International Journal of Computer Trends and Technology
ISSN: 2231 – 2803 / https://doi.org/10.14445/22312803/IJCTT-V68I12P106

Volume 68 Issue 12, 34-39, December 2020

Original Article

Improving Women's Safety by accelerating Spatio Temporal Crime Prediction of Hotspots

Satvik Shukla[1], Hari Purnapatre[2], Gitalee Jadhav[3], Rasika Gohokar[4]

[1,2,3,4] *Pune Institute of Computer Technology Survey No. 27, Near, Trimurti Chowk, Dhankawadi, Pune, Maharashtra 411043*

Received Date: 16 October 2020
Revised Date: 25 November 2020
Accepted Date: 27 November 2020

Abstract *- Crimes against women are a common social problem affecting the quality of life of women. Crimes could occur everywhere. However, it is common that criminals work on crime opportunities they face in the most familiar areas for them. By providing a machine learning approach to determine the criminal hotspots and find the type, location, and time of committed crimes, we hope to make our community safer for the women living there and the ones who will travel there. With the increase of crimes, law enforcement agencies demand advanced geographic information systems and new machine learning approaches to improve crime analytics and prediction to protect their communities better. We aim at building an alert system for women's safety, using machine learning prediction models. These models will help to achieve a deeper understanding of criminal hotspots. The alert system will function through an Android application that will deliver women alerts if they enter a neighborhood susceptible to danger. The alerts will be based on a static database obtained as an output of the machine learning prediction.*

Keywords *- Spatio-temporal crime prediction, machine learning, the alert system.*

I. INTRODUCTION

Crimes could occur everywhere. However, it is common that criminals work on crime opportunities they face in the most familiar areas for them. By providing a machine learning approach to determine the criminal hotspots and find the type, location, and time of committed crimes, we hope to make our community safer for the women living there and the ones who will travel there. Crimes against women are a common social problem affecting the quality of life of women. With the increase of crimes, law enforcement agencies demand advanced geographic information systems and new machine learning approaches to improve crime analytics and prediction to protect their communities better. Apart from the key features of location and time, predictions with better accuracies can be made if we include extra features. To support this, the dataset that we have used is the Chicago crime data from 2001 to 2018. In combination with this dataset, we are also using the Census

Data - Selected socioeconomic indicators in Chicago, 2008 – 2012. This dataset contains a selection of six socioeconomic indicators of public health significance and a "hardship index," by the Chicago community area, for the years 2007 – 2011.

II. RELATED WORK

There has been an extensive amount of work done based on crimes. This work has involved large amounts of datasets that have primarily worked on two factors of location and time of the crime [1]. These projects' main components have included data pre-processing, analysis, and model building [1]. Data pre-processing consists of the steps of data cleaning, reduction, integration, and transformation. The stage of analysis follows this stage. Statistical analysis is conducted on the attribute values of the dataset. A variety of graphs are created to give a better understanding of the data. Each graph comes up with the percentage of crime occurrences regarding a particular aspect. The algorithms used in this paper are Apriori, Decision trees, and Naive Bayes. To extract frequent patterns of crimes, the Apriori algorithm is applied. Then Naive Bayesian classifier and decision tree classifier is used to build two different classification models for the dataset. The purpose of the classifiers is to predict the potential crime type in a specific location within a particular time in the future. Both the classification models are examined, and the one which gives better accuracy in prediction is chosen. Another set of approaches included in our survey used the data mining approach for crime prediction[3]. The approach is comparing two types of classifications: the K-NN classifier and the Naive Bayes classifier. In the K-NN classifier, two different techniques were performed; the Uniform technique and inverse technique. While in the Naive Bayes, Gaussian, Bernoulli, and Multinomial techniques were tested.

III. PREPARING AND ANALYZING THE DATASET

Preparing the data will consist of cleaning the data. Data cleaning will take care of the missing values in the dataset. Along with the missing values, it will also aim to remove inconsistent and noisy data.

A. Processing data

Pre-processing involved concatenation of the Chicago crime dataset with the Census data. This was followed by replacing empty cells in the dataset with NA values and replacing those NA values with the appropriate measure of central tendencies like the mean. For the process of filtration, we used two filters as follows:

- Filter based on the type of location of the crime: Only crimes that happen in the type of locations like streets, gas stations, sidewalks, and alleys are considered. Crimes that might happen indoors are out of the scope of our problem statement.
- Filter based on crime type: Only crime types that will be directly related to the traveler, like Battery, Assault, Sexual Assault, theft, and sex offense, are considered. Crimes that might not be directly related to travelers, like gambling, prostitution, and narcotics, aren't taken under consideration.

B. Data Analysis

Implementing script to calculate frequencies of distinct values for every attribute. We can then generate a variety of graphs for the visualization. This visualization can further help us to find relationships between criminal hotspots and the social conditions manually.

We explore trends in the general crime rate based on the location, and we plotted graphs as follows:

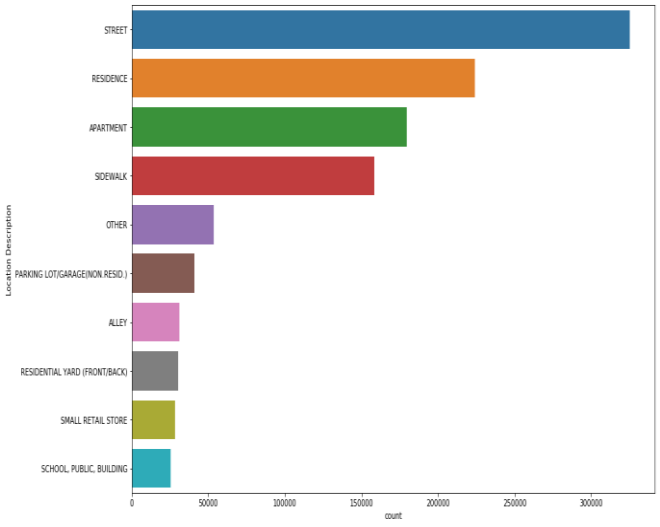

Fig. 1 Histogram of crime based on the type of location

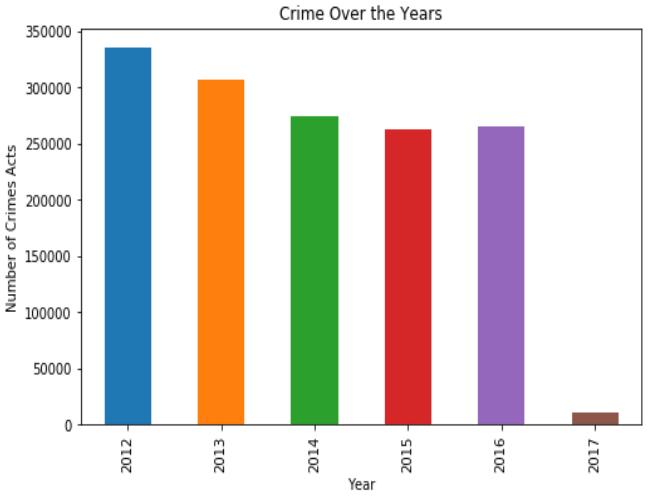

Fig. 2 Histogram showing number of crimes over the years

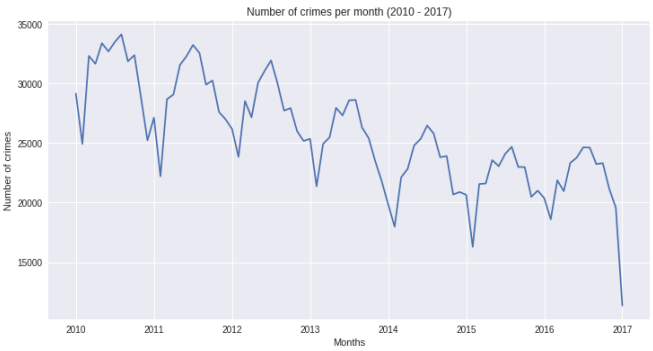

Fig. 3 A line graph showing the frequency of crime over the years

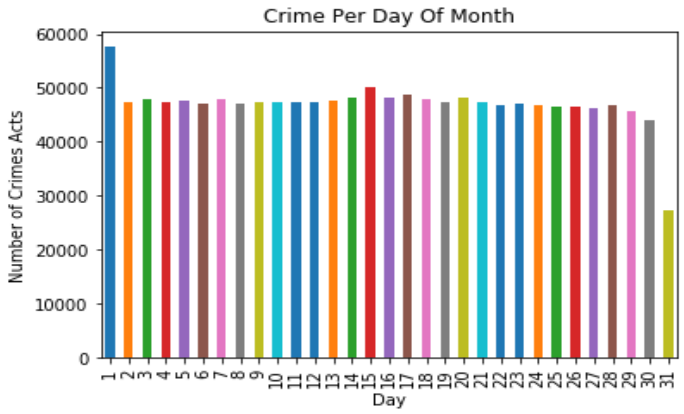

Fig. 4 A histogram showing crimes per days of the month

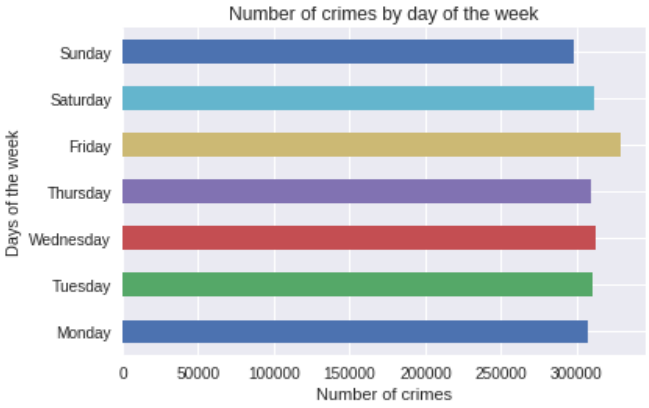

Fig. 5 Histogram showing crimes per day of the week.

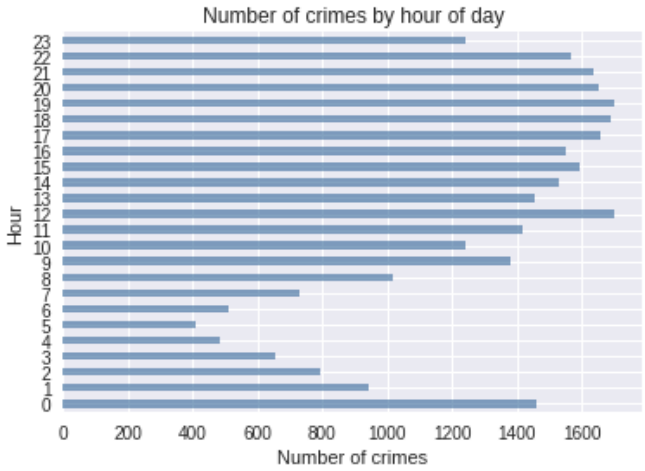

Fig. 6 Number of crimes by the hour of the day

IV. Hotspot Prediction Algorithms and Methodologies

Finding relationships between crime elements highly helps in predicting potential dangerous hotspots at a certain time in the future. Therefore, our proposed approach aims to focus on three main crime data elements: the type of crime, the occurrence time, and the crime location. Apart from these main features, we will also work on extracting supporting features by using the Apriori algorithm and then applying classification methods like Conditional Decision trees to predict potential crime types in a specific location within a particular time. The challenge while predicting the crimes was that the frequency of certain crime types like Battery was overpowering more serious crimes like Assault. This led the usual machine learning models like SVM and K-Means to show frequency-dependent results without considering the weight of the crime.

The objectives are:

- Assign weights and sort of the priority to the classes of the crimes.
- Take both the frequency and the intensity of the crime into consideration while trying to predict the crime people need to be most aware of.
- Show the crime prediction for each pair of community areas and increase prior algorithm accuracy.

A. Apriori Algorithm

Apriori is an algorithm for frequent itemset mining and association rule learning over transactional databases. It proceeds by identifying the frequent individual items in the database and extending them to larger and larger item sets as long as those itemsets appear sufficiently often in the database. The frequent itemsets determined by Apriori can be used to determine association rules which highlight general trends in the database. The pseudo-code for the algorithm is given below for a transaction database T and a support threshold of ϵ. The usual set-theoretic notation is employed, though note that T is a multiset. C is the candidate set for level k. At each step, the algorithm is assumed to generate the candidate sets from the preceding level's large itemsets, heeding the downward closure lemma. Count [c] accesses a field of the data structure representing candidate set c, which is initially assumed to be zero. Many details are omitted below; usually, the most important part of the implementation is the data structure used to store the candidate sets and count their frequencies.

Apriori(T, ε)

$L_1 \leftarrow$ { large 1 - itemsets }

$K \leftarrow 2$

While $L_{k-1} \neq 0$

$C(k) \leftarrow \{c=a \cup \{b\} \mid a \in L_{k-1} \wedge b \notin a, \{s \sqsubseteq c \mid |s| = k-1\} \sqsubseteq L_{k-1}$

for transactions $t \in T$

$D_t \leftarrow \{c \in C_k \mid e \sqsubseteq t\}$

for candidates $c \in D_t$

count[c] ← count[c] + 1

$L_k \leftarrow \{c \in C_k \mid count[c] >= e\}$

$K \leftarrow K + 1$

return $\cup_k L_k$

As some crime types are more frequent, they will overpower other less frequent crimes, and we will not obtain a satisfactory number of hotspots for the less frequent crime types. Thus we used Apriori separately for all unique crime types in our dataset. That way, the crime type was kept constant on the LHS, and we found the most frequent patterns in location and time as a combination. Chicago is divided into various crime beats by their police department. Thus we have also used the beat number as a qualifier for the location. We have aimed at identifying 10% beats as hotspots. We adjusted the support and confidence for all crime types accordingly to yield those many numbers of hotspots.

B. Conditional Decision Trees

Similar to traditional decision trees, conditional inference trees also recursively partition the data by performing a univariate split on the dependent variable. However, what makes conditional inference trees different from traditional decision trees is that conditional inference trees adapt the significance test procedures to select variables rather than selecting variables by maximizing information measures (e.g., Gini coefficient), As an improvement of the recursive partitioning procedure used in traditional decision trees, conditional inference trees separate the variable selection from the splitting procedure. This results in basically three steps in the conditional inference tree procedure.

- The first one concerns variable selection.
- The second one chooses the splitting methodology.
- The last one is the recursive application of the first two steps.

Since all types of crime do not have equal weight, for example, sexual Assault will have more
weight over Battery or theft, the crimes that need to be predicted to be delivered as a part of the alert message will vary. Conditional inference trees are very useful in this scenario as we can assign weights to the classes that need to be predicted.

1) Arabic numeral followed by a right parenthesis. The level-3 heading must end with a colon. The body of the level-3 section immediately follows the level-3 heading in the same paragraph. For example, this paragraph begins with a level-3heading.

VI. IMPLEMENTATION

The implementation stage was further divided into the following two stages:

- Identification of hotspots.
- Prediction of the possible crimes in the hotspots identified.

For the Identification of hotspots, Apriori helped us find the most frequent relations between the community area and the time interval on the one hand and the crime type on the other from all the crime data points we had in the Chicago dataset. While going about using the Apriori algorithm, we kept the crimes constant on the RHS to find out which all space-time pairs had the most support and confidence. Thus, we run the Apriori algorithm thrice for each crime separately.
The support and confidence were adjusted to get 20% of the combinations to be identified as hotspots. This ratio can be tweaked as and when required.

The output generated in the form of the rules of the Apriori algorithm taken and manually
transformed into a simple CSV file with the following fields:

1) Community area number
2) Hour
3) Crimes

Table 1: Support and confidence for crimes.

Sr. no	Crime	Support	Confidence
1	Sex Offence	0.00002	0.2
2	Assault	0.00001	0.48
3	Battery	0.002	0.50

The dataset was divided into testing data and training data. The output obtained from the Apriori algorithm was used as the testing data to predict the crimes that can happen in the hotspots. The dataset was divided into testing data and training data. The output obtained from the Apriori algorithm was used as the testing data to predict the crimes that can happen in the hotspots. The challenge while predicting the crimes was that the frequency of certain crime types like Battery was overpowering other important crimes like Assault. This led the usual machine learning models like SVM and K-Means to Battery as the crime in most cases, which was undesirable. A solution to this challenge would be to assign weights and thus a sort of priority to the classes of the crimes, and that's exactly what we opted to do with the help of the Conditional Decision Tree algorithm. The weights assigned to the crimes were as follows:

Table 2: Weights assigned to crimes

Sr. no	Support	Confidence
1	Sex Offence	3
2	Assault	2
3	Battery	1

VII. RESULTS AND EVALUATIONS

Once the hotspots were identified using Apriori algorithm, we needed to find an algorithm that predicts the crimes women should beware of. In this process, we carried out the following algorithms :

A. SVM

SVM classified the hotspots in our dataset obtained from the Apriori algorithm into various crime types. In machine learning, support vector machines are supervised learning models with associated learning algorithms that analyze data for classification and regression analysis. Given a set of training examples, each marked as belonging to one or the other of two categories, and an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier.

B. MULTIPLE LINEAR REGRESSION

Multiple linear regression attempts to model the relationship between two or more explanatory variables and a response variable by fitting a linear equation to observed data.

Every value of the independent variable x is associated with a value of the dependent variable y. The population regression line for p explanatory variables x1, x2, ... , XP is defined to be y = 0 + 1x1 + 2x2 + ... + pxp. This line describes how the mean response y changes with the explanatory variables. The observed values for y vary about their means y and are assumed to have the same standard deviation. The fitted values b0, b1, and bp estimate the parameters 0, 1, and p of the population regression line.

Since the observed values for y vary about their means y, the multiple regression model includes a term for this variation. In words, the model is expressed as DATA = FIT + RESIDUAL, where the "FIT" term represents the expression 0 + 1x1 + 2x2 + ... pxp. The "RESIDUAL" term represents the deviations of the observed values y from their means y, normally distributed with mean 0 and variance. The notation for the model deviations are. Formally, the model for multiple linear regression, given n observations, is yi = 0 + 1xi1 + 2xi2 + ... pxip + i for i = 1,2, ... n.

C. KNN

The k-nearest neighbors' algorithm (k-NN) is a non-parametric method used for classification and regression in pattern recognition.[1] In both cases, the input consists of the k closest training examples in the feature space. The output depends on whether k-NN is used for classification or regression:

In k-NN classification, the output is a class membership. An object is classified by a plurality vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors (k is a positive integer, typically small). If k = 1, then the object is assigned to that single nearest neighbor's class.

In k-NN regression, the output is the property value for the object. This value is the average of the values of k nearest neighbors.

D. CONDITIONAL DECISION TREE

Similar to traditional decision trees, conditional inference trees also recursively partition the data by performing a univariate split on the dependent variable. However, what makes conditional inference trees different from traditional decision trees is that conditional inference trees adapt the significance test procedures to select variables rather than selecting variables by maximizing information measures (e.g., Gini coefficient).

Table 3. Algorithm Accuracy

Algorithm	**Accuracy**
SVM	61.6%
Multiple Linear Regression	86.78
KNN	55.5%
Conditional Decision tree	82%

VIII. CONCLUSION

We have successfully carried out a literature survey that has given us an insight into the various existing and proposed methodologies. We have also performed a requirement analysis that has covered our research's functional and non-functional requirements. This has also given us a clearer idea of the kind of dataset we will be dealing with. We have also studied algorithms and found the ones that are best suitable for our problem statement. Using these algorithms, we have built and trained machine learning models that have identified hotspots and their corresponding crimes.

IX. REFERENCES

[1] Tahani Almanie, Rsha Mirza and Elizabeth Lor, Crime Prediction Based On Crime Types And Using Spatial And Temporal Criminal Hotspots, International Journal of Data Mining & Knowledge Management Process (IJDKP).5(4), 2015.

[2] Fondazione Bruno Kessler, Trento (Italy)]MIT Media Lab, Cambridge, MA (United States) Once Upon a Crime: Towards Crime Prediction from Demographics and Mobile Data.ICMI '14: Proceedings of the 16th International Conference on Multimodal Interaction (arxiv.org/abs/1409.2983)10 Sep 2014

[3] Mohammad Al Boni and Matthew S. Gerber, Department of Systems and Information Engineering, University of Virginia, Charlottesville, Virginia, Area-Specific Crime Prediction Models,2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA) Dec. 18, 2016

[4] Noora Abdulrahman and Wala Abedalkhader, Knn Classifier and Naive Bayes Classifier for crime prediction in San Francisco context International Journal of Database Management Systems (IJDMS) 9(4) 2017.

[5] Ying-Lung Lin, Tenge-Yang Chen, Using Machine Learning to Assist Crime Prevention, 6th IIAI International Congress on Advanced Applied Informatics 1 (2017) 1029-1030

[6] Surendiran,R., and Alagarsamy,K., 2013. "Privacy Conserved Access Control Enforcement in MCC Network with Multilayer Encryption". SSRG International Journal of Engineering Trends and Technology (IJETT), 4(5), pp.2217-2224.

[7] Ginger Saltos, Ella Haig, Exploration of crime prediction using data mining, International Journal of Information Technology and Decision Making International Journal of Information Technology & Decision Making 16(05) (2017) 1155-1181.