International Journal of Engineering Applied Sciences and Technology, 2021
Vol. 5, Issue 10, ISSN No. 2455-2143, Pages 99-103
Published Online February 2021 in IJEAST (http://www.ijeast.com)

AN OVERVIEW OF CRIME ANALYSIS, PREVENTION AND PREDICTION USING DATA MINING BASED ON REAL TIME AND LOCATION DATA

Okeke Ogochukwu C
Department of Computer Science
Chukwuemeka Odumegwu Ojukwu University,
Uli, Anambra State, Nigeria.

Oranyelu Forster O.
Department of Computer Science
Chukwuemeka Odumegwu Ojukwu University,
Uli, Anambra State, Nigeria.

***Abstract*— Crime analysis and prevention is a systematic approach for identifying and analyzing patterns and trends in crime. The system can predict regions which have high probability for crime occurrence and can visualize crime prone areas. With the increasing advent of computerized systems, crime data analysts can help the Law enforcement officers to speed up the process of solving crimes. Using the concept of data mining, real time and location data, the system can extract unknown, useful information from an unstructured data. Here we have an approach between computer science and criminal justice to develop a data mining, real time and location data procedure that can help solve crimes faster. Instead of focusing on causes of crime occurrence like criminal background of offender, political enmity etc. we are focusing mainly on crime factors of each day. To have a better response towards criminal activity, it is very important that one should understand the patterns in crime. I analyses these patterns by taking crime datasets from the Nigeria Police zone six (6) Calabar, Cross River State and town planning bodies. This dataset includes different streets of the city of Calabar. The major aim of this mission is to expect which category of crime is most probably to take place at a detailed time and places in Calabar.**

Keywords: **Naive Bayes, Apriori algorithm, Decision tree, crime, Mongo DB.**

I. INTRODUCTION

Crime rate increases on a daily basis. Crime as the word suggests is the violation that people does, and it is usually performed against the laws and it is punishable. Crime cannot be predicted since it is not systematic. Also the modern technologies and hi-tech methods help criminals in achieving their goals. According to the Nigeria police, crimes like burglary, arson and it likes have been decreased while crimes like sex filming, rape, robbery, fraud, kidnapping etc. are increasing rapidly. Though, crime victims might not be easily predicted but the time and location can be prediction based on the probabilities of its occurrence. The predicted results cannot be assured of 100% accuracy but these results shows that the software program helps in reducing crime rate to a certain extent by providing security in vital areas where crimes can easily occur. Developing such a powerful crime analytics tool we have to collect crime records and evaluate them.

Criminal and sociology scholars are analysing the pattern of criminal activity and its relationship with the area. Researchers have shown that many crook activities are taking place in a region. This is called a hotspot. Machine learning can be used to become aware of hotspots by way of data pushed approach.

It is only within the last few decades that technology made spatial data mining a practical solution for wide audiences of Law enforcement officials which is affordable and available. Since the availability of criminal records is limited, the collection crime data from various sources like newspapers, new websites, blogs, social media etc. This huge data is used as a record for creating a crime record database. So the main challenge is developing a better, efficient crime pattern detection tool to identify crime patterns effectively.

This paper will solve problems of Law enforcement bodies not having clue as to when crimes will be committed, what type of crime will be committed, which group of people will commit the crimes, and the location a crime will be committed.

The objective of this study is to develop a system that will analyse and predict crime so that Law enforcement bodies will be able to; have clues of when crimes will be committed, what type of crime will be committed, which group of people will commit the crimes and the exact location the crime will take place. These will help the Law enforcement bodies arrive early and possibly put an end to the upcoming crime.

II. LITERATURE REVIEW

Criminal activities are common around the world. Therefore, researchers have completed many works on this subject matter.

RELATED WORK

Researches have been analysing the relation among criminal activities and socio-economic variables like unemployment, earnings level, level of schooling and so forth. Researchers like Alkesh Bharati et al worked-on Crime Prediction and Analysis Using Machine Learning where the author used machine learning and data mining for prediction of crimes in Chicago [1]. The datasets include information like vicinity description, type of crime, date, time, range, longitude, etc. They used the K-Nearest Neighbor (KNN) classification and plenty of different algorithms to test for crime prediction. A classifier that gives higher accuracy is used for further training. Also, Ankit Sangani et al worked on similar paper on Crime Prediction and Analysis where they used Simple K-Means clustering techniques and algorithm for predicting Crimes. This model tried to figure out crime trends based on crime zone but not based on a certain period [2]. Therefore, they can extend their model by using more datasets with more features such as season, date, time so that this model will be more beneficial of police using as well as safety of normal people.

In countries like England, Cambridge Police Department have done a similar one named Series Finder for finding the patterns in burglary. For achieving this they used the modus operandi of offender and they extracted some crime patterns which were followed by offender. The algorithm constructs modus operandi of the offender. The M.O. is a set of habits of a criminal and is a type of behaviour used to characterize a pattern. The data included means of entry (front door, window, etc.), day of the week, characteristics of the property (apartment, house), and geographic proximity to other break-ins. Using nine known crime series of burglaries Series Finder recovered most of the crimes within these patterns and also identified nine additional crimes. The predicted result showed more than 80% accuracy. So the same concept we are applying here i.e. find unknown patterns from known data and facts. It's the first mathematically principled approach to the automated learning of crime series.

Crime Prediction using Ensemble Approach by Ayisheshim Almaw et al investigates on hidden crime data mining .They used ensembled classification learning methods. They used Naïve Bayes classification and artificial network to test for crime analysis. They identify the crime trends and patterns and predicts the type of crimes might occur next in a specific locality of longitude and latitude ensemble classification learning methods [3]. They used Naïve Bayes classification and artificial network to test for crime analysis. They identify the crime trends and patterns and predicts the type of crimes might occur next in a specific locality of longitude and latitude in a specific schedule of time and season. Therefore, they can further implement the required combined techniques for improving a better crime prediction of a single classifier model by integrating multiple models by using more datasets with more features. Some researchers like Christian Tabedzki et al. used the random forest model. By the usage of random forest and k-nearest neighbor, researchers obtained the first-rate accuracy across 39 different categories. Since the information was very noisy, consequently they thought random rest might provide great effects [4].

In addition, Clifton Phua et al. offer a new fraud detection technique that is built by thinking about existing fraud detection research. Three distinctive algorithms are used at the identical skewed statistics. Using these algorithms, the accuracy obtained is greater than 90% [5]. Chandrasekar et.al. implemented Gradient Boosted trees and Support Vector Machines in their other paper and showed that their work on implemented algorithm gives high accuracy. They have worked on especially centred on one-of-a-kind cities. Last works related to crimes in different cities, expected many kinds of crimes occurring in the metropolis. The paper has discovered that the maximum suitable classifiers that possible practice on those sorts of datasets can be tree-based methods.

Hyeon-Woo Kang et al. analyzed crime occurrences by the usage of multimodal records in which they have applied deep learning. They found out that DNN version provides greater precision values in predicting crime prevalence than other prediction fashions when they are compared with other works. Their present crime predicting methodology for finding occurrences is not able to produce statistics based on the unique form of at a selected time [6]. Prediction of Hourly Effect of Land Used on Crime by Irina Matijosaitiene et al predicted future crimes based on the time using Manhattan. They used the random forest and logistic regression for their predicting model of crimes based on exact time features when most of the crimes happened. They achieved high accuracy on random forest algorithms [7]. They also tried to analyze hot spot feature to predict for controlling crimes in different areas. In their future work, they can also add specific longitude and latitude of a specific area where crimes mostly committed. Survey of Crime Analysis and Prediction by Mookiah et al studied crime connected variables, which showed the influencing factor of crimes and tried to figure out the rate of crimes [8].

Boni et al worked-on Area-Specific Crime Prediction Models where they predicted crimes by using zip code located specific area [10]. In their model, they used crime datasets from Chicago where they tried to discover two types of model one is hierarchical models based on certain locality and another is regularized multitasking for this model [9]. Naïve Bayes Approach for the Crime Prediction in Data Mining by Mrinalini Jangra et al used Machine learning techniques and regression techniques for predicting and analyzing of criminal datasets related for predicting crimes in India based on time and location. This model gives uniform characteristics result

International Journal of Engineering Applied Sciences and Technology, 2021
Vol. 5, Issue 10, ISSN No. 2455-2143, Pages 99-103
Published Online February 2021 in IJEAST (http://www.ijeast.com)

for noisy data but KNN regression techniques and Naïve Bayes are utilized for this model and Naïve Bayes gives the high accuracy for this predictive model [10]. Combining two datasets - 1990 US Law Enforcement Management and Administrative Statistics (LEMAS) and crime records 1995 FBI Uniform Crime Reporting (UCR), Iqbal et al. found an accuracy of 83.95% when used Decision Tree and Naive Bayesian set of rules is used to predict the category of crimes for various states of the USA. Similarly, Dash et al. analyze Chicago city crime facts fused with different social information sources the usage of community analytic strategies to expect a criminal activity for the following 12 months. They experimented with polynomial, auto-regressive and support vector regression methods and determined that the first-rate of support vector regression considerably outperforms other methods. Sathyadevan et al endorse an approach for detecting and figuring out crime using data mining and machine learning techniques. They implemented k-means clustering for crime detection, where it generates two crime-clusters. To increase k-means achieved results, they added GMAPI, which embeds Google maps through NetBeans. The accuracy that they receive from their model is of 93.62% and 93.99%, respectively [11]. Paper on Different Approaches for Crime Prediction system by Varshitha D N et al. predicted crimes based on location and time from previous crimes data records. They applied deep learning technique to predict crimes and analyzed sentimental techniques too [12]. Varvara Ingilevich et al used criminal datasets of the city of Saint-Petersburg to implant the type of the crimes and tried to predict the possible crimes in future on specific urban area depending on time to decrease the crimes of that area [13]. They used logistic regression, linear regression, gradient boosting for this predicting, forecasting classification model. In addition, they compared with the predictive results and figured out that gradient boosting gives the high accuracy for this model. They can further extend their work by focusing on longitude and latitude of a specific urban area with more features. Sadhana and Sangareddy have used twitter records and sentiment evaluation to predict crime in real time. They extensively utilized this fact to map the awareness of crime occurrences and discover huge scale hotspots. 90%.

This paper aims to predict the crime occurrences in Chicago based on time and location. The algorithms are trained with time attributes such as “Day”, “Month”, “Year”, “Hour”, “Minute”, “Second” and with location attributes such as “Location”, “Location Description”, “Block”, “Latitude” and “Longitude”.

III. Experiment and Result

Because of the limitations of the existing system, there is need for an automated approach which will go a long way in order to put an end to crimes and criminality at large. The proposed system will use dataset collected from the Nigeria police force zone six Calabar. The datasets include facts on crimes that has happened in Cross River State from 2005 to 2020 respectively.

Methodology

There are different attributes of the dataset. The attributes that are used in this paper is given in the table; Location, Description, Area code, Street Location, Landmarks, Population density, Latitude, Longitude, Month, Day, Hour, Minute, Second, Primary.

The algorithms used to train the dataset are; Apriori, Random Forest, Decision Tree, and different ensemble methods like Bagging, AdaBoost, and Extra Trees.

The following steps are followed for all the implemented algorithms;

i. Data Collection
ii. Pattern Identification
iii. Prediction

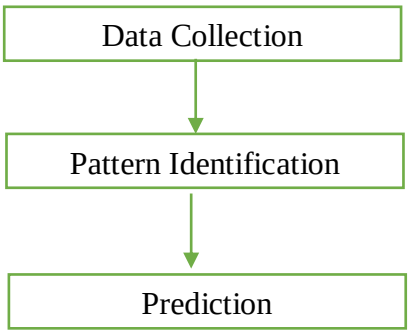

Fig 1 proposed system

i. **Data Collection**: I am collecting data from different web sites like news sites, blogs, social media, RSS feeds etc. The collected data is stored into database for further process. Since the collected data is unstructured data we use Mongo DB. Crime data is an unstructured data since the no of field, content, and size of the document can differ from one document to another the better option is to have a schema less database. Also the absence of joins reduces the complexity. Other benefits of using an unstructured database is that:

   a. Large volumes of structured, semi-structured, and unstructured data
   b. Object-oriented programming that is easy to use and flexible.

The advantage of No SQL database over SQL database is that it allows insertion of data without a predefined schema. Unlike SQL database it not need to know what we are storing in advance, specify its size etc.

- Attributes in our dataset with string type are “Date”, “Location”, “Location Description” etc. Using python this paper assigned numeric values for those capabilities.
- Since time is considered as the main factor thus “Date” has been split into “Day”, “Month”, “Year”, “Hour”, “Minute”, “Second” attributes.

International Journal of Engineering Applied Sciences and Technology, 2021
Vol. 5, Issue 10, ISSN No. 2455-2143, Pages 99-103
Published Online February 2021 in IJEAST (http://www.ijeast.com)

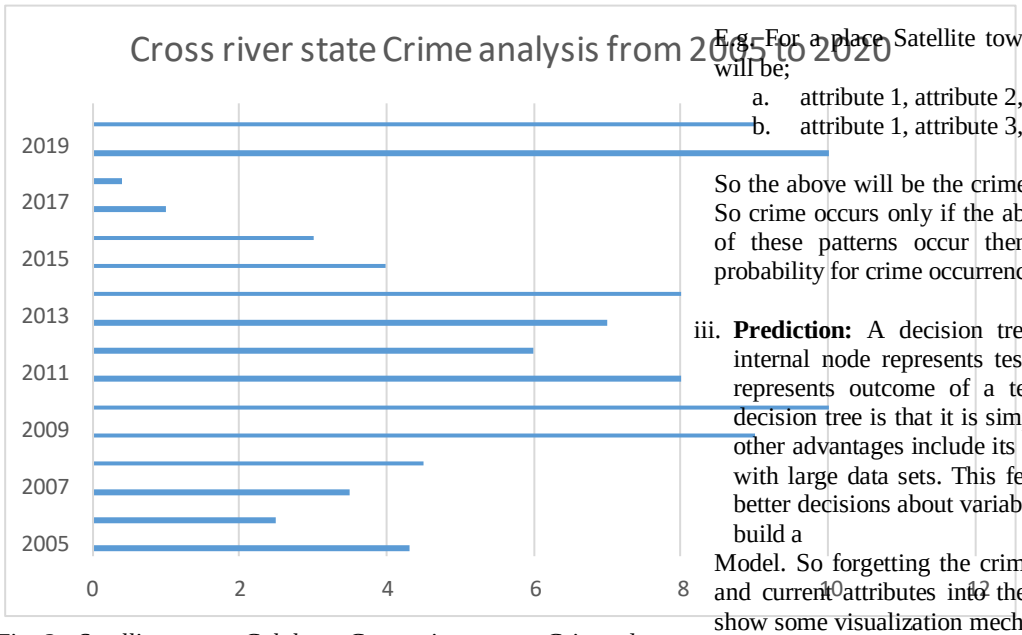

Fig 3. Satellite town, Calabar, Cross river state Crime data analysis from 2005 to 2020 by percentage

In Fig 3. Y axis represent year and X axis represents the percentage frequency of crimes over the year from 2005 to 2020 respectively. There was high rate of robbery, burglary, rape from 2005 to 2015 when cybercrime took pace and making physical crime fall low. But in recent time from 2019 crime rate in Satellite town, Calabar, Cross river state began to elevate again probably because internet fraud victims has come to a realization of all tricks used by fraudsters. This made the fraudsters to hit the street once again to commit all sort of crimes like robbery, murder, rape etc.

ii. **Pattern Identification:** The second phase is the pattern identification phase where we have to identify trends and patterns in crime. For finding crime pattern that occurs frequently we are using Apriori algorithm. Apriori can be used to determine association rules which highlight general trends in the database. The result of this phase is the crime pattern for a particular place. Here corresponding to each location we take the attributes of that place like VIP presence, weather attributes, area sensitivity, notable event, presence of criminal groups etc. After getting a general crime pattern for a place, when a new case arrives and if it follows the same crime pattern then we can say that the area has a chance for crime occurrence. Information regarding patterns helps police officials to facilitate resources in an effective manner. They can avoid crime occurrence by providing security/patrolling in crime prone areas, fixing burglar alarms / CCTV etc.

Take a sample list of 100 news for a place and apply Apriori algorithm in it. It will mine the frequent crime patterns for a place. So if there is a pattern in which crime occurred then we assume that if again that pattern occurs in a place then there is probability for crime occurrence in that place. We are considering several attributes for crime pattern detection.

E.g. For a place Satellite town, Calabar the pattern after mining will be;

a. attribute 1, attribute 2, attribute 3, attribute 4
b. attribute 1, attribute 3, attribute 4, attribute 5

So the above will be the crime pattern for Satellite town, Calabar. So crime occurs only if the above patterns occur on a day. If any of these patterns occur then only we can say that there is probability for crime occurrence.

iii. **Prediction:** A decision tree is similar to a graph in which internal node represents test on an attribute, and each branch represents outcome of a test. The main advantage of using decision tree is that it is simple to understand and interpret. The other advantages include its robust nature and also it works well with large data sets. This feature helps the algorithms to make better decisions about variables. Corresponding to each place we build a

Model. So forgetting the crime prone areas we pass current date and current attributes into the prediction software. The result is show some visualization mechanisms. Fig 2 shows the example of a decision tree model. Below shown is the example of decision trees of two different places.

Area sensitivity	**Notable event**	**VIP presence**	**Criminal group**	**Crime**
Yes	Yes	Yes	No	Yes
Yes	Yes	No	Yes	No
No	No	No	Yes	No
Yes	No	No	No	No
Yes	Yes	Yes	Yes	Yes
No	Yes	No	No	No

Table 1. Decision tree

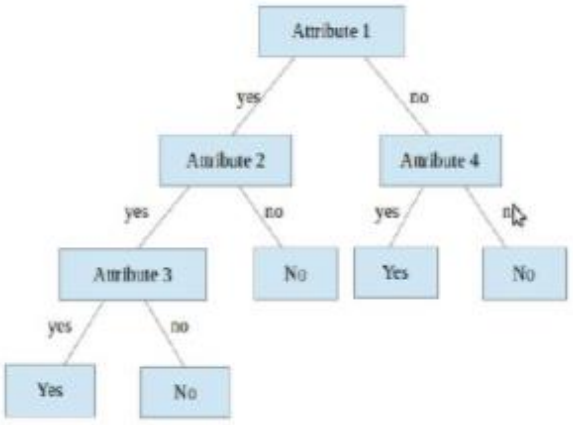

Fig 2 a decision tree

The working of decision tree seems to be little confusing but it's really easy. Consider a variety of plant species. We classify them according to order, genus, species etc. Instead we have to classify them into a common category as shrubs and trees. If a new species is identified then we have to classify this into any of the two categories. Basically we categorize it based on its characteristics i.e. we have a set of questions to check whether it satisfies the conditions. If first condition is satisfied then we check the next case and if the first condition itself is not satisfied then there is no need to check the rest. So the series of questions and their answers can be organized in the form of a decision tree. The tree has three types of nodes:

a. A Root node, that has incoming edges and zero or more outgoing edges.
b. Internal nodes, each of which has one incoming edge and two or more outgoing edges.
c. Leaf node or end node, each of which has exactly one incoming edge and no outgoing edges.

This supervised machine learning technique builds a decision tree from a set of class labelled training samples and by using this tree, tests the new samples. It is a predictive model which uses a set of binary rules to calculate the class value. The tree determines:

- Which variable to split at a node.
- Decision to stop or split.
- Assign terminal node

IV. RESULTS

This paper uses dataset which contains both the mixture of categorical and numeric values. Thus, the paper mainly focuses on those algorithms which can work on the combination of both categorical and numeric values. Also, keeping in mind that, the algorithm performs well for our classification problem. Therefore, several algorithms are chosen to serve the purpose such as Decision Tree, Random forest and several ensemble methods such as Bagging, AdaBoost and ExtraTree Classifier. The main motive of this paper is to use algorithms on these datasets to classify the type of crime occurring based on time and location. The chosen algorithms are applied where it provides a simple and fast way of learning a function. This is where the algorithm maps data x to outputs y, where x is a mixture of categorical and numeric variables and y is the categorical value for classification. The applied algorithm gives better performance for any classification problem. The result after reducing the classes is shown in the below table for all algorithms

V. CONCLUSION

This paper uses five different types of algorithms to predict the type of crime that might occur based on time and location. The algorithm involving trees showed that the predicted results is very much closer to the actual results. Thus, the dataset used, provides the maximum correct result with higher accuracy when implemented with different tree classifiers. The stated results in this paper show that Bagging method works best and AdaBoost works least well for predicting crimes using time and location. The results in this paper provides similar results when implemented with tree-based algorithms. Therefore, this paper expects to get more variation in the results when implemented with other classifying algorithms in the future.

VI. REFERENCE

[1] Alkesh Bharati and Dr Sarvanaguru RA.K. (2018). Crime Prediction and Analysis Using Machine Learning.
[2] Ankit Sangani, Vijaya Pinjarkar and Chirag Sampat. (2019). Crime Prediction and Analysis, 2nd International Conference on Advances in Science & Technology.
[3] Ayisheshim Almaw and Kalyani Kadam. (2018). Survey Paper on Crime Prediction using Ensemble Approach, International Journal of Pure and Applied Mathematics, (118 (8), 133-139)
[4] Christian Tabedzki, Amruthesh Thirumalaiswamy and Paul van Vliet. (2018). Yo Home to Bel-Air: Predicting Crime on The Streets of Philadelphia
[5] Clifton Phua, Damminda Alahakoon and Vincent Lee. (2004). Minority Report in Fraud Detection: Classification of Skewed Data, Sigkdd Explorations, (6(1), 51-5).
[6] Hyeon-Woo Kang and Hang-Bong Kang. (2018). Prediction of crime occurrence from multimodal data using deep learning. DOI= https://doi.org/10.1371/journal.pone.0176244
[7] Irina Matijosaitiene, Peng Zhao, Sylvain Jaume and Joseph W. Gilkey Jr. (2018). Prediction of Hourly Effect of Land Use 127 on Crime, International Journal of Geo-Information, 8, 16, https://doi.org/10.3390/ijgi801001
[8] Mookiah, L., Eberle, W. and Siraj, A., (2015), April. Survey of crime analysis and prediction. In The Twenty-Eighth International Flairs Conference.
[9] Al Boni, M. and Gerber, M.S., (2016), December. Area-specific crime prediction models. In 2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)(pp. 671-676). IEEE. https://doi.org/10.1109/ICMLA.2016.0118
[10] Mrinalini Jangra and Shaveta Kalsi. (2019). Naïve Bayes Approach for the Crime Prediction in Data Mining, International Journal of Computer Applications, 178 (4).
[11] Sathyadevan, S. and Gangadharan, S., (2014), August. Crime analysis and prediction using data mining. In 2014 First International Conference on Networks & Soft Computing (ICNSC2014) (pp. 406-412). IEEE.
[12] Varshitha D N, Vidyashree K P, Aishwarya P, Janya T S, K R Dhananjay Gupta and Sahana R. (2017). Paper on Different Approaches for Crime Prediction system, International Journal of Engineering Research & Technology (IJERT), 5(20)
[13] Varvara Ingilevich and Sergey Ivanov. (2018). Crime rate prediction in the urban environment using social factors, Procedia Computer Science, (136, 472–478)