Int. j. inf. tecnol.
https://doi.org/10.1007/s41870-018-0260-7

ORIGINAL ARTICLE

Analysis and prediction of crime patterns using big data

Ravi Kumar[1] · **Bharti Nagpal**[1]

Received: 5 July 2017 / Accepted: 26 November 2018

Abstract Nowadays crimes are increasing at a high rate which is a great challenge for the police department of a city. A huge amount of data on different types of crimes taking place in different geographic locations is collected and stored annually. It is highly essential to analyze data so that potential solutions for solving and mitigating the crime incidents and predicting similar incident patterns for future becomes possible. Then it can be carried out using big data and various machine learning techniques in conjunction. The paper introduced a solution to the crime prediction problem using Naive Bayes classifier, which includes finding the most likely criminal of a particular crime incident when the history of similar crime incidents has been provided with the incident-level crime data. The incident-level crime data is provided as a crime dataset which includes incident date and location, crime type, criminal ID and the acquaintances are the attributes or crime parameters. The acquaintances are the suspects whose names are either directly involved in the incident or indirectly the acquaintances of the criminal. Acquiring a real-time crime dataset is a difficult process in practice due to confidentiality principle. So, crime dataset are used for the inputs using the state of the art methods. The proposed system is tested for the crime prediction problem using the data learning, and the experimental results show that the proposed system provides better results and finding of the potential solutions and crime patterns.

Keywords Crime patterns · Big data · Naive Bayes classifier · Incident-level crime data

1 Introduction

Nowadays crimes are increasing at a high rate which is a great challenge for the police department of a city. Over the last few decades, there has been a huge amount of data that has been gathered by the different law enforcement organizations. Data does not pertain to only a single type of crime, but it contains information of different types of criminal incidents that have taken place in different states and cities across a particular country. Solving the crimes has been the right way of the justice the criminal. And data brings a greater amount of challenges and opportunities for both researchers as well as analyzers. Researchers can find relationships in between the attributes of the data so they can guide the police to catch the criminal. Analysis of crime data needs a different approach for finding the trends and the unique patterns in the crime reports. Crime prediction is the major task in crime analysis that main aim is to discover the type of crime is going to happen in which place. The prediction of crime becomes a difficult and complex task, particularly when the data pertaining to the criminal incident, such as the profile of the criminal, social network of operation of the criminal involved in the crime, and any kind of geographical data such as the date of the incident and the geographical location of the criminal activity also gets included in the analysis.

In any crime based incident, the most important step is to find the criminals involved in the incident. Identifying

✉ Bharti Nagpal
bharti_553@yahoo.com

Ravi Kumar
ravikumar1772@gmail.com

[1] Department of Computer Science, Ambedkar Institute of Advanced Communication Technologies and Research, Delhi, India

Published online: 18 December 2018

can be done by analyzing the data that has been obtained from the crime scene and the analysis is done using the big data analytics by the researchers and the analyzers [18]. It is essential to discover the suspects of the incidents in order to optimize the usage of human and the technical resources. Crime-analysis tools have been developed and are being used by the security forces for solving such incidents. However, with the help of tool, the most important requirement is that the date and the geographical location of the incident are needed for the software to work properly. However, it is manually not possible to evaluate such huge amount of data using the tool and human resources. Thus, here steps in big data, which can be used in the crime analysis for finding patterns and predicting future incidents.

Now the level of lives are stored in data generation, in which complexity of data is increasing either it is in social post form or a simple multimedia message. In the past, the data is in structured or unstructured form but now it is in mixture form of different types which is making them more complex. The big data is a term in which information is aggregated at a very large scale whether it is by business, web service companies or retailer related. Databases have the ability to store the huge amount of information using different services provided by vendors and do some analytics on the big data. But with the machine learning, applications are able to teach the machine to learn data and predict the patterns which help in the decision making for the system.

1.1 Research methodology

In this paper, R programming is used to derive the trends of the crime as well as the prediction of the type of crime which is going to happen in a certain location. Crime data is in categorical form by which it is easy to apply analysis process on the dataset. The proposed framework is tested for the crime prediction using machine learning technique to stop the crime so that it should not be repeated in the same place next time and try to catch the criminal.

The second section of the paper describes a related literature review of existing work. The third section describes the theoretical background of the concepts that are used in the framework of analysis. The fourth section presents the results of the experimental analysis. Finally, the fifth section concludes the paper.

2 Literature review

This section is about the related literature survey in which different technique or algorithms are proposed by the researchers and how these techniques used to assure the better analysis of the crime prediction model.

In literature review, crime dataset can be divided into incident-based data or aggregate-based data [12, 14, 20, 25]. In the incident-based data, information is in a detailed form about the crime incidents, but in aggregate based data, information given is about only the group of crime incident for a certain process which summarizes the data about certain incidents of that group. However, for crime prediction, the data needs only is to be incident type related data, so aggregate level data cannot be used for similar problem. Incident-based information are available in a limited way because information about incidents is always confidential [8, 26]. So, crime prediction related information is hidden and it can be exposed with date and location of crime incidents.

In simplification, the work related to the analysis of crime data give exposure to techniques like visualization, supervised learning, statistical approaches and unsupervised learning [3–7]. Visualization represents the graphical view of between time and crime data such as the mapping through GIS, crime profiling, criminal prediction [1, 18, 20]. Similarly, finding the relation between the crime data and unsupervised learning technique like clustering are used mostly [9]. Clustering are used in crime data as analysis of number of crimes by dividing the crime over the different numbers of clusters to easily analysis, crime mapping or profiling, recognition of crime pattern [1, 3, 8, 9].

Deshmukh et al. [11] has proposed a criminal identification system in which they compared the J48, JRIP and Naïve Bayes algorithm against the sample criminal data get the best algorithm for identifying the criminal for the particular crime. In another paper, the author has collected the real dataset of crime from his country police department and trained the machine on the linear regression model by which they forecast the crime for different types of crime as dacoity, murder, robbery, child repression, kidnapping for the different regions of Bangladesh [12].

Kiani et al. has applied a complete theoretical model on different techniques of data mining like clustering and classification for the crime dataset of the police department of England over some period of time. For improving the quality of the values as well as the model they provide weights to attributes with the help of clustering and for optimization purpose Genetic algorithm [9, 10].

Ja video et al. has applied an approach to find the important entities from crime reports of police which are full narrative reports in plain text by automatically entering the crime data into the database. Also applied a clustering method named as SOM for analysis of crime and matching process [13, 22].

Almanie et al. has proposed a framework through which they are finding spatial and hotspots of the criminal by comparing the datasets with some statistical analytics.

After the analytics, they used data mining technique finding the patterns of the hotspots and different classifiers to predicting crime type [14].

Malathi et al. have concentrated on MV algorithm and apriori algorithm for using the missing values over the dataset of crime and finding useful patterns for knowledge accuracy of predictive crime [15].

Saeed et al. has used data mining techniques to compare classification techniques to predict the crime and also used machine learning algorithms to the crime dataset to predict the outcomes of a particular crime incident [16].

McClendon et al. has implemented regression algorithms and decision stump algorithm over the crime dataset for finding the best accurate machine learning technique to predict the violent crime. According to its implementation linear regression comes out to be the best algorithm for predicting crime [17].

3 Theoretical background

R

R is a software that contains a wide variety of tools that most data scientists use the most common in their workflow that includes data frame which helps to manipulate data easily. In the starting era of R, it is only computing language for statistics purposes but now supports data mining, analytics with a large number of visualization tools that helps to plot different characteristics of the dataset [23]. The biggest trend of R are the packages that are written over a past decade [19]. It supports different packages of data analytics and statistical analytics. So, a large number of functions that accounted are mainly one line function to run. While having a lot of usefulness, R has limitation too. It is limited to restricted to a single thread that tells from starting to ending it works analytics only on one thread. It is also restricted to the amount of memory that available on the machine.

3.1 SparkR

SparkR is a library package giving a light-weight frontend of Apache Spark, which is an open source cluster processing system [4]. It gives an interface to programming whole clusters with implicit information parallelism and adaptation to failure which is called a fault tolerance. The SparkR API enables the clients to intuitively run the employments from the R shell on a bunch. The passage point into SparkR is the SparkContext which interfaces the R program to a Spark cluster. SparkR.init is utilized to make a process to which the name of an application and connected spark library packages are passed. SparkR gives distributed parallelism for the operation like filtering, separation, collection, and so on. Accordingly, significantly bigger datasets can be effectively taken care of than with R. It likewise bolsters disseminated machine getting the hang of utilizing MLlib [24] (Fig. 1).

3.2 Machine learning

Machine learning is basically an artificial intelligence (AI) wherein machine can learn on its own code ability without providing the explicit programmer. The main motive is that when a problem is counter it doesn't write program again, but it changes its own code according to its new scenario that discovered.

Itself learned what has to be learned from provided data scenario, past expressions and learning from past experiences it comes up with a new situation.

3.3 Types of machine learning

3.3.1 Supervised learning

In supervised learning, supervised means monitoring something constantly. Supervised learning is one of the machine learning algorithms in which a known dataset which is also known as train dataset is used to make a prediction. For example, in a classroom teacher teach a student a difficult concept with the example.

In algorithm, there are a set of inputs and also corresponding responses so it can predict new responses on the basis of input responses. The supervised learning algorithm is mainly used to build a model that take values from inputs that help to predict the values for a new dataset.

3.3.2 Unsupervised learning

Unsupervised learning is used to create patterns or results from the dataset which contains labeled input data. In this,

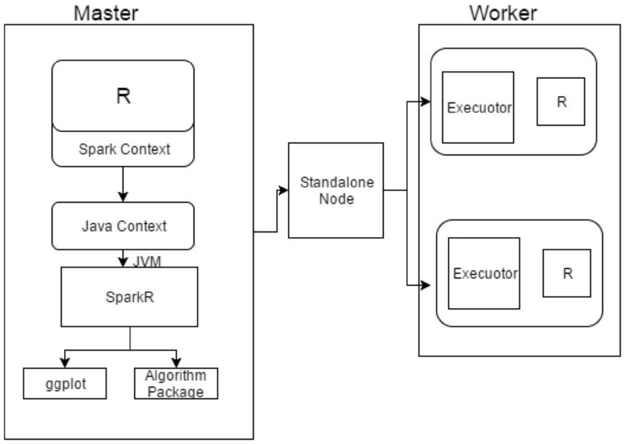

Fig. 1 The structure of SparkR cluster [3]

only inputs are provided not answers so because of having no answer or data machine can only find pattern or structure from inputs only. For example, a student learning on their own, they have books, they are trying to figure out what thing is on their own. The problem in unsupervised learning could be when system don't know whether there is a correlation in the data or structure in the data, then the job of the algorithm is to find the pattern in the data.

3.3.3 Reinforcement learning

Reinforcement learning is another type of machine learning in which it is totally based on the behavior psychology so that how can a software do an action to maximize the reward notion. The machine trying to take a decision. Machine aim is to maximize the rewards when it does action, it gives the proper result of maximum reward it has.

3.4 Naïve Bayes

Naïve Bayes is one of the classification technique that represents a supervised machine learning technique which is based on the most popular Bayes' Theorem proposed by Thomas Bayes (1702–1761).

In Naïve Bayes, there is an independent assumption made between the predictors [21]. Basically, as a classifier, Naïve Bayes always make the assumption that availability of one characteristic in a particular class is unrentable to another available characteristic in the same class. Conditional independence is the assumption that provided earlier.

Assumption can be explained as:

$$\boldsymbol{P}(\boldsymbol{H}|\boldsymbol{D}) = \boldsymbol{P}(\boldsymbol{D}|\boldsymbol{H})\boldsymbol{P}(\boldsymbol{H})/\boldsymbol{P}(\boldsymbol{D})$$

$$\boldsymbol{P}(\boldsymbol{H}|\boldsymbol{D}) = \boldsymbol{P}(\boldsymbol{d}_1|\boldsymbol{c}) \times \boldsymbol{P}(\boldsymbol{d}_1|\boldsymbol{H})\ldots\boldsymbol{P}(\boldsymbol{d}_n|\boldsymbol{H}) \times \boldsymbol{P}(\boldsymbol{H})$$

$\boldsymbol{P}(\boldsymbol{H}|\boldsymbol{d})$ represents the posterior probability of H class and the predictor is d and the attributes. $\boldsymbol{P}(\boldsymbol{H})$ is the probability of class H, which is also known as prior probability. $\boldsymbol{P}(\boldsymbol{D}|\boldsymbol{H})$ is the likelihood and also the probability of given predictor class. $\boldsymbol{P}(\boldsymbol{D})$ is the probability of predictor class which is also called as prior probability.

Naive Bayes provides previous responses, observed data, and learning technique. It provides a clear view for computing and understanding many learning techniques. For example, a vegetable is considered to be a tomato only if it has a color red, round in shape and has some 2–3 inches in length. So, all these characteristics are depending on each other or existence of another characteristic. In the end, all these characteristics differently give the probability that vegetable is a tomato. The reason it is called as 'Naïve'. It particularly needed when inputs are of a high value of dimensionality. In naive Bayes, while estimating the parameters the likelihood should be maximum in the method so that it performs better in complex problems. It is mainly used to make decision and interpretation statistics which works with the conclusion of probability. It is used to predict events of future from prior based events (Fig. 2).

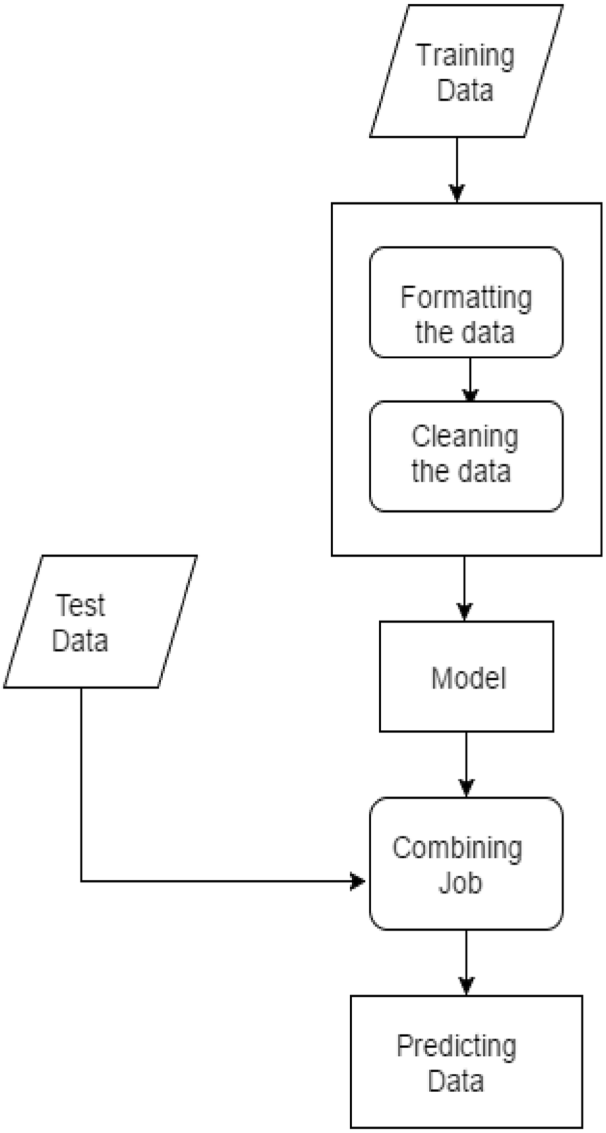

Fig. 2 The job process of Naïve Bayes [4]

4 Results

The Table 1 is a training dataset, which gives descriptions of the date and time of the crime, the type of crime taking place, the location where the crime has taken place and the day on which it took place. In order to be more accurate about the location of the crime, the Latitudes (*X*) and Longitudes (*Y*) of the location have been specified as well. The training dataset is stored in the Spark cluster, the

Table 1 Training dataset acquired from Cheltenham crime data [7]

Date and time	Address	City	*Y*	*X*	Day_of_week	Incident_type
1/1/2011 12:49	MT VERNON & CHELTENHAM AVE	ELKINS PARK	40.0769444	– 75.1269444	Saturday	Traffic
1/1/2011 12:49	W CHELTENHAM AVE & LAKESIDE	ELKINS PARK	40.0587165	– 75.1317043	Saturday	Traffic
1/1/2011 14:35	400 Block ACCOMAC RD	WYNCOTE	40.0878328	– 75.1440846	Saturday	Property crime
1/1/2011 18:41	300 Block LIMEKILN TPK	GLENSIDE	40.100996	– 75.16377	Saturday	Breaking and entering
1/2/2011 12:41	2300 Block W CHELTENHAM AVE	WYNCOTE	40.071591	– 75.154416	Sunday	Theft
1/3/2011 0:36	500 Block JEFFERSON AVE	CHELTENHAM	40.0661785	– 75.094054	Monday	Robbery
1/3/2011 9:28	500 Block COTTMAN AVE	CHELTENHAM	40.0669177	– 75.0934468	Monday	Disorder
1/4/2011 11:24	300 Block GERARD AVE	ELKINS PARK	40.068979	– 75.123101	Tuesday	Community policing

Table 2 Test dataset generated from Cheltenham crime data [2]

Id	Date and time	Day_of_week	City	Address	*Y*	*X*
1	1/1/2017 9:45	Sunday	WYNCOTE	2400 Block SHOPPERS LN	40.0764547	– 75.1543659
2	1/1/2017 10:05	Sunday	GLENSIDE	100 Block BICKLEY RD	40.099277	– 75.152725
3	1/1/2017 15:13	Sunday	ELKINS PARK	OLD YORK RD & 02:00:00:00	40.07479	– 75.1294946
4	1/1/2017 15:47	Sunday	WYNCOTE	2000 Block BL CHELTENHAM AVE	40.0692777	– 75.1502952
5	1/1/2017 17:19	Sunday	WYNCOTE	2400 Block SHOPPERS LN	40.0764547	– 75.1543659
6	1/1/2017 18:50	Sunday	ELKINS PARK	500 Block SHOEMAKER RD	40.0764079	– 75.1252675
7	1/2/2017 8:54	Monday	PHILA	7600 Block WASHINGTON LN	40.0751799	– 75.1506404
8	1/2/2017 9:52	Monday	GLENSIDE	7800 Block COBDEN RD	40.0922301	– 75.1794567

cleaning and formatting techniques categorizes the incidents on the basis of the time of the incident into the morning, afternoon and evening depending on when the crime took place.

The Table 2 shows the dataset which is used for the purpose of testing after the classifier model is built using the training data set and the classification algorithm. The complete dataset is split into training and test data sets. 70% of the data from the original dataset is used for training the algorithm for building the classifier model and the remaining 30% is used for the purpose of testing using that classifier model. Dataset also shows the date and time of the incident, location (along with the latitude and longitude) and the type of incident taking place.

4.1 Visualization of crime trends

See Figs. 3, 4, 5, 6, 7, 8.

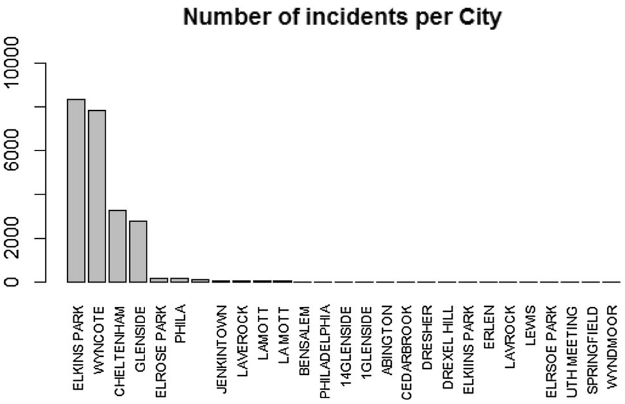

Fig. 3 Analysis of crime over the cities of Cheltenham [5]

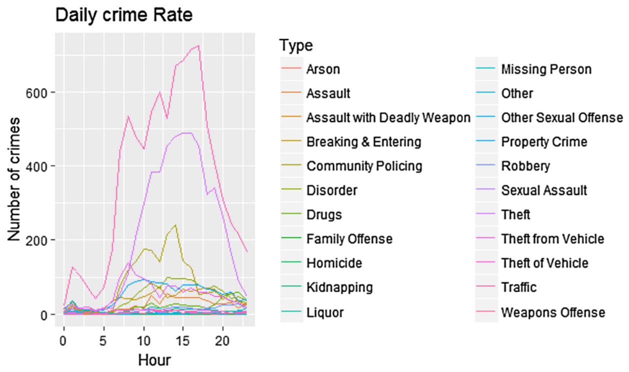

Fig. 4 Hourly analysis of each crime [20]

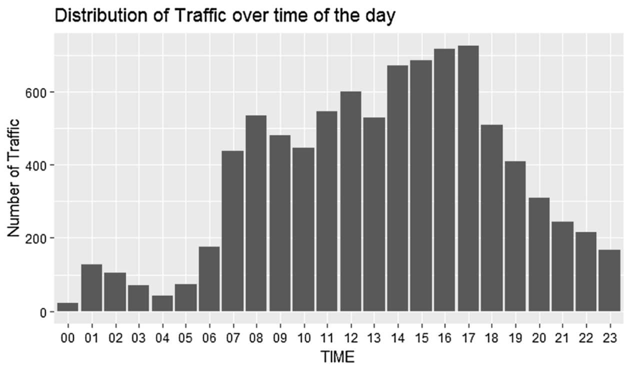

Fig. 5 Hourly analysis of traffic crime [5]

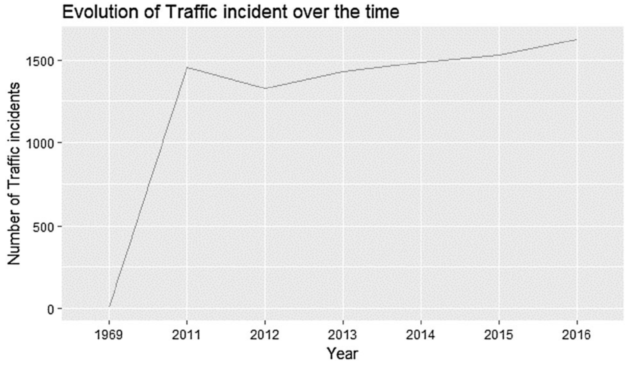

Fig. 6 Traffic crime trend over the years [3]

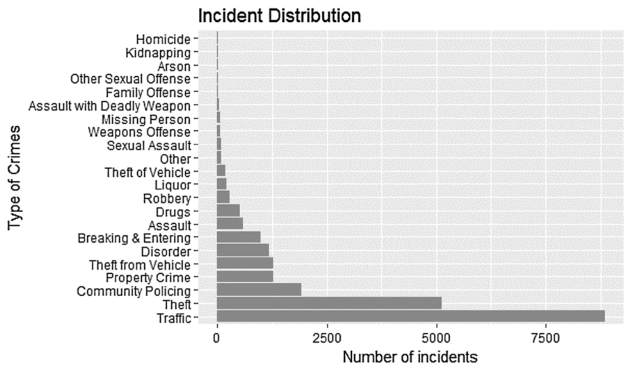

Fig. 7 All crime trends in Cheltenham [3]

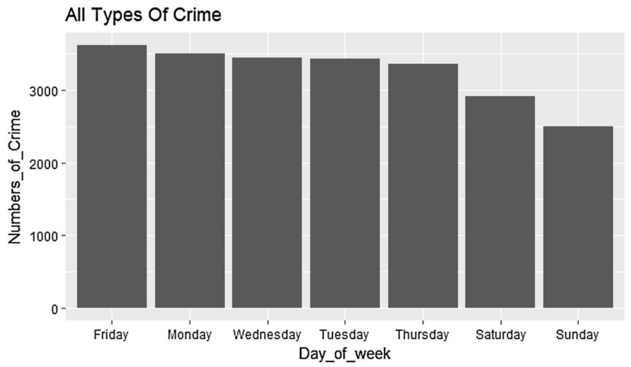

Fig. 8 Analysis of crime day wise [2]

Table 3 Predicted Crime patterns of the test data

Id	City	Shift	Incident_Pred
1	WYNCOTE	Night	Property crime
2	GLENSIDE	Night	Traffic
3	ELKINS PARK	Afternoon	Property crime
4	WYNCOTE	Afternoon	Property crime
5	ELKINS PARK	Evening	Property crime
6	PHILA	Afternoon	Property crime
7	GLENSIDE	Afternoon	Traffic
8	LAVEROCK	Afternoon	Disorder

Table 4 Reported accuracy of each class in crime prediction using Naive Bayes classifier [15]

Class	Accuracy
Disorder	0.506493
Property crime	0.744647
Theft from vehicle	0.5000000
Traffic	0.5605139
Drugs	0.4969136

4.2 Predictive analytics

The Table 3 presents the dataset, which has been predicted by the classifier algorithm after building the classifier model using the training dataset. The dataset contains information about the city where the crime has taken place, the shift of the day when the incident takes place and the type of incident taking place. From the predicted data, it is evident that most of the incidents have taken place in the afternoon and only one of the incident occurred during the evening, while the others occurred during the night shift. It can also be observed that ‘property crime’ is the type of crime that is the most prevalence of all the crimes. The other incidents taking place are Traffic, disorder, crime related to drugs and crime related to the theft of a vehicle.

The Table 4 shows the statistics regarding the different time of crime incidents taking place in a particular country.

These incidents have been predicted on the basis of accuracy. In table, it can be seen that the 'property crime' has the highest prediction accuracy of 74.4% as compared to other types of crimes predicted. It also has the highest sensitivity of 76.6% as compared to others.

5 Conclusion

The purpose of above study was to explore the applicability of data mining techniques for detection of crime patterns. Paper provides information about crime based incidents, showing the percentage different types of crime incidents that occur in Cheltenham, the areas which are more prone and sensitive to crime and the percentage of each type of crime incidents that occurs in each area. Area types considered in work are: slums, residential, commercial, VIP zones, travel points and markets and crime types considered are: heinous crimes, non-heinous crimes and special and local laws. It provides information that which areas are sensitive towards crime and also an association between areas and crime types. Analysis can help Cheltenham Township Police Department in analyzing crime profiles and finding potential solutions to mitigate similar incidents in future. It also helps into detect patterns and trends of crime incidents for the future purpose. Finally, an attempt has also been made to extract crime profiles hence police officials can make use of these profiles in their day-to-day battle against crime.

6 Future scope

This framework can be enhanced with different security services like data reliability, privacy etc.

Using artificial intelligence it can be further enhanced so as to increase the accuracy and to reduce certain types of crime incidents.

References

1. Gera P, Vohra R (2014) City crime profiling using cluster analysis. Int J Comput Sci Inf Technol 5(4):5145–5148
2. Vural S, Gok M, Yetgin Z (2013) Generating incident-level artificial data using GIS-based crime simulation. International Conference on IEEE Electronics, Computer, and Computation (ICECCO' 2013), pp 239–242
3. Brunsdon C, Corcoran J, Higgs G (2007) Visualising space and time in crime patterns: a comparison of methods. Comput Environ Urban Syst 31(1):52–75
4. Xiang Y, Chau M, Atabakhsh H, Chen H (2005) Visualizing criminal relationships: comparison of a hyperbolic tree and a hierarchical list. Decis Support Syst 41(1):69–83
5. Jain LC, Seera M, Lim CP, Balasubramaniam P (2014) A review of online learning in supervised neural networks. Neural Comput Appl 25(3):491–509
6. Corsini P, Lazzerini B, Marcelloni F (2006) Combining supervised and unsupervised learning for data clustering. Neural Comput Appl 15(3):289–297
7. Enzmann D, Podana Z (2010) Official crime statistics and survey data: comparing trends of youth violence between 2000 and 2006 in cities of the Czech Republic, Germany, Poland, Russia, and Slovenia. Eur J Crim Policy Res 16:191–205
8. Vural MS, Gok M, Yetgin Z (2014) Analysis of incident-level crime data using clustering with hybrid metrics. GAUJ Appl Soc Sci 6:8–20
9. Kiani R, Mahdavi S, Keshavarzi A (2015) Analysis and prediction of crimes by clustering and classification. Int J Adv Res Artif Intell (IJARAI) 4(8):11–17
10. Yamuna S, Sudha Bhuvaneswari N (2012) Data mining techniques to analyze and predict crimes. Int J Eng Sci (IJES) 1(2):243–247
11. Deshmukh SR, Dalvi AS, Bhalerao TJ, Dahale AA, Bharati RS, Kadam CR (2015) Crime investigation using data mining. Int J Adv Res Comput Commun Eng (IJARCCE) 4(3):22–24
12. Awal MA, Rabbi J, Hossain SI, and Hashem MMA (2016) Using linear regression to forecast future trends in crime of Bangladesh. In 5th International Conference on Informatics, Electronics, and Vision (ICIEV)
13. Keyvanpoura M, Javidehb M, Ebrahimia MR (2011) Detecting and investigating crime by means of data mining: a general crime matching framework. Proc Comput Sci 3:872–880
14. Almanie T, Mirza R, Lor E (2015) Crime prediction based on crime types and using spatial and temporal criminal hotspots. Int J Data Min Knowl Manag Process (IJDKP) 5(4):1–19
15. Malathi A, Baboo SS (2011) An enhanced algorithm to predict a future crime using data mining. Int J Comput Appl 21(1):1–6
16. Saeed U, Sarim M, Usmani A, Mukhtar A, Shaikh AB, Raffat SK (2015) Application of machine learning algorithms in crime classification and classification rule mining. Res J Recent Sci (ISCA) 4(3):106–114
17. McClendon L, Natarajan Meghanathan N (2015) Using machine learning algorithms to analyze crime data in machine learning and applications. Int J (MLAIJ) 2(1):1–12
18. Hussain S, Lee S (2015) Visualization and descriptive analytics of wellness data through Big Data. IEEE Conf Digital Inf Manag (ICDIM). https://doi.org/10.1109/icdim.2015.7381878
19. Venkataraman S, Yang Z, Liu D, Liang E, Falaki H, Meng X, Xin R, Ghodsi A, Franklin M, Stoica I, Zaharia M (2016) SparkR: Scaling R Programs with Spark SIGMOD
20. Vural MS, Gok M (2016) Criminal prediction using Naive Bayes theory. Neural Comput Appl. https://doi.org/10.1007/s00521-016-2205-z
21. Jung YG, Kim KT, Lee B, Youn HY (2016) Enhanced Naive Bayes Classifier for real-time sentiment analysis. In: International Conference on IEEE information and communication technology convergence (ICTC)
22. Krishnamurthy R, Satheesh Kumar J (2012) Survey of data mining techniques on crime data analysis. Int J Data Min Tech Appl 01(02):117–120
23. Mena J (2003) Investigative data mining for security and criminal detection. Butterworth-Heinemann Press, Oxford, pp 15–16
24. Meng X, Bradley J, Yavuz B, Sparks E, Venkataraman S, Liu D, Freeman J, Tsai DB, Amde M, Owen S, Xin D, Xin R, Franklin MJ, Zadeh R, Zaharia M, Talwalkar A (2016) MLlib: machine learning in Apache Spark. Journal of Mach Learn Res 17(34):1–7
25. Poulsen E, Kennedy LW (2004) Using dasymetric mapping for spatially aggregated crime data. J Quant Criminol 20(3):243–262
26. Kumar S, Toshniwal D (2016) A novel framework to analyze road accident time series data. J Big Data 3:8. https://doi.org/10.1186/s40537-016-0044-5