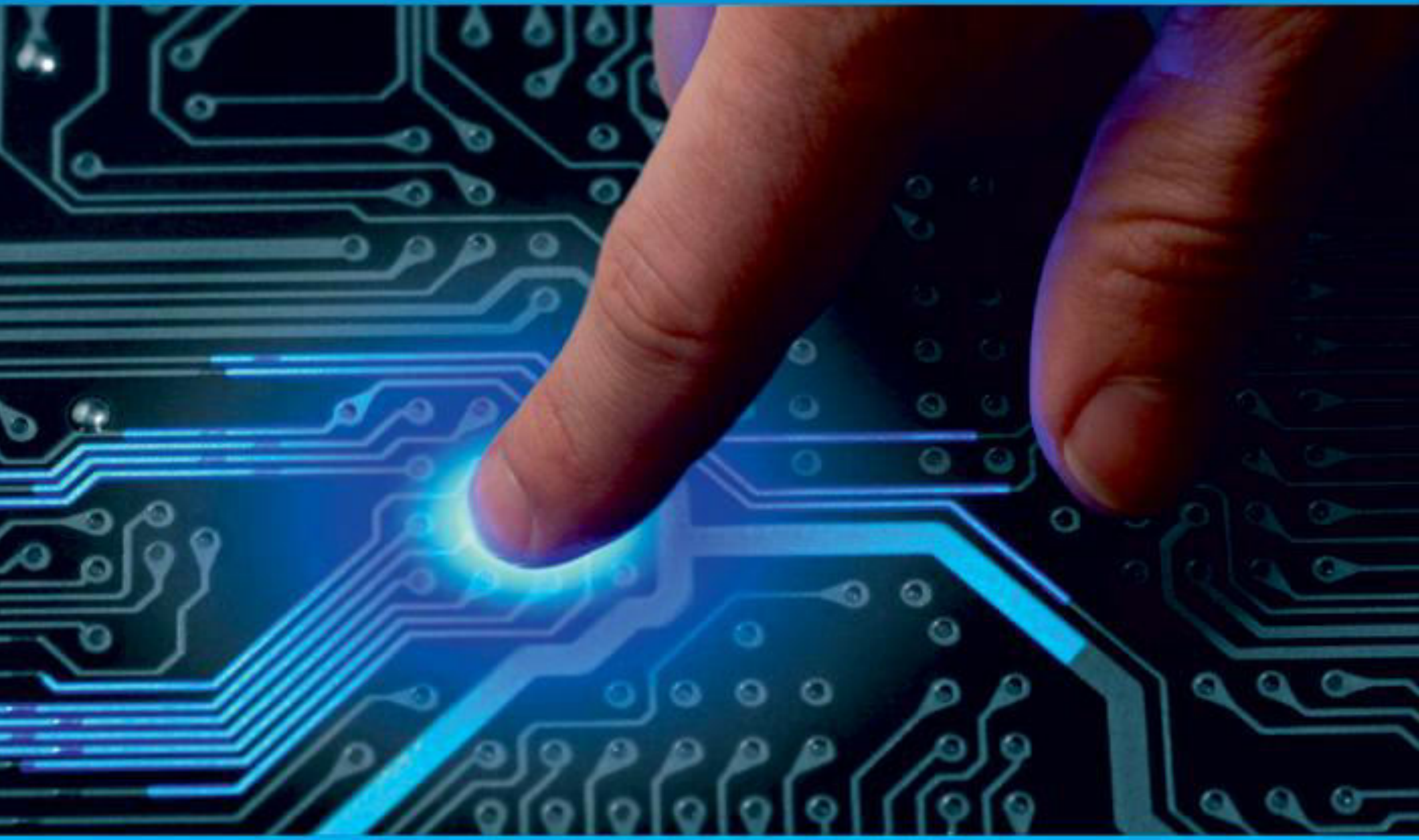
IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 9, Issue 6, June 2021

INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA

Impact Factor: 7.542

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542 |

|| Volume 9, Issue 6, June 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0906281 |

Crime Hotspot Prediction using Machine Learning

Mr.P.Siva Prasad, Valeti Ramya, Vemuri Sravanya, Yenugula Bhanoday ,Vemuri Krishna Teja

Assistant Professor, Department of Information Technology, Vasireddy Venkatadri Institute of Technology, Nambur, Guntur, Andhra Pradesh, India

U.G. Students, Department of Computer science and Engineering, Vasireddy Venkatadri Institute of Technology, Nambur, Guntur, Andhra Pradesh, India

ABSTRACT: Crimes are an increasing danger to mankind. Prediction of hotspots plays a vital role in case investigation and to aware the citizens about dangerous locations. The project focuses on finding spatial and temporal criminal hotspots. It analyses two different real-world crimes datasets for Denver and Los Angeles and provides a comparison between the two datasets through a statistical analysis supported by several graphs. Apriori algorithm is used to produce interesting frequent patterns for criminal hotspots. In addition, Decision Tree classifier and Naïve Bayesian classifier are used in order to predict potential crime types. To further analyze crimes datasets, an analysis study by combining our findings of Mumbai crimes dataset with its demographics information in order to capture the factors that might affect the safety of neighbourhoods. The results of this solution could be used to raise people's awareness regarding the dangerous locations and to help agencies to predict future crimes in a specific location within a particular time.

I. INTRODUCTION

Crimes are a common social problem affecting the quality of life and the economic growth of a society [1]. It is considered an essential factor that determines whether or not people move to a new city and what places should be avoided when they travel [2]. With the increase of crimes, law enforcement agencies are continuing to demand advanced geographic information systems and new data mining approaches to improve crime analytics and better protect their communities [3].

Although crimes could occur everywhere, it is common that criminals work on crime opportunities they face in most familiar areas for them [4]. By providing a data mining approach to determine the most criminal hotspots and find the type, location and time of committed crimes,we hope to raise people's awareness regarding the dangerous locations in certain time periods.

Therefore, our proposed solution can potentially help people stay away from the locations at a certain time of the day along with saving lives. In addition, having this kind of knowledge would help people to improve their living place choices. On the other hand, police forces can use this solution to increase the level of crime prediction and prevention. Moreover, this would be useful for police resources allocation. It can help in the distribution of police at most likely crime places for any given time, to grant an efficient usage of police resources [5]. By having all of this information available, we hope to make our community safer for the people living there and also for others who will travel there.

II. MACHINE LEARNING

Machine Learning is the study of computer algorithms that improve automatically through experience and by the use of data.[1]It is seen as a part of artificial intelligence. Machine learning algorithms build a model based on sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to do so.[2] Machine learning algorithms are used in a wide variety of applications, such as in medicine, email filtering, speech recognition, and computer vision, where it is difficult or unfeasible to develop conventional algorithms to perform the needed tasks.[3]

A subset of machine learning is closely related to computational statistics, which focuses on making predictions using computers; but not all machine learning is statistical learning. The study of mathematical optimization delivers methods, theory and application domains to the field of machine learning.[4] Data mining is a related field of study, focusing on exploratory data analysis through unsupervised learning.[5] In its application across business problems, machine learning is also referred to as predictive analytics.

III. RELATED WORK

Reference paper 1

Authors implemented Machine leaning algorithms like Linear Regression, Additive Regression, and Decision Stump algorithms using the same limited set of features, on the groups and crime un-normalized dataset to conduct a relative revision among the violent crime forms from this particular dataset and actual crime statistical data for the state of Mississippi that has been delivered by neighborhoodscout.com. The scope of this project was to prove how operative and precise machine learning algorithms can be at forecasting violent crimes, there are other applications of data mining in the realm of law enforcement such as decisive criminal "hot spots," creating criminal profiles, and learning crime trends. Exploiting these applications of data mining can be a long and dreary process for law enforcement officials who have to sift through huge capacities of data.

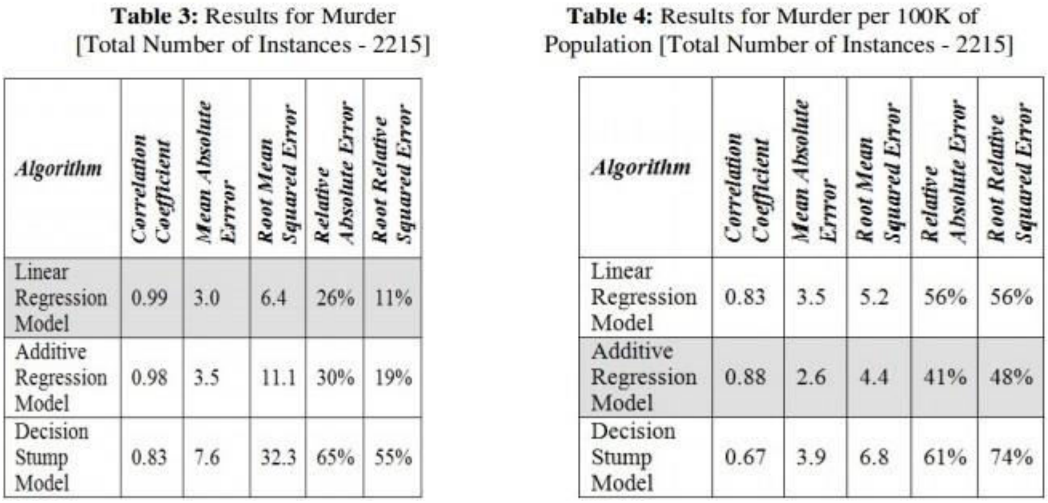


Table 3: Results for Murder [Total Number of Instances - 2215]

Algorithm	*Correlation Coefficient*	*Mean Absolute Error*	*Root Mean Squared Error*	*Relative Absolute Error*	*Root Relative Squared Error*
Linear Regression Model	0.99	3.0	6.4	26%	11%
Additive Regression Model	0.98	3.5	11.1	30%	19%
Decision Stump Model	0.83	7.6	32.3	65%	55%

Table 4: Results for Murder per 100K of Population [Total Number of Instances - 2215]

Algorithm	*Correlation Coefficient*	*Mean Absolute Error*	*Root Mean Squared Error*	*Relative Absolute Error*	*Root Relative Squared Error*
Linear Regression Model	0.83	3.5	5.2	56%	56%
Additive Regression Model	0.88	2.6	4.4	41%	48%
Decision Stump Model	0.67	3.9	6.8	61%	74%

Figure 1: Existing System

IV. PROPOSED SYSTEM

Problem statement

Our study aims to find spatial and temporal criminal hotspots using a set of real- world datasets of crimes. We will try to locate the most likely crime locations and their frequent occurrence time. In addition, we will predict what type of crime might occur next in a specific location within a particular time. Finally, we intend to provide an analysis study by combining our findings of a particular crimes dataset with its demographics information.

Steps in preprocessing

We strongly believe that finding relationships between crime elements could highly help in predicting potential dangerous hotspots at a certain time in the future. Therefore, our proposed approach aimed to focus on three main elements of crimes data, which are the type of crime, the occurrence time and the crime location. We tried to extract all possible interesting frequent patterns based on the crime variables. Then, we applied some classification methods in order to predict potential crime types in a specific location within a particular time.

We performed the following preprocessing steps on the two datasets:

Data Cleaning

There are some missing values in some attributes such as last_occurance_date andincident_address in Denver dataset. However, we found that all attributes containing missing values are not of our key attributes. Therefore, we did not need to clean them. All key attributes in (Table 1 and Table 2) were completed with cleaned values in both datasets. In addition, we did not found any noisy or inconsistent values in these attributes.

Data Reduction

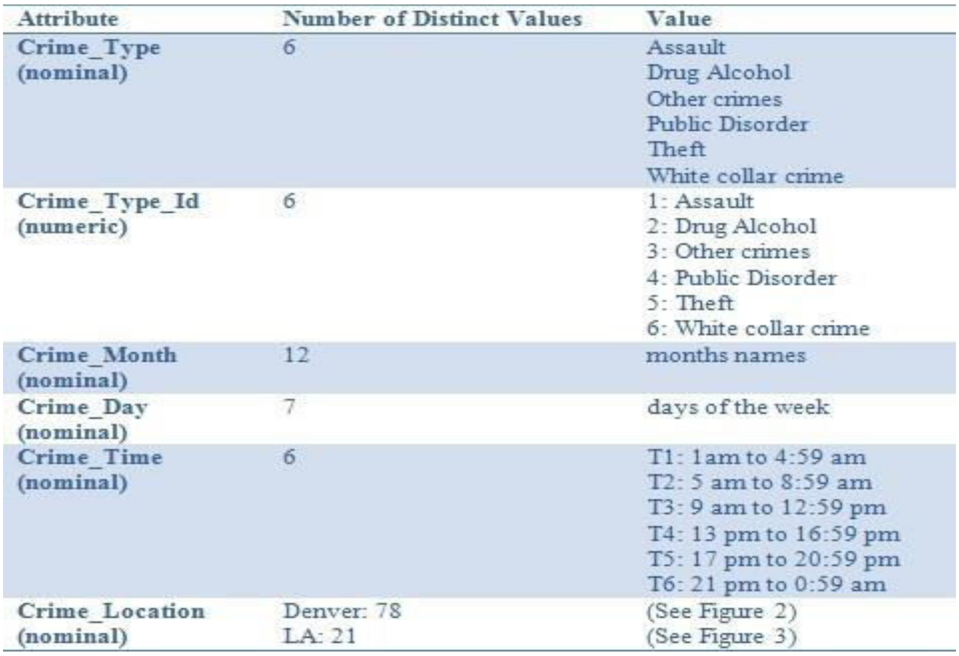
For both crime datasets, we needed to apply data reduction. We implemented dimensionality reduction using attribute subset selection. For example, among the available 19 attributes in Denver crimes dataset, we just selected four of them. The selected attributes are the related ones or the key attributes for our mining purpose (see Table 1). We removed all the other irrelevant attributes from the dataset.

Data Integration

We performed several steps of data integration for our datasets. First, to avoid different attribute naming, we unified the key attribute names for both crime datasets as follow: Crime_Type, Crime_Date, and Crime_Location. Crime_Location represents the neighborhood attribute for Denver dataset whereas the Area attribute for Los Angeles dataset. Our mining study requires analyzing the date and time info on different granularities. Therefore, we used the Crime_Date attribute, which contains date and time crime info, to generate three more attributes: Crime_Month, Crime_Day, and Crime_Time. We adopted the military time system, and we considered the hour part without paying attention to the minutes to get more of frequent patterns. In addition, we initiated Crime_Type_Id attribute to give an id for each of the 14 crime categories (See Table1). We used this attribute for both datasets to get integrated crime types.

Data Transformation and Discretization

We finished our data integration process by having 24 different distinct values for the Crime_Time attribute and 14 types for the Crime_Type attribute. We realized that it is necessary to reduce the diversity of these two attribute values. Thus, we applied data transformation to both attributes by mapping their values to fall within smaller groups. Our goal was to get more frequent patterns and to increase the model accuracy. For the crime types feature, we minimize the type list by grouping them into six new types. For the crime time feature, we mapped its values into 4-hour intervals.Table1 illustrates the resulted attributes after data preprocessing.



Attribute	Number of Distinct Values	Value
Crime_Type (nominal)	6	Assault Drug Alcohol Other crimes Public Disorder Theft White collar crime
Crime_Type_Id (numeric)	6	1: Assault 2: Drug Alcohol 3: Other crimes 4: Public Disorder 5: Theft 6: White collar crime
Crime_Month (nominal)	12	months names
Crime_Day (nominal)	7	days of the week
Crime_Time (nominal)	6	T1: 1am to 4:59 am T2: 5 am to 8:59 am T3: 9 am to 12:59 pm T4: 13 pm to 16:59 pm T5: 17 pm to 20:59 pm T6: 21 pm to 0:59 am
Crime_Location (nominal)	Denver: 78 LA: 21	(See Figure 2) (See Figure 3)

Table 1 : Data after preprocessing

Models Building

In order to extract frequent patterns from Denver and Los Angeles crimes, we applied the Apriori algorithm on both datasets. Then we use Naïve Bayesian classifier and decision tree classifier to build two different classification models for each dataset. The purpose of the classifiers is to predict the potential crime type in a specific location within a particular time in the future. We aimed to examine every model then choose the model that gives the best accuracy in prediction. In this section, we provide a brief description of each model used.

Apriori Algorithm

Apriori is one of the basic algorithms for mining frequent patterns. It scans the dataset to collect all itemsets that satisfy a predefined minimum support. Our goal of using this model is to find all possible crime frequent patterns regardless of the committed crime type. We wanted to come up with a list of all crime hotspots along with its related frequent time. Hence, we implemented the algorithm on location and time features and excluded the crime type feature. Additionally, to obtain more frequent patterns we applied constraint-based mining by restricting the extraction process on the frequent patterns having this formula of three specificitemsets (Location, Day, Time).

We implemented this model using an open source tool [13]. We conducted multiple experiments using different minimum support values for each dataset. Then we selected the optimum choice. For Denver, the minimum support value was 0.0012, which corresponds to 277 absolute frequencies. For Los Angeles, the minimum support value was 0.0018, which corresponds to 354 absolute frequencies.

Naïve Bayesian Classifier

Naïve Bayesian classifier is a supervised learning algorithm, which is effective and widely used. It is a statistical model that predicts class membership probabilities based on Bayes' theorem (Formula 3). It assumes the independent effect between attribute values. While our selected crime features have an independent effect on each other, this classifier was an ideal choice $P(H|X) = P(X|H)\,P(H)/\,P(X)$

We constructed this model using Scikit–Learn that provides a set of open source data-mining tools for Python. We applied Multinomial Naïve Bayes, which is used for multinomial distributeddata that conforms to the categorical features in our datasets. The crime features contain (month, day, time, location) of the crime while we selected the crime type to represent the class label. We randomly divided the dataset into 80% of data as a training set and 20% of data as a testing set. We trained the same classifier on the training data for each of Denver and Los Angeles datasetsto obtain two different models ready for crime type prediction in each of the two cities.

Decision Tree Classifier

Decision Tree classifier is our second used supervised learning algorithm. It creates a model to predict the class label values by learning simple decision rules implied from the data features.

We created this model for both datasets using Scikit–Learn another library tool allocated for decision tree induction. To measure the quality of the split, we applied the entropy function for the information gain on Los Angeles training dataset. Since the generated tree was complex, we restricted the decision tree to have ten maximum leaf nodes. The tree shows that the Time attribute is selected as the root node to split the data.

V. RESULTS

In this section, we summarize the key results that we obtained from applying the Apriori and Bayesian classifier models on the two datasets. Then, we provide an analysis study through combining our findings of Denver crimes dataset with its demographics information.

Crime Frequent Hotspots

The first goal of our study was finding spatial and temporal criminal hotspots. We have successfully achieved this goal using Apriori algorithm on both Denver and Los Angeles datasets. We have extracted all the interesting patterns based on our predefined thresholds. We found that Denver has 62 interesting frequent patterns while Los Angeles has 59 patterns. Table 4 and Table 5 report our Apriori results for Denver and Los Angeles crime frequent patterns. The frequent itemsets ordered by the location, day of the week, and the time period.

With these different frequent itemsets, we are able to conclude the most likely crime locations along with their frequent occurrence day and time. Table 6 indicates that Five-Point, Capitol Hill, CBD, Montebello, Union Station, Stapleton,and Westwood are the hotspots that have most crimes frequent patterns in Denver. It is obvious that Five-Point has the largest number of patterns compared to other locations while CBD comes next. In addition, we can find that Wednesday is the peak day of crimes occurred in CBD. It is also interesting to notice that Union Station has frequent patterns only on weekend days (Friday, Saturday, Sunday) four hours before and after midnight. In Los Angeles, we can see that most likely crimes happen at 77th Street, Southwest, Pacific, N Hollywood,Southeast, Northeast, and Van Nuys respectively. Among all crime patterns, the highest frequent one occurs in 77th Street on Monday around 9pm to 1am. Both 77th Street and Southwest areas have crimes everyday, and their crimes are more likely to happen from 8 am to midnight.

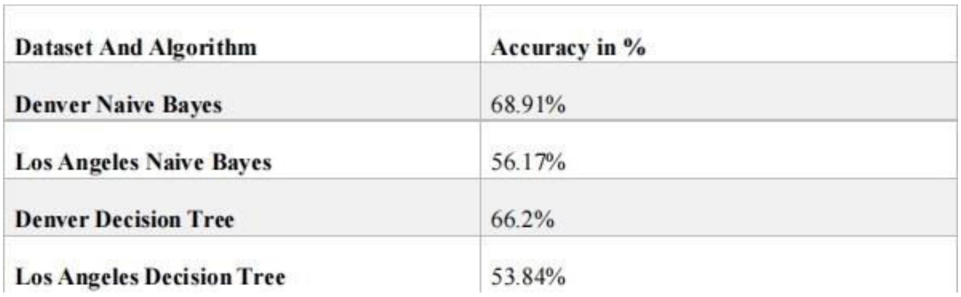
Crime Prediction

The second target for our study was to predict the crime type that might occur in a specific location within a particular time. The Bayesian classifier enabled us to reach this target with a reasonable accuracy. To predict an expected crime type, you need to provide four related features of the crime. The required features are: the occurrence month, the occurrence day of the week,the occurrence time and the crime location. All features can be submitted in their nominal values. The provided occurrence time should be in the form of time period interval from T1 to T6 (See Table 3). For Denver, the location has to be one of its 78 neighborhoods (See Figure 2). For Los Angeles, the location should be one of its 21 areas (See Figure 3). Every given result is a number from 1 to 6 that indicates the predicted crime type for a given set of crime features. Table 3 gives the corresponding crime type for each number.

VI. CONCLUSION AND FUTURE WORK

We generated many graphs and found interesting statistics that showed the baseline to understand Denver and Los Angeles crimes datasets. Then, we applied Apriorialgorithm to find frequent crime patterns in both cities. After that, we applied Decision Tree and Naïve Bayesian classifiers to help predicting future crimes in a specific location within a particular time. We achieved 51% of prediction accuracy in Denver and 54% prediction accuracy in Los Angeles. Finally, we provided an analysis study by combining our findingsof Denver crimes' dataset with its demographics information. We aimed to further understand our models' findings and to capture the factors that might affect the safety of neighborhoods. As a future extension of our work, we plan to apply more classification models to increase crime prediction accuracy and to enhance the overall performance. It is also a helpful extension for our study to consider the income information for neighborhoods in order to see if there are relationships between neighborhoods income level and their crime rate. Moreover, we intend to analyse Los Angeles demographics information with its crime findings. Furthermore, we want to study other crimes datasets from new cities along with their demographics datasets.

Last but not least, we hope by publishing this paper starting a trend of crimes prediction, which can help law enforcements and keep our community safer for everyone.



Dataset And Algorithm	Accuracy in %
Denver Naive Bayes	68.91%
Los Angeles Naive Bayes	56.17%
Denver Decision Tree	66.2%
Los Angeles Decision Tree	53.84%

Table2: Results

REFERENCES

1. Bogomolov, B. Lepri, J. Staiano, N. Oliver, F. Pianesi and A. Pentland, 'Once Upon a Crime: Towards Crime Prediction from Demographics and Mobile Data', CoRR, vol. 14092983, 2014.
2. R. Arulanandam, B. Savarimuthu and M. Purvis, 'Extracting Crime Information from Online Newspaper Articles', in Proceedings of the Second Australasian Web Conference - Volume 155, Auckland, New Zealand, 2014, pp. 31-38.
3. Buczak and C. Gifford, 'Fuzzy association rule mining for community crime pattern discovery', in ACM SIGKDD Workshop on Intelligence and Security Informatics, Washington, D.C., 2010, pp. 1-10.
4. M. Tayebi, F. Richard and G. Uwe, 'Understanding the Link Between Social and Spatial Distance in the Crime World', in Proceedings of the 20th International Conference on Advances in Geographic Information Systems (SIGSPATIAL '12), Redondo Beach, California, 2012, pp. 550-553.
5. S. Nath, 'Crime Pattern Detection Using Data Mining', in Web Intelligence and Intelligent Agent Technology Workshops, 2006. WI-IAT 2006 Workshops. 2006 IEEE/WIC/ACM International Conference on, 2006, pp. 41,44.
6. Crimereports.com, 2015. [Online]. Available: https://www.crimereports.com. [Accessed: 20- May- 2015].
7. S. Chainey, L. Tompson and S. Uhlig, 'The Utility of Hotspot Mapping for Predicting Spatial Patterns of Crime', Security Journal, vol. 21, no. 1-2, pp. 4-28, 2008.
8. Data.denvergov.org, 'Denver Open Data Catalog: Crime', 2015. [Online]. Available:

http://data.denvergov.org/dataset/city-and-county-of-denver-crime. [Accessed: 20- May- 2015].

9. Imgh.us, 2015. [Online]. Available: http://imgh.us/neighborhood_map.jpg. [Accessed: 20- May- 2015].
10. O. Knowledge, 'Crime — Datasets - US City Open Data Census', Us-city.census.okfn.org, 2015. [Online]. Available: http://us-city.census.okfn.org/dataset/crime-stats. [Accessed: 20- May- 2015].
11. Laalmanac.com, 'City of Los Angeles Planning Areas Map', 2015. [Online]. Available: http://www.laalmanac.com/LA/lamap3.htm. [Accessed: 20- May- 2015].

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

☎ 9940 572 462 6381 907 438 ✉ ijircce@gmail.com

www.ijircce.com

Scan to save the contact details