A PREDICTIVE MODEL FOR MAPPING CRIME USING BIG DATA ANALYTICS

Saoumya[1], Anurag Singh Baghel[2]

[1]*Gautam Buddha University, Uttar Pradesh, India*
[2]*Gautam Buddha University, Uttar Pradesh, India*

Abstract

Crime reduction and prevention challenges in today's world are becoming increasingly complex and are in need of a new technique that can handle the vast amount of information that is being generated. Traditional police capabilities mostly fall short in depicting the original division of criminal activities, thus contribute less in the suitable allocation of police services. In this paper methods are described for crime event forecasting, using Hadoop, by studying the geographical areas which are at greater risk and outside the traditional policing limits. The developed method makes the use of a geographical crime mapping algorithm to identify areas that have relatively high cases of crime. The term used for such places is hot spots. The identified hotspot clusters give valuable data that can be used to train the artificial neural network which further can model the trends of crime. The artificial neural network specification and estimation approach is enhanced by processing capability of Hadoop platform.

Keywords— *Crime forecasting; Cluster analysis; artificial neural networks; Patrolling; Big data; Hadoop; Gamma test.*

----------***----------

1. INTRODUCTION

Police will greatly benefit by a software that will be able to intelligently analyse a constantly updating database of crime incidences and its description, providing accurate predictions of where the crime is most likely occur and at what time . This will help in optimum police resource allocation. In doing so, one of the drawback is the fact that crime occurrences are generally sparse with respect to the type of crime, time and space at which the incidence occurs, and the randomness subjected to it. Apart from this the ability to process unstructured data was limited till now but with the advent of big data we can explore a new approach for making predictions. A crime analysis tool must be able to identify crime patterns accurately and efficiently for future forecasting and accurate crime pattern. However, in the present scenario, there are few challenges as followed

1) Increasing size of the information that has to be stored and analysed.
2) Different techniques that can analyse with accuracy and efficiency of this increasing volume of data on crime.
3) Varied methods and infrastructure that are used for recording data on crime.
4) The available data are inconsistent and incomplete and are making the task increasingly difficult formal analysis
5) Due to complex nature, it takes more time

This paper delineate a forecasting framework on the Hadoop platform that will be able to predict the near likely crimes. This work vary from other previous studies which basically describes the hot-spot methods and their statistical significance. Researchers have mainly focused on mapping and analysing crime divisions but in this paper the identified hot-spots clusters are used as the foundation for predictive algorithm. Thus hotspot visualization aids in crime prediction. Depending on backdated events it is used to recognise the areas of high occurrences of crime incidences so that appropriate police resources can be deployed at the identified locations. The software also harness the capability of Hadoop to process large amount of data in half the time as compared to other systems that have been studied till now.

The approach presented in this paper have three key stages as shown in Fig 1. The first is the geographic distribution of crime data analysis which identifies spatial clusters having greater risk of crime. In the second step a clustering algorithm is used to determine the quality of each identified cluster.

Fig. 1. Predictive Process Model

The final step involves the prediction which deploys an artificial neural network (ANN) model which is based on classification and regression tree predictive specification.

The paper basically shows how geographical clusters of crime data can train artificial neural networks to facilitate predictive modelling and how the same can be done on the

Hadoop platform, which renders the capability to process more data at a fast rate as well as using varied sources of data. So firstly the data is collected from the past records and it is mapped. Kernel density estimation is used to make the clusters from the mapped data. The Gamma Test (GT) is used estimate how much potential each cluster has to facilitate prediction. Then the identified clusters are fed into the artificial neural network which makes prediction about the future crime. The paper is concluded with a discussion of the results obtained and scope for future research.

2. CRIME PREDICTION THEORY

Many researchers have tried explaining why crimes occur in certain areas or is there any pattern that can be concluded from the past events. One such theory that answers these questions is the crime prevention theory [1]. According to it crime does not happen in random fashion, it is either opportunistic or planned. It states that any criminal activity occurs when there is intersection of work space of a target and the offender. The people's work space is comprised of places he/she visits in day to day routine, like workplace, educational institutes, shopping malls, recreational areas etc. All these specific locations of the offender or victim are also called nodes. Personal paths, the routes people takes every day connects with various nodes creating a circumference of personal space. This personal area is also person's awareness space. Thus Crime Pattern Theory states that a crime involving two people can only occur when the personal spaces of both intersects at one point or another. Thus we can say that crimes are not completely random, they can be studied and analysed to provide likable predictions. It may not be as accurate as the ones shown in the movie minority repot though but to some extent predictions can be made. Fig 2 depicts this phenomenon.

Simply put if an area provides the opportunity to the offender for crime and it is within the personal awareness space of the victim then crime will happen. Thus areas that are secluded and does not have any proper patrolling provides greater opportunity for crime.

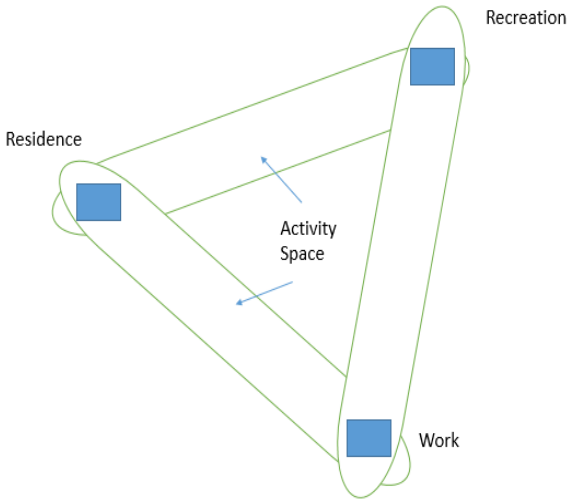

Fig. 2. Activity and Awareness Space of Criminals

Although places like shopping and recreational areas are the places where the offender and the victim are likely to meet. The reason being that a large amount of people visit there places and the offenders can easily mark their potential victims. The study of human behaviour is outside the scope of our study as we are only interested in finding a pattern that can prevent further crimes. Like one of the example being identification of places where many people fall victim to chain snatching or pick pocketing. This is mainly concentrated in certain areas only. Thus crime pattern theory provides an organized way to proceed in the direction of prediction exploring the patterns of crime and analysing them.

3. ARTIFICIAL NEURAL NETWORK FOR CRIME PREDICTION

Designing prediction models with artificial neural networks is a well-studied area. Many researchers have contributed in this field. In one study it is concluded that in the crime level forecasting methods, the models possess the characteristic of being autoregressive with input and output are generally Counts of crime: multiple inputs and a single output [2]. These types of models are used in this paper.

In order to test the artificial neural network it is subjected to a series of input vectors whose result is known to the tester but not to the corresponding network. So to determine the robustness of the training process the tester uses the answers given by the network, describing the determined level of crime, provided the input. If the tester feels that the robustness of the network is sufficient, then the network is said to have true predictive capabilities and is fit to use. Thus for the prediction we feed those input to the network for which we do not know the answers and assume that the provide output by the network is reliable.

4. HADOOP PLATFORM

The system is built on Apache Hadoop which is an open-source software framework used for storing and large-scale processing of data-sets on clusters of hardware. It is used as the basic platform so that large amount of data can be processed thus leading to more accurate predictions. The prediction algorithm is written in the mapper and reduce program which is processed on the multi cluster environment resulting in faster results [3].

5. THE CRIME OCCURRENCE DATA

The crime data taken into account in this paper are 100,000 criminal incidents spanning 5 year in area which measures roughly 457, 23400,040 m2. This research uses crime data that has happened before a particular date, to create hotspot cluster maps, and to test its robustness for forecasting when and where the crime are most likely to happen next. Four crime types are taken into consideration namely-burglary, street crime, theft from vehicles and theft of vehicles. The data sets used in the database of crime incidents are variables like time, day, month, weather, and location which are mapped as geographical coordinates [4]. In a

comparative study undertaken, KDE was the technique that consistently outshone the other techniques, amongst the crime types, street crime hotspot maps are generally found more accurate at predicting where near likely street crime would occur as compared to others. Hence KDE is used for the hotspot mapping.

5.1 Training Data from Crime Clusters

For the analysis of crime clusters, Kernel density estimation (KDE) was used which is stated as the most appropriate spatial analysis technique for mapping crime data. KDE is widely approved method that can be attributed to the fact that it is growing very rapidly and its availability is unquestionable (one of the example being MapInfo add-on Hotspot Detective), others factors include the deduced efficiency of hotspot mapping and the user friendly layout of the resulting map in comparison to other techniques. Point data (offences) are aggregated within a user defined search radius and a continuous surface that represents the volume or density of crime events across the needed area is calculated. A map is a smooth surface, which shows the variation of the density of crime / point across the study area, without adhering to geometric shapes like circle or ellipse. It also provides flexibility in configuring various parameters such as grid size and radius search, however, despite many useful recommendations, there is no universal doctrine on how to use these and in what circumstances. The following steps are followed.

1) Raw data analysis,
2) cluster analysis and geographic representation,
3) allocation of centroids to clusters,
4) allocation of crime incidents to cluster boundaries

5.2 Raw Data Analysis

A simple search algorithm is used which identifies the small areas that have higher than average incidence crime [5]. The algorithm uses a counting function that iterates through crime data incrementing the counter whenever any incidence is within the scan range. Clusters of the crime incidences are shown in Fig 3.

Two coordinates C1 and C2 are used which projects the coordinates of the crime incidence.

$$\text{If scan radius} < \sqrt{(C1-X)^2+(C2-Y)^2}$$

Where X and Y are coordinates of centroid.

5.3 Cluster Analysis and Geographic Representation

A heuristic approach is used in this step to identify the count of crime occurrences required for a cluster to be considered as salient. The heuristic rules makes use of the fact that crime incidents are generally clustered within small geographical areas known as hotspots. A graphical representation of dispersion of heuristically generated data is used to increase the radius of the zone associated with that centroid [6]. As the density of the centroids sample increases, so does the radius of influence associated with that centroid. User interaction and experimentation resulted in the radius of influence, which are set to count * 45, where count is the number of crimes related to the centre of gravity during the first stage of analysis. This leads to the identification of stakeholders.

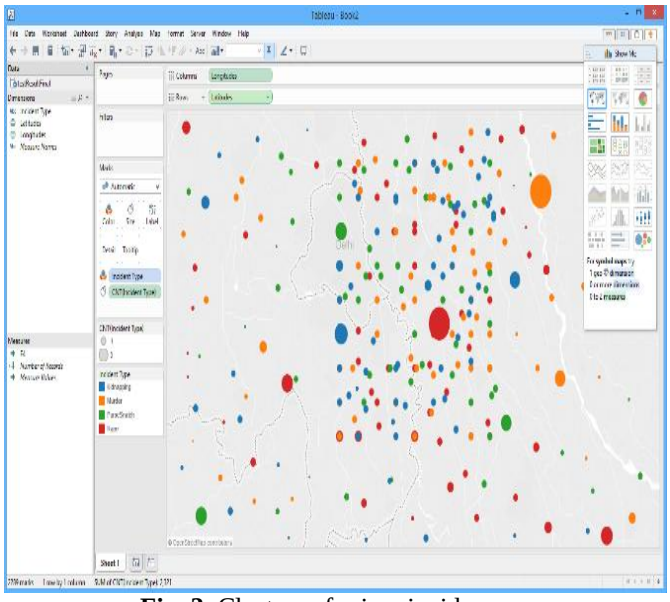

Fig. 3. Clusters of crime incidences

5.4 Allocation of Centroids to Clusters

Next, the centroids that are to be grouped together to form clusters are identified. The gravity and density parameters, along with a centroid list generated in previous stage, forms the basis for this iterative procedure.

Cluster	Centroid	Crime count
1	3,6,7	3778
2	2,5	3232
3	17,34	323
4	56,67,89	1782
5	119,134	234
6	159	897

Fig. 4. Crime incidence by cluster

5.5 Allocating Incidents to Cluster Boundaries

Finally, each group is filled with data ready for a series of training neural networks, one for each group. Each record contains a unique identifier crime, the cluster to which it belongs, the population's economic situation and the day of the week during which the crime was committed. Fig 4 shows the cluster and its corresponding crime count. In addition, each group record has a unique identifier, a list of its member centroids and a total count of crime.

6. GAMMA TEST FOR CLUSTER ANALYSIS

Considering the results of cluster analysis, autoregressive techniques were used to model the data grouped. The autoregressive model was selected as the preferred method for the problem of short-term prediction [7]. Below associated methodology is discussed.

7. FORECASTING USING ARTIFICIAL NEURAL NETWORK

The implementation of an ANN model requires consideration of accurate model parameters that affect the efficiency and stability of the model. These include decisions concerning the number of input / output nodes and hidden layers, training algorithm selection, and volume of data to be used for training and testing.

7.1 The Neural Network Architecture

Tree-structured prediction is used which is becoming increasingly popular in a vast domain of applications. A tree is a graph prediction which is associated with the following statistical model [8].

A characteristic function of the set is used. The ANN has two inner layers, with the specified activation functions. The number of neurons in the first layer is same as the number of nodes in the tree, it makes functions at the nodes such that each neuron has output 1 or 0 depending on whether or not the function defining the branching has yes or no answer. The second layer contains many neurons as there are leaves in the tree and the output of each neuron is 1 or 0 depending on whether the subject with predictor x is assigned or not to the corresponding leaf. The output layer simply realizes equation. The best topology of nodes in the hidden layer is also empirically determined. Previous research indicated that using a single hidden layer is sufficient to learn any complex nonlinear function suggest that two hidden layers can produce more efficient architectures.

Initial large number of nodes in the hidden layer was incrementally reduced to a minimum while maintaining an acceptable prediction capabilities. The shallow gradient (shown for the town centre cluster) suggested that a relatively few number of hidden nodes compared to 2 N 11 rule would be sufficient to model the underlying function, and it turned out to be the case. The standard gradient descent method for adjusting weights are replaced with conjugate gradient descent using earlier gradient measures to improve the error minimization process .

7.2 Terminating the Training Procedure

In the field of artificial neural network overfitting is widely accepted problem but gamma test can suitably measure and remove the noise from the data thus determining where exactly the training should be stopped, which is very useful for the researchers [9]. Overfitting mainly occurs because the ANN attempts to fit all data fed into the network, including the noise present. If we know the measure of noise which is included in the data sets it will be a lot easier to determine the point at which the training will be stopped, because the network will try to fit the useful data before the noise. Thus, the GT statistic G gives the MSE value at which training needs to be stopped [10].

7.3 Partitioning the Inputs into Test and Training Sets

First we need to determine the number of input required to model the output. Once it is calculated, the data needs to fit into the designed architecture. This architecture is used and M-test is performed to determine whether the number of inputs are sufficient to model the discussed algorithm. An asymptotic level for the Gamma statistic (which approximates to the inherent noise of the output) points out that the data is sufficient and establish a point at which we can segregate the test data and the training data. It is a very useful technique because it allows the data to be divided into two sets rather than three. Consequently, Paris, Ware Wilson, and Jenkins (2002) has demonstrated that there is no need of the validation data, whose main use is to determine at what point the training data will result in overfitting thus allowing the maximum possible use of data efficiently. Therefore the choice of appropriate quantity of data required for modelling is given by the M-Test.

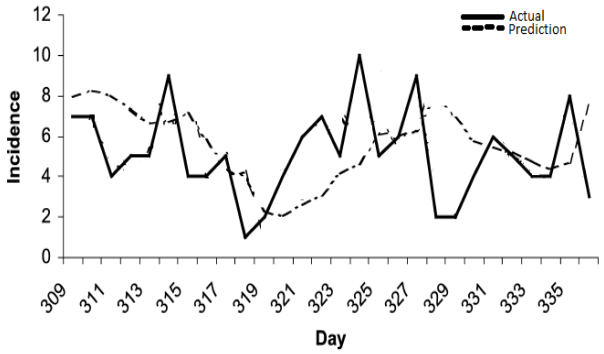

Fig. 5. Forecast and Incidence of crime

8. DISCUSSION OF RESULT

The results commensurate the expectations seeing the gamma test's output. Importantly, the noise which represents the exceptional incidence levels was evidently excluded giving reliable results. The use of the gamma test was represented by a pre model evaluation technique. The city centre cluster offered the best predictive model using the artificial neural network, cluster seven generated poor models. The incorporation of the statistical techniques improved the results. Using the Hadoop multi cluster platform reduced the time of processing. Now further experiments can be done which can cover a longer period of time duration as well as incorporate the diverse nature of data. Fig 5 represents the accuracy of prediction.

9. CONCLUSION AND FUTURE WORK

The paper describes a forecasting framework keeping in mind the geographical areas of concern that may transcend traditional policing limits. This paper sheds light upon the method of developing a practical system which can be used for an effective operational policing environment which in turn can decrease the drawbacks of inefficient techniques and head towards a more dynamic methodology. The computer created Hadoop procedure will utilise a

geographically mapped crime occurrences-scanning algorithm to map the clusters with relatively high levels of crime which are known as hot spots. The mapped clusters give sufficient data which is analysed by the gamma procedure to test the fitness of the data required for predictive modelling. By using the results obtained from the gamma test, the Hadoop model based on artificial neural network is implemented. As the previous study states, the artificial neural network generally displays a superior ability to model the trends within each cluster.

Further development will extend to the modelling of more detailed scenarios to facilitate prediction based on detailed input crime. Thus, the impact on the area for an upcoming holiday, where the weather is expected to be warm, could be evaluated. The aim here was to extract an underlying generalized model of crime incidents. However, specific locations best modelled independently of the other data on specific times of the year. Although there considerations can be modelled separately if sufficient amount of high quality data is backing it. As well as many statistical tools can study the exceptional events that will provide greater insight to how these events change the levels of crime and ultimately defining the rules that will modify incidence count significantly.

REFERENCES

[1] Balkin, S. D., & Ord, J. K. (2000). Automatic neural network modelling for univariate time series. International Journal of Forecasting, 16, 509–515.

[2] Woodworth JT, Mohler GO, Bertozzi AL, Brantingham, "Non-local crime density estimation incorporating housing information",.Phil. Trans. R. Soc. A 372: 20130403, September 2014.

[3] Han hu, Yonggang wen, Tat-seng chua, and Xuelong li," Toward Scalable Systems for Big Data Analytics: A Technology Tutorial", 1-School of Computing, National University of Singapore, Singapore 117417, 2014.

[4] Big Data Startup," The Los Angeles Police Department Is Predicting and Fighting Crime with Big Data," http://www.bigdata-startups.com/BigData-startup/los-angeles police-department-predicts-fights-crime-big-data/, May 2013.

[5] Chainey, S., & Reid, S. (2002). When is a hotspot a hotspot? A procedure for creating statistically robust hotspot maps of crime. In Kidner, D. B. et al. (Ed.), *Socio-economic applications in geographical information science*. London: Taylor and Francis, pp. 21–36.

[6] D.Usha, Dr.K.Rameshkumar," A Complete Survey on application of Frequent Pattern Mining and Association Rule Mining on Crime Pattern Mining", International Journal of Advances in Computer Science and Technology ISSN 2320 – 2602 Volume 3, No.4, April 2014.

[7] Hirschfield, A., 2001. Decision support in crime prevention: data analysis, policy evaluation and GIS. In: Mapping and Analysing Crime Data - Lessons from Research and Practice, A., Hirschfield & K, Bowers (Eds.). Taylor and Francis, 2001, pp. 237-269.

[8] Antonio Ciampi and Yves Lechevallier, Statistical Models and Artificial Neural Networks: Supervised Classification and Prediction Via Soft Trees, McGill University, Montreal, QC, Canada INRIA-Rocquencourt, Le Chesnay, France

[9] W. Chang et al., "An International Perspective on Fighting Cybercrime," Proc. 1st NSF/NIJ Symp. Intelligence and Security Informatics, LNCS 2665, Springer-Verlag, 2003, pp. 379-384.

[10] Jonathan J. Corcoran, Ian D. Wilson, J. Andrew Ware, P redicting the geo-temporal variations of crime and disorder, International Journal of Forecasting 19 (2003) 623–634.