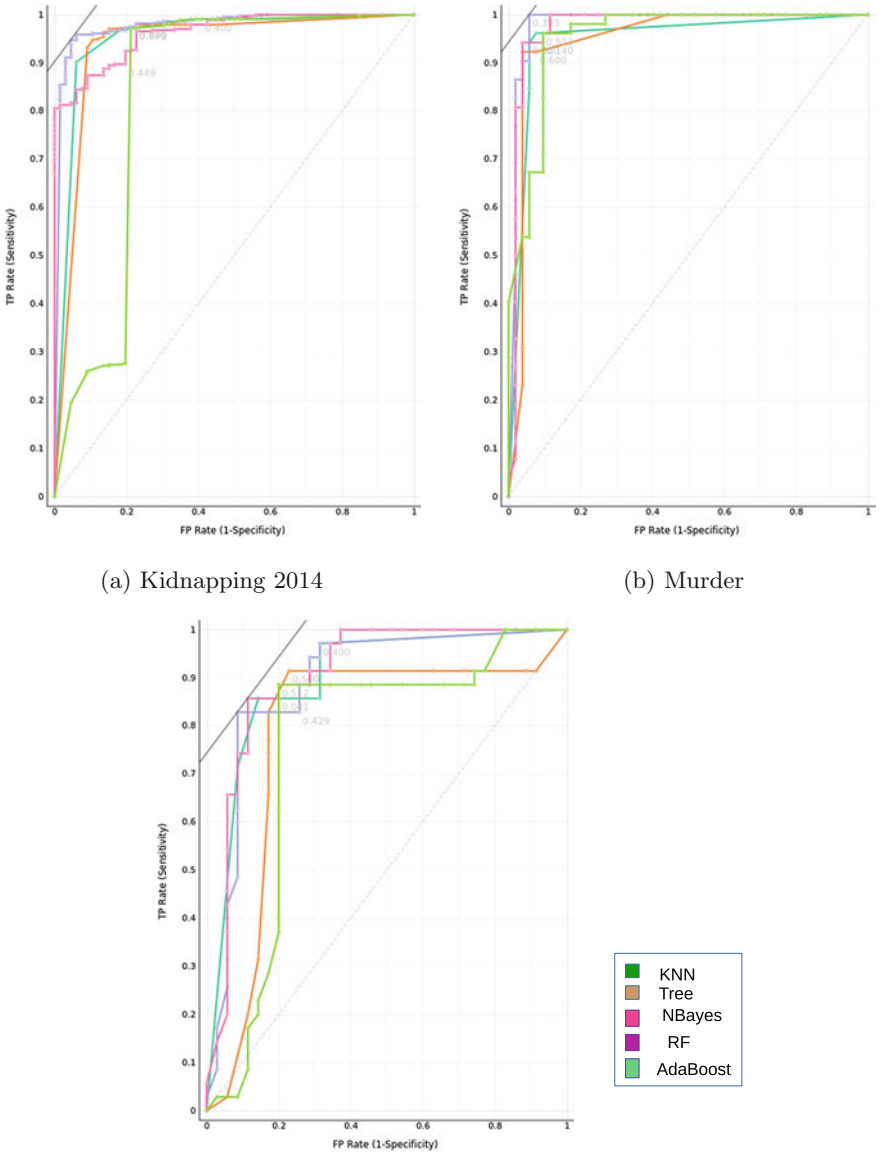
Classifier	AUC	CA	P	R	F	Correctly classified	Misclassified
KNN	98.0	97.0	97.0	97.0	98.0	98.0	2.0
Tree	97.0	96.0	97.0	97.0	97.0	97.9	2.1
Random Forest	98.0	96.0	98.0	98.0	98.0	98.0	2.0
Naive Bayes	98.0	97.0	98.0	98.0	98.0	97.0	3.0
AdaBoost	97.0	98.0	98.0	98.0	98.0	98.0	2.0

Table 9 Performance measure in (%) for prediction of the dowry death cases that are pending investigation from previous years

Classifier	AUC	CA	P	R	F	Correctly classified	Misclassified
KNN	91.2	88.9	88.9	88.9	88.9	88.9	11.1
Tree	92.5	91.7	91.4	91.8	91.7	89.5	10.5
Random Forest	97.5	91.7	91.4	91.8	91.7	89.5	10.5
Naive Bayes	97.5	88.9	88.9	88.9	88.9	88.9	11.1
AdaBoost	95.0	91.7	91.4	91.8	91.7	89.5	10.5

few crime instances. The ROC curve shows the trade-off between sensitivity and specificity, and most of the curves are to the top of the ROC space which denotes better classification accuracy for learning techniques like Random Forest and Naïve Bayes.

As an outcome of the research work, this study provides knowledge on how many people are getting victimised each year, and it reflects the motivation behind the crimes, as it predicts that several males are being kidnapped for slavery or any other unlawful activity, whereas most of the females are being kidnapped and sent to the middle east part of the globe for prostitution or most often their body parts are being sold. Most of the children are kidnapped for adoption and begging. For murder cases, most of the females of age 18–30 years are raped and murdered every year.



(a) Kidnapping 2014 (b) Murder

(c) Rape (d) index

Fig. 2 ROC curves for predicting different instances of crime

Regarding prediction in rape cases, it is observed that though number of rape cases has gone up till date, but more number of girls are being raped by their relatives (incest) and friends. The prediction in dowry death cases emphasises on three extreme issues that reflect how government takes action against the victims and their relatives. Though cases are being filed everyday, timely actions are not taken resulting into huge number of cases that are pending both investigation and trial. Very few cases have been convicted through all these years 2001–2014. Due to lack of facts and evidences, some cases have been declared false by the law. Thus, the present crime analysis is an effective approach so that law enforcement section can consider it and take preventive measures to reduce the crime rate.

4 Conclusion and Future Work

Application of several data mining techniques can be beneficial for achieving insights on the crime patterns which will help the law enforcement prevent the crime with proper crime prevention strategies. The present work has used a Python-based software called 'Orange' [10] for the learning techniques and demonstrates a simple yet effective approach for crime prediction from Indian crime data. It utilises few existing learning algorithms that provides an insight of the attributes present in the data set and later helps in predicting the crime trend in India for the year 2001–2014. The results show high prediction accuracy for most of the cases. The present work also demonstrates a comparative study of the classification algorithms used for crime analysis. Not only it analyses the crime trend, it also reflects on the victims, their ages, number of people getting victimised every year, and most importantly it reflects on the actions taken by the government dealing with dowry death cases. Though the present work reflects the crime scenario of Indian states, this method can also be applied for analysing the global crime scenario. As a future work, other crime types can be considered and other classification methods can also be employed for analysing the crime in an extensive manner.

References

1. Anisha Agarwal, Dhanashree Chougule, A.A.D.C.: Application for analysis and prediction of crime data using data mining. In: Proceedings of IRF-ieeeforum International Conference. (2016) 35–38
2. Chandrasekar, A., Raj, A.S., Kumar, P.: Crime prediction and classification in San Francisco city
3. Subhash Tatale, N.B.: Crime prediction based on crime types and using spatial and temporal criminal hotspots. International Journal of Data Mining & Knowledge Management Process **5**(4) (2015) 1–19
4. Subhash Tatale, N.B.: Criminal data analysis in a crime investigation system using data mining. Journal of Data Mining and Management **1**(1) (2016) 1–13

5. Lawrence McClendon, N.M.: Using machine learning algorithms to analyze crime data. Machine Learning and Applications: An International Journal (MLAIJ) **2**(1) (2015) 1–12
6. Yu, C.H., Ding, W., Chen, P., Morabito, M.: Crime forecasting using spatio-temporal pattern with ensemble learning. In: Pacific-Asia Conference on Knowledge Discovery and Data Mining. (2014) 174–185
7. Abba Babakura, Md Nasir Sulaiman, M.A.Y.: Improved method of classification algorithms for crime prediction. In: Proceedings of International Symposium on Biometrics and Security Technologies (ISBAST). (2014) 250–255
8. S. Yamuna, N.B. Chang: Datamining techniques to analyze and predict crimes. The International Journal of Engineering And Science (IJES) **1**(2) (2012) 243–247
9. Sathyadevan, S., Devan, M.S., Surya Gangadharan, S.: Crime analysis and prediction using data mining. In: 2014 First International Conference on Networks Soft Computing (ICNSC2014). (Aug 2014) 406–412
10. Demšar, J., Curk, T., Erjavec, A., Črt Gorup, Hočevar, T., Milutinovič, M., Možina, M., Polajnar, M., Toplak, M., Starič, A., Štajdohar, M., Umek, L., Žagar, L., Žbontar, J., Žitnik, M., Zupan, B.: Orange: Data mining toolbox in python. Journal of Machine Learning Research **14** (2013) 2349–2353