Crime Analytics: Exploring Analysis of Crimes through R Programming Language

Pilar M. Fandino, Jose B. Tan Jr.

Technological Institute of the Philippines, Manila, Philippines.
Email: pilarmfandino@gmail.com; jose.tanjr@yahoo.com

Abstract - The complexity of geographical crime patterns creates challenges for many laws enforcing agencies as well for the community. Crime analytics is an intelligent crime analysis designed to aggregate crime related data. We use different algorithm in Machine Learning to collect data sets. Our focus in this study is to design a crime analysis system to identify trends and patterns. This study provides solutions to reveal valuable records that can be used effectively for analyzing and recording the information. The application of R programming language gives a new way to connect to enormous volumes of police crime data, where R streamlines the processing and interpretation of crime analysis. The study used the comprehensible histogram to show crime rates in every district. R explores predictions related to crime patterns, which depicts the crime committed. It is a tool that can help law enforcement effort to set various stages of development to check the potential problems before they become disastrous. It is needed to emphasize that R can help in the analysis of data concerning small units or districts.

Keywords - crime analytics, data mining. Aggregate, machine learning

I. INTRODUCTION

A crime-free nation is an important factor in fostering investment and economic growth. A record shows that common crimes have a record of an average of 5.6% (SWS 2017). This resulting 2 points below from the previous percentage which is 7.6% in 2016 record property crime rate (SWS 2017). Data shows there was steady drop rate. A total of 227, 757 crimes were recorded from January to June lower by 47, 945 compared to 275, 702 crimes (Albayalde 2017).

The evolution of technology leads to digital information. This advancement leads the interest and the capacity to keep safe the community. These massive volume of data helps us to fight crime. It will give way to predict possible behavior or patterns of crime. Classifying threats in the community to prevent the future attacks or terrorizing.

Studies by Thomas Rich and by Cynthia Mamalian and Nancy LaVigne of the National Institute of Justice have found that the majority of departments using crime mapping are creating automated pin maps and generating hotspot maps. Along a range of difficulty, pin maps and hot spots analysis are the low end. For more advanced analysis, some knowledge of cartography and geographic information system is required. The complexity of these latest innovations may well prevent their adoption by all with the most technologically advanced departments.

Data mining denotes lineage of data or knowledge discovery and analysis of patterns into useful information. A tool that can be used for crime detection. The emersion of space and time enhance the detecting crime patterns, these patterns give situational awareness among the security agents and prevent the potential problematic areas (Neil and Gorr, 2017). Crime patterns is very useful to locate a possible committed crime to happen. Using the historical data, it can preempt measure may applied to prevent the next crime. Detecting crime patterns is challenging. Data mining is one of the evolving fields that can play with the large volume of datasets. The crime intensity of a location, in real-time basis (Bolla, 2014) should gain insights to aggregate knowledge for it.

Data mining in crime literature has rapidly increase for the past years, it turns out to be mandatory to improve an impression to lessen the crime rate. This systematic review focuses on techniques and technologies used in the previous studies of data mining in crime. It is classified into different categories and presented using visualizations. Crime data mining research indicates challenges related to it.

The main purpose of this study is to identify the patterns and behavior of accuracy of existing crime with the use of R Programming. This will determine the areas of opportunities to improve, algorithm and procedures used. It can also assist us to determine the best fit process to be use for further study.

II. EXISTING STUDIES

To determine and understand the behavior of the algorithm used from the different literature presented. This shows the strength and weaknesses for the study it may applies. The following are the existing work:

The use of clustering and classification in data mining algorithms to detect the criminal and crime spot. Data stored in databases to classify the activities of criminals. The identification of criminals will be based on crime spot and witnesses. These criminal “hotspots” will help the law enforcers to avoid or to lessen criminals. [Sukanya, 2012].

Spatial is related to space, this method helps us to identify the patterns and the threshold. These patterns predict the potential crime spots. The assessment of the
datasets through statistical analysis conducted by the Apriori algorithm and produces interesting frequent results of criminal hotspots pattern. The decision tree classifier and Naïve Bayesian use to predict potential types of crimes. The paper brings together the analysis of the study through the combination of the demographics information datasets of crime to factors the capture information that affects the neighborhood. Those findings will give a result to this solution and the be used as the awareness of the people regarding the dangerous locations or to help the agencies to predict the future crimes happen in specific locations.

The K-mean clustering and geographical information system (GIS) also use to detect the crime hotspots over the use of data mining tools to create a structure. This unique feature of this of this study is the application of factor. The mapping of the information serves as the foundation of the parameters of the crime. The analysis of the factor application is used to uncover the hidden structure of the data that is present yearly. This procedure can help to predict where the crime is located.

Agarwal et al. used the rapid miner tool for analyzing the crime rates and anticipation of crime rate using different data mining techniques. Their work done is for crime analysis using the K-Means Clustering Algorithm. The main objective of their crime analysis is to extract the crime patterns, predict the crime based on the spatial distribution of existing data and detection of crime. Their analysis includes the tracking homicide crime rates from one year to the next. (Jyoti Agarwal, Renuka Nagpal and Rajni Schgal, 2013).

Most of the literature introduce a method to map out crime in various locations. It provides database, crime parameters and structure of the correlations among the defining variables.

III. ALGORITHMS USED AND THEIR ANALYSIS

The following table I shows the systematic review of literature, algorithm used by the studies presented, their strength and weaknesses and their findings.

<table>
<thead>
<tr>
<th>Authors</th>
<th>Algorithm/Architecture/Framework Used</th>
<th>Strength</th>
<th>Weaknesses</th>
<th>Findings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Detecting Hot Spots on Crime Using Data Mining and Geographical Information System</td>
<td>K-Mean Clustering Factor Analysis</td>
<td>The unique feature of this study is the application of factor, k-mean clustering and Geographical Information System (GIS) analysis as data mining tools to develop the hidden structure present in the data for each year.</td>
<td>Factor analysis is extended with the techniques of Varimax/Quartimax criterion for orthogonal rotation. Even though the results obtained by both the criterions were very similar, the Varimax rotation provided relatively better clustering of crime data. Consequently, only the results of Varimax rotation are reported here. We have decided to retain 76</td>
<td>It shows that with only two of the seven HIMs it is possible to have a position independent object classification algorithm that allows differentiating and classifying</td>
</tr>
</tbody>
</table>

Table I clearly depicts the different systematic review of literature and some studies that prove the strength, weaknesses, and findings of the different studies found. It was also indicated the application to where it was apply and the conclusion in which it further discussed the output of the study. It was noted as well some of the disadvantages of the different studies which help the proponent to identify as to what algorithm and application should she used and why this kind of application be used.

IV. METHODOLOGY USED

This study allows the researcher to explore the data mining using the R. R programming is an open source program that can be use looking in to the data. For this research the proponent use crime data from Chicago as the dummy data, to make analysis more manageable by applying the historical data. Data Mining is a new approach in data collection. This algorithm design composed of
different sections, such as data mining, Extraction and Classification, Pattern Identification, Prediction and Visualization.

Extraction and Classification are used to get the data source, it is recommended to use a classifier for statistical method. The algorithm will classify the data. Using this step, we can create a model that suits on the training the data. In the context of classification estimates the possibility of given point data falling into a certain object. This classifier method finds way to predict the finding parameters to feature individual probability. In this section, the data will be extracted to get the summary or the description. To identify the fields involved and the levels of data sets.

Figure 1. Exploring Variable

Figure 1 the researcher explores on some functions of R like the variable. This figure tells that we have 22 variables in the data file. The variable name converts names t lower case since R is predominantly done in lower case.

Figure 2. Size and Structure of data file.

Figure 2 shows the data file containing of 779919 observations and 22 variables. The figure has 9 variable integers such as the year, the beat, the district, the ward, the community, the Area, the x-coordinate, y-coordinate and ID. The case number, date, block and the others are character.

Pattern Identification, see Figure 3, is the trend identification section of the study. It identifies the trends and patterns of crime. The Apriori algorithm is used in this section to determine the association rules that will highlight the ruling in trends data file, it corresponds with the location that represents every attribute, such as the x coordinates, y coordinates, ward, district.

Figure 3. Local data in frame.
After determining the pattern, a new set of rules will arrive, however, if it is the same, we can predict that there’s another crime occur.

In this section the law enforcer will take the opportunity to prevent it by providing the necessary safety measures. In this case the police or any analyst take this as the detection for crime. The mounting information available to process by the technologies has enabled the law enforcement agencies to aggregate data into various crimes. The investigation of law enforcement agencies used the classifications techniques to apply on these data and to form decision-aid tools to facilitate investigations.

The implementation and analysis method of prediction can be done using the decision tree. Decision is parallel to a graph which had nodes that represent every attribute. The researcher proposes a decision tree-based classification model to predict the possible crime. The law enforcement agencies used this model to determine the crime patterns and predicting the future trends [Nasridinov, Park 2013].

Visualization section in Figure 4 shows graphical representation trends of crime. It is an art to transform the data into useful information. It shows the occurrences of the dates when the crime happened. It makes easier for view to interpret the information. Below shows the graphical representation of Total Crime by the District.

Heat map displays the probable regions to prevent crimes by taking preventing mechanism x-axis plotted the district while the y-axis plotted the rate of the crime. The heat map shows the regions which has the high volume of crimes. The data is based from the historical. Deering District has the high level of crime rate.

V. CONCLUSION

The flexibility of R gives us the new direction to have an effective and accurate win predicting the crime pattern in a location. Different algorithm can be the based training set of the data file. The relatively insignificant performance of the certain attribute factor for the randomness of the crimes and the features of the algorithm used, make the branches if the decision tress to build more rigid and accurate results if the test set follows the model pattern. On the other hand, the linear regression algorithm could handle randomness of test cases to make samples to certain attributes and to avoid error incur that leads to error in prediction. The effectiveness and accuracy of this project can improve by using machine learning algorithms that can predict intense crimes.

The application of data mining gives the realm and the curve of learning trends in crimes. The utilization of this application can be long and tedious due to the large volumes of datasets, but this will help the law enforcer to get the accurate data needed. The precision could infer and create new knowledge on how to slow down or lessen the crime
rate as well as the safety and security of the community and its people.

Clustering technique is used to cluster the similar type of crimes together, based on the clusters’ result the burglary type of crime hotspot will be identified. This result will help to reduce the burglary type crime. In future all type of crimes’ hotspot will be identified, through this the crime activities will be reduced.

REFERENCES