International Journal of Current Engineering and Technology
E-ISSN 2277 – 4106, P-ISSN 2347 – 5161

Available at http://inpressco.com/category/ijcet

Research Article

Crime Data Analysis and Prediction using Machine Learning

Mansi .S. Bagale and Dr (Mrs.) S .K, Wagh

Department of Computer Engineering Modern Education Society's College of Engineering, Pune, India

Received 10 Nov 2020, Accepted 10 Dec 2020, Available online 01 Feb 2021, **Special Issue-8 (Feb 2021)**

Abstract

An act done by any person which is against the laws of the particular country or region is called crime. A person who does this is called a criminal. Almost without exception, people in India think that the crimes are increasing at all time high levels. Crime may be considered as one of the following types for e.g. theft of motor vehicle, robbery of money, murder, rape, assault, , false imprisonment, kidnapping, Robbery of business property, Card fraud, Homicide, Domestic burglary, Robbery on public transportation, Rape, Firearm injuries, Robbery in subway etc. In India the police maintain a criminal record which is accessible to the masses. The basic idea of what things are called "crimes" is that they are thought to be things that might cause a problem for another person. Data mining is a process that extracts useful information from huge amount of crime related data stored in databases. The crime data is extracted from the official portal of National Crime Records Bureau (NCRB) of India. Before training of the model data pre-processing will be done following this feature selection and scaling will be done so that accuracy obtain will be high. The various classification and other algorithms will be tested for crime prediction and one with better accuracy will be used for training. Visualization of dataset is done in terms of graphical representation of many cases for example at which area criminal activities are high. The main purpose of this project is to give a jest idea of how machine learning can be used by the law enforcement agencies to predict, solve crimes and identify patterns of crime at a much faster rate in order to reduce the crime rate.

Keywords: *Classification, crime, patterns, visualization etc.*

Introduction

priority by the government. The improvement in computer technology has lead to development of emergent techniques that has revolutionized the functioning of classical and old style in most of public and private sector. Criminology is an area that focuses the scientific study of crime and criminal behavior and aims to identify crime characteristics. It is one of the most important fields where the application of machine learning techniques can produce important results. Crime analysis, is a part of criminology, a task that includes exploring and detecting crimes. The main purpose of crime analysis is to provide the data needed to investigator or police from huge amount of information stored in the department to guide them in right direction in order to prevent crime and control the criminal activities which may occur in future. Crime is committed by people of all age groups. Earlier only males were reported committed crimes whereas now reports prove that both men and women commit crimes. Problem is of identifying techniques that can accurately and efficiently analyze the growing volumes of crime data. Also the data available is inconsistent and are incomplete thus making the task of formal analysis a far more difficult. There is a change in the trends of crime and it is very difficult to find new trends and patterns in crime Hence it is very essential to find techniques to reduce crime and enable the police officials to easily catch the criminals.

Literature Survey

A. Literature analysis

In this proposed work different clustering algorithm such as k-mean clustering ,agglomerative clustering are used for analyzing the location of the crime in order to reduce city crime rates. Various visualization techniques are used to create visual images which aid in the understanding of complex, often massive representation of data[1]. The objective of this proposed work is to analyse dataset which consist of numerous crime and predicting the type of crime which may happen in future depending on various condition. The crime data set is taken from official portal of Chicago police. K-NN and Various other algorithms will be tested for crime prediction and one with better accuracy will be used for training .The whole aspect of this paper is to give jest idea of how machine learning can be used by the law enforcement agencies to detect ,predict and solve crimes at a much

faster rate and thus reduce the crime rate[2]. The NCRB(National crime records bureau) collects and maintains criminal data and publish the crime data. In this k-mean clustering algorithm is used on criminal dataset. WEKA a Software is used to construct cluster zones. It builds model with high, low, medium crime zone. zones of state. This information can be helpful to police to increase or decrease level of preventive actions[3]. In this proposed work data is collected from government sources in csv format. This data is preprocessed in R. The technologies used for mining various crime pattern and analysis are WEKA tool and R tool. The output is represented in graphical form such as charts indicating high or low crime region[4]. Crime investigation [5] is a range of essential significance in police division. Investigation of wrong doing information can help us examine wrong doing design, between related clues and critical cases relations between the violations. That is the reason information mining can be extraordinary guide to breaking down, envision and foresee wrong doing utilizing informational collection. Dataset is grouped on the premise of some predefined condition. Here gathering is done by different sorts of wrongdoings against ladies occurring in various states and urban areas of India. Wrong doing mapping will help the organization to arrange methodologies for the counteractive action of wrongdoing, further utilizing information mining strategy information can be anticipated and imagined in a different frame with a specific end goal to give better comprehension of wrongdoing examples. [6] In this proposed work data is collected from Chicago Portal from 2010 to 2012 i.e. 2 years related to different crime that has been committed in different region of Chicago city. Two approaches has been used. First one is clustering using kmean, to identify different places of crime for which WEKA tool is used. Second one is Spatial mining to locate hot spot of crime. Hot spots detection will help to detect, locate and .solve crimes at much faster rate. [7]The proposed model the dataset contains both categorical and numerical value collected from Chicago Police Department system in CSV format. The algorithm that is used to train the dataset are Random Forest, Decision Tree and different ensemble methods such as AdaBoost ,bagging and Extra trees. The main motivation of this model is to use algorithm on these dataset to classify the type of crime occurring based on time and location for better performance and highest accuracy

Proposed Methodology

A. Existing system

In existing system various techniques are proposed to analysis crime data. Using machine learning the extraction of new information is predicted using existing dataset. Many approaches for analysis of crime and prediction has been performed with help of WEKA tool, RapidMiner , R etc. But the existing work does not have the facility of feature engineering or exploratory data analysis. Specific crimes are analysed one by one to get better insights. EDA helps to discover insights which were not noticed or worth investigating but it can be very much informative for the crime investigation department.

B. Architecture

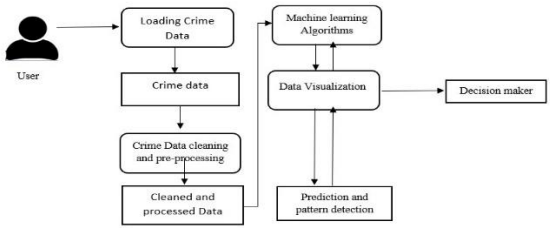

Fig. 1. System Architecture

The proposed architecture consist of the following stages: Loading Crime data(Dataset Collection): The user is the data analyst who is responsible for data collection. The aim is to perform crime data analysis and prediction using features present in the available dataset. The dataset is collected from National Crime Record Bureau(NCRB) official site of India. This dataset contains complete information about different aspects of crimes happened in India . The file format for the data is .CSV. File is loaded using Pandas library containing read.csv function in Python. The state wise data from 2001 to 2012 is considered. In the dataset different types of crimes(attributes) are considered like murdered, rape, kidnapping, dacoity, robbery, burglary, dowry deaths etc. There are 9017 instances in the dataset. The crime in each district is recorded from 2001 to 2012.The total IPC (Indian Penal Code) crime for each district is given in the dataset. Hence after loading the crime data ,it is available for data cleaning and pre-processing.

Data Pre-Processing: The next step in this model is data preprocessing which includes filling the missing values, data cleaning and transformation of data. The missing values are replaced by mean/mode value of the corresponding attribute instance.

Application of machine learning algorithms: Once the preprocessing is completed classification and clustering algorithms are applied based on requirements. The requirement can be anything from selecting the high or low crime prone area to finding the crime patterns in particular area based on previous crime records. The classification algorithm e.g. Naive Bayes works on supervised learning concept in which random sampling need to be carried out i.e. diving the data into test and train sample for e.g. 80% train samples and 20% test samples to train the classifier to identify the new unknown crime record. Whereas clustering algorithm e.g. k-mean is based on unsupervised learning algorithm which separates the

crime records based on the number of group to be created. Classification algorithms such as decision tree, Naïve Bayes, KNN, SVM is applied on the training dataset and implementation is done using python.

Visualization: The results can be visualized using appropriate graphs or maps showing sensitive areas of having high probability of crimes. Matplotlib library from sklearn is used for analysis of crime dataset. Exploratory Data Analysis is used to get the full understanding of the data and draw attention to its most important features.

Pattern identification: Next phase in the methodology is pattern identification that is used to find the sequence of crimes which are similar in nature and belongs to same class. Identification of meaningful patterns can help police to develop effective crime prevention and crime reduction strategies record. Whereas clustering algorithm e.g. k-mean is based on unsupervised learning algorithm which separates the crime records based on the number of group to be created.

Prediction: The use of frequent patterns is used to drive models that can predict future crime.

Decision maker: The analyst provide result of data analysis to crime investigation department. This information can help law enforcement agencies in enhancing intervention strategies via effective preparedness.

C. Selection of machine Algorithms i. Naïve Bayes

Naïve Bayes is a statistical classifier. It is capable of predicting class membership probabilities. Here a membership probability means the chances that a given tuple belongs to a particular class. Naïve Bayes classification is based on independence assumption. This classifier assumes that the presence or absence of particular attribute is independent of presence or absence of any other attribute. For example a person can be considered as a suspect if he is 6.5 feet tall, has round face and white hair. Naïve Bayes classifier works on each attribute independently and concludes that probably it is a suspect. Classification is performed using parameter estimation such as calculating mean, variance. However maximum likelihood method can be used for parameter estimation. Naïve Bayes classification is based on Bayes‟ theorem with naïve class conditional independence. Class conditional independence means the effect of an attribute value on a given class is independent of the values of other attributes.

The equation for Bayes‟ theorem is given as follows:

- Where, $P(A)$ - prior probability of A. It is "prior" in the sense that it is independent of any information about B.
- $P(A|B)$ - conditional probability of A, given B. It is also called the posterior probability because it is derived from or depends upon the specified value of B.
- $P(B|A)$ - conditional probability of B given A. It is also called the likelihood.
- $P(B)$ - constant.

ii. Decision Tree

The decision tree is a hierarchical data structure based on the concept of divide-and-conquer strategy. It is an productive nonparametric method, which can be used for both classification and regression. Decision trees classify the examples by arranging them down the tree from the root to some leaf node, with the leaf node providing the classification to the example. Each node in the tree acts as a test case for some features, and each edge descending from that node corresponds to one of the possible answers to the test case. The method is recursive in nature and is repeated for each and every subtree routed at the new nodes.

Entropy

Information theory is a measure to state the degree of disorganization in a system known as Entropy. If the sample is completely homogeneous, then the entropy calculated is zero and if the sample is an equally divided into 2 equal parts, it has entropy of one. Entropy is calculated using the formula:

$$\text{Entropy} = -p \log_2 p - q \log_2 q \quad (1)$$

Information Gain

Information class measures how much information a attribute gives about the class.

Attribute that perfectly partition should give maximal formation.

Unrelated Attributes should give no information.

iii. K-nearest neighbors (*KNN)*

One of the type of supervised machine learning algorithm is K-nearest neighbors (KNN) algorithm , which is used for both regression as well as classification problems. K-nearest neighbors (KNN) algorithm uses 'feature similarity' to predict the values of new item .It means that the new item is assigned a value based on measurement of similarity between data points in the training set. Its working consist of the following steps − 1. Load the training and test data

2. K value i.e. the nearest data point is chosen. It should be integer.

3. For each data point in the test data do the following steps:-

   3.1 The distance between test data and every row of training data is calculated using Euclidean or Manhattan or Hamming distance. Euclidean is the most frequently used method to calculate distance

3.2 Now, based on the distance value arrange them in ascending order,.

3.3 Next, from the sorted array it chooses the topmost K rows .

3.4 Lastly it allocate a class to a data point based on most often class assigned to these rows

4 End *iv.* Support Vector machine (SVM)

SVM model is basically a representation of different classes in a hyperplane in multidimensional space. The hyperplane will be generated in an iterative manner by SVM so that the error can be minimized. The aim of SVM is to split the dataset into classes to look out for marginal hyperplane which is
maximum

The followings are useful terms in SVM :-

• Support Vectors − These are datapoints that are nearest to the hyperplane, It is also called support vectors. These data points are used to define the Separating line. • Hyperplane − It is a plane or space which is partitioned between a set of objects belonging to different classes. • Margin − It is defined as the distance in between two lines on the closet data points of different classes. It is measured as the perpendicular distance from the line to the support vectors. Large margin is considered as a good margin and small margin is considered as a bad margin.

The main purpose of SVM is to split the datasets into classes to look out fora maximum marginal hyperplane (MMH) and it is done in the following two steps

• Initially separating lines i.e. hyperplanes are created repetitively by SVM, thus separating the classes in better way • After that the hyperplane which isolates the classes accurately is chosen

D. Mathematical model

Classification is a process that requires few predefined functions. These function choose the train data over sample data which is given by user .

Given a class label c. The conditional independence is considered as follows:

$P(S \setminus C= c) = \prod P(Si \setminus C = c)$

Where set S = {S1,S2,....,Sn} consists of n attribute .

Consider the following steps:- 1.Input dataset is Di total number of training data points is n.

2.Calculate the prior probability P(Cj) for each class Cj in Dataset Di : $P(Cj)= \frac{\sum ti \rightarrow cj}{\sum ti}$ $i=1$ n

3.Calculate the class conditional probabilities P(Sij\Cj) for each attribute values in dataset Di . $P(Sij\setminus Cj) = \frac{\sum Si \rightarrow Cj}{\sum ti \rightarrow Cj}$ $i=1$

4. Each training datapoint ti in training data D is required to be classified with maximum posterior probabilities. $P(ei \mid Cj) = p (Cj) \prod k = 1 \rightarrow p P(Sij \mid Cj)$

5. Iterate through steps 2 to 4 until all the training datapoints ti in D are classified correctly.

Analysis

For performing analysis of the models the performance metrics values need to be calculated. Performance metric calculation consists of few formulas which is used to achieve the performance values of dataset.
Some of the important formulas are:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

$$\text{Error rate} = 100 - \text{Accuracy}$$

$$\text{Precision} = \frac{TP}{TP+FP}$$

$$\text{Recall} = \frac{TP}{TP+FN}$$

$$\text{F-score} = \frac{2}{\text{Recall}^{-1} + \text{Precision}^{-1}}$$

Where TP=True positive, TN=True Negative, FP=False Positive, FN=False negative, N = No. of instances

Confusion matrix

It is used to describe the performance of classification model on set of test data for which true values are unknown.

	Predicted class (P)	Predicted class (N)
Actual Class (P)	TP	FN
Actual Class (N)	FP	TN

We compared the result of four models using decision tree, KNN, Naïve Bayes, SVM algorithms. By comparing the models it s found that there are areas with high crime rate and also there are some areas that has evolved over time span.

Result and Discussions

Since crime is increasing at an alarming rate globally it is important to control it. In order to reduce crime rate we need to study the crime rate at different places of the country. The required patterns and their relationship are searched in the dataset by using machine learning algorithms. This techniques helps in providing overview of the dataset and hence helps in searching, handling and retrieving the desired information. The figures below show the detailed comparative results in terms of accuracy.

It is concluded from the analysis that all the models are performing fair enough i.e. Naïve Bayes, Decision tree, SVM, KNN. But among the four considered models , decision tree is observed functioning as best model with greater accuracy. Performance evaluation is measured considering only the important attributes from the dataset . It is observed that decision tree classifier performs better than other classifier for the considered dataset and features.

Table. 1. Comparative Result of analysis

Algorithm used	*Accuracy*
Naïve Bayes	86.86
SVM	48.83
KNN	33.25
Decision Tree	98.77

Conclusion and Future Work

The crime rate in the world is increasing day by day due to many reason such as increase in poverty, unemployment, corruption etc. If the crime has increased necessary measures can be taken by the officials to study why the crime rate has increased and also how to reduce crime rate in that region. Many papers have been studied, only those papers with background in crime analysis and prediction are compared. Each paper has their own advantages and disadvantages. Each paper has its own individual approach for solving crime. Using python the accuracy of the proposed model will be measured and verified. The proposed model will be helpful for the law enforcement agencies in takin necessary steps to reduce crime. In future we may predict crime hot spot that will help in the deployment of police at most likely places of crime for any given window of time.

References

[1]. AdelAli Alkhaibari , Ping-Tsai Chung, Ping-Tsai Chung, "Cluster Analysis for Reducing City Crime Rates", in Long Island Systems, Applications and Technology Conference (LISAT), 2017

[2]. Alkesh Bharati, Dr Sarvanaguru RA.K ,"Crime Prediction and Analysis using Machine Learning" In International Research Journal of Engineering and technology (IRJET), Vol no. 05, pp. 2395-0072, 2018.

[3]. Lalitha Saroja Thota, Suresh Babu Changlasetty, "Cluster based Zoning of Crime Info", IEEE International Conference on Security and Privacy in Computing and Communications (Trust Com), 2017.

[4]. Sunil Yadav, Meet Timbadia, Nikhilesh Yadav, Rohit Vishwakarma. "Crime Pattern Detection, Analysis and Prediction", In International Conference on Electronics, Communication and Aerospace Technology, 2017.

[5]. S.R Deshmukh, Arun Dalvi, Tushar Bhalerao, Ajinkya Dahale, Rahul Bharati , Chaitali R. Kadam, "Crime investigation using Data mining", In International Journal of Advance Research in Computer and Communication Engineering (IJARCCE), Vol.4,Issue 3,March 2015

[6]. Ayidh alqahtani, Ajwani Garima, Ahmad Alaid, "Crime analysis in Chicago City", in 10th IEEE International Conference on Information and Communication System (ICICS), 2019.

[7]. Jesia Quader Yuki, Md.Mahfil Quader Sakib, Zaisha Zamal, Khan Mohammad Habibullah, Amit Kumar Das "Predicting Crime Using Time and Location Data", ICCCM 2019 Association for Computing Machinery.

[8]. Anand Joshi, A.Sai Sabitha, Tanupriya Chaudary, "Crime Analysis using k-mean Clustering", In International Conference on Computational Intelligence and Networks, IEEE, Vol. 18, No. 3, March 2017.

[9]. Benjamin Fredrick David, A.Suruliandi, "A Survey on Crime Analysis and Prediction using Data Mining" Techniques, ICTACT Journal on Soft Computing, April 2017 vol. 07, ISSUE.03

[10]. Shyam Varan Nath, "Crime Pattern Detection using Data Mining", Proceedings of IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology Workshops, pp. 1-4, 2006.

[11]. National Crime records bureau, Ministry of home affairs, India, Software for data analysis, Crime Info(Crime in India) , http://ncrb.nic.in/index.htm

[12]. J. Han, M. Kamber and J. Pei and M. Kamber, Data Mining, Concepts and Technologies, 3rd Edition, The Morgan Kaufmann, 2011