A Survey on Data Mining Techniques for Crime Hotspots Prediction

Neha Patel[1] Prof. Shivani V. Vora[2]
[1]P.G. Student [2]Assistant Professor
[1,2]Department of Computer Engineering
[1,2]CGPIT, UkaTarsadiya University, Mahuva, Surat, Gujarat, India.

Abstract— A crime is an act which is against the laws of a country or region. The technique which is used to find areas on a map which have high crime intensity is known as crime hotspot prediction. The technique uses the crime data which includes the area with crime rate and predict the future location with high crime intensity. The motivation of crime hotspot prediction is to raise people's awareness regarding the dangerous location in certain time period. It can help for police resource allocation for creating a safe environment. The paper presents survey of different types of data mining techniques for crime hotspots prediction.

Key words: Data mining; crime hotspot; prediction; accuracy

I. INTRODUCTION

Crime is a social problem that affecting the people's life and economic development of a society. The prediction of crime is difficult [1]. The occurrence of crime is related to a variety of socio-economic and crime opportunity factors, like population, economic investment and arrest rate.

Crime usually has spatial and temporal characteristics related to the population, environment, economic factors, politics, and social events. The victims of crime may not be predicted but the place that has probability of an occurrence of crime it may be predicted. Analysis of large amount of crime data is a difficult task. Data mining is a useful process to handle amount of data and to reduce the manpower. The most effective spatial-temporal analysis for the understanding of the contained relationships among crime events is the analysis of crime hot spots. The analysis of crime hot spots contributes to the conflict and prevention of crimes by allowing the planning of strategies that optimize the distribution of police resources. As police resources are limited in some areas, the planning of such allocation becomes an important task. Two type of crime hotspots are there: crime general and crime specific. Wide range of crime types are occur in a particular area then is known as crime general hotspot. One or several types of crimes are occur in a particular area then it is called as a crime specific hotspot. The crime hotspots prediction is done using crime location, crime time and which type of crime is done. We can also find the area for specific crime. The main motivation of crime hotspots prediction is to create a safe environment and for police resource allocation.

In the next section-II the general steps for crime hotspots prediction and classification techniques are introduced. Where five classification techniques are described. Section-III represents the comparative analysis of three classification techniques which are support vector machine, decision tree and naïve bays.

II. LITERATURE SURVEY

The general steps for crime hotspot prediction are:

1)Data collection
2)Preprocessing
3)Feature selection
4)Classification
5)Prediction
6)Visualization

A. Data Collection

In data collection step crime data is collected from different sources like news sites, blogs, RSS feed and mainly from police records. For unstructured data Mongo database is used [2].

B. Preprocessing

There are few techniques for data preprocessing. This techniques are data cleaning, reduction, integration, discretization, transformation and feature selection. It intends to reduce some noises, incomplete and in consist data.

The data cleaning is used to decrease noise and handle missing values. There are a number of methods for handling records that contain missing values such as omitting the incorrect fields(s) or entire record that contains the incorrect field(s), automatically entering or correcting the data with default values, deriving a model to enter or correct the data, replacing all values with a global constant and using the imputation method to predict missing values.

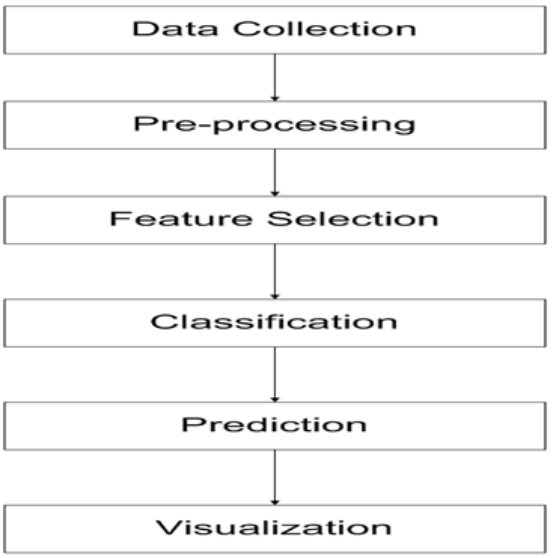

Fig. 1: steps for crime hotspot prediction [2][3].

Data reduction is necessary to remove irrelevant attributes from dataset. For example according to Almanie, Mirza and Lor [4], the authors performed data reduction in terms of number of instances. They observed Denver crimes dataset contained a set of traffic accident instances. The attribute "Is_Crime" suggests whether the instance belongs to a crime or accident. The author used the attribute

"Is_Crime" to filter the instances and remove all the irrelevant ones [4].

Data integration steps is used for different purposes. In Almanie, Mirza and Lor [4] the authors used to avoid different attribute naming, they 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.

The data transformation is used to reduce the diversity of attribute values by mapping their values to fall within smaller group. For example burglary and robbery crimes are included in theft crime type [4].

C. Feature Selection

Feature selection is a part of data preprocessing. Feature selection is used to remove the irrelevant or redundant attributes. Feature selection has several objectives such as enhancing model performance by avoiding over fitting in the case of supervised classification [1]. The main attributes like crime type, location, crime time in feature selection process.

D. Classification

After preprocessing and feature selection phases, the numbers of attribute was meaningfully extract and are now more precise for building the data mining models. Classification as a famous data mining supervised learning techniques are used to extract meaningful information from large datasets and can be adequately used to predict unknown classes. There are various classification algorithms, such as Support Vector Machines (SVM), k-Nearest Neighbor (k-NN), Decision Tree, Weighted Voting and Artificial Neural Networks. All these techniques can be applied to a dataset for discovering sets of models to forecast unknown class labels [1][3].

E. Prediction

In order to quantitatively predict the crime status, many data mining methods can be used. In this study, a classification task is applied for prediction.

F. Visualization

The crime prone areas can be graphically represented using a heat map which indicates level of activity, generally darker colors to represents low activity and brighter colors to represents high activity [1][2][4].

The steps for crime hotspots prediction are introduced. Now the survey of different data mining techniques for prediction are presented. The objective of prediction is to forecast the value of an attribute based on value of other attributes. In the prediction techniques first a model is created based on data distribution and then that model is used to predict future on unknown value. The basic data mining techniques are introduced which are used to predict crime hotspots.

1) Support Vector Machine

The support vector machines are supervised learning models with associated learning algorithms that analyze the data and recognize the patterns that is used for classification and regression analysis. A support vector machine construct a hyperplane or set of hyperplanes in a high or infinite-dimensional space, which can be used for classification, regression and other tasks.

The support vector machine technique is used for support vector machine [3]. The following approach is: The data used for this research was taken from a variety of city agencies. Each data contains the type of event, the location with longitude and latitude, and the time and date of the incident. This data was classified from different classification techniques. The area which have the crime rate above the predefined rate are positive or members of hotspot class and area with crime rate below the predefined rate are negative or non-members of hotspot class. This labelled data set used as the training set in SVM classification. The technique gives good results in most cases but it is computationally expensive so it runs slow.

2) Naïve Bayesian

Naïve Bayes classifier is a supervised learning algorithm. It is effective and widely used. It is a statistical model that predicts class membership probabilities based on Bayes' theorem [5]. The naïve bayes classifier model is fast to build. It can be modified with new training data without having rebuild the model. This classifier is very simple to construct and it may be easily apply to huge data sets [1]. Predictive accuracy is generally of this classifier in most cases. It consider each attributes separately when classify new instance. It based on the Bayes rule of conditional probability [6][5].

$$P(H|X) = P(X|H)\, P(H)/ P(X)$$

Where,

1) P (A) is the prior probability of A. It is "prior" in the sense that it does not take into account any information about B.
2) P (A|B) is the conditional probability of A, given B. It is also called the posterior probability because it is derived from or depends on the specified value of B.
3) P (B|A) is the conditional probability of B given A. It is also called the likelihood.

3) Decision Tree

Decision tree is a flow chart like structure in which each internal node represents a 'test' on the attribute, each branch represent outcome of the test and each leaf node represents a class label. It is simple to understand and to interpret. It is able to handle both numerical and categorical data. It is also able to handle multi-output problems. It performs well even if its assumptions are somewhat violated by the true model from which the data was generated. For crime hotspot prediction generally J48 algorithm is used [2][3][4].

The decision tree can be unstable because small variations in data might results in completely different tree being generated. Another disadvantage is its complexity [6].

4) Artificial Neural Network

In 1943, McCulloch and Pitts, gives the first model of artificial neuron. According to Nigrin, A neural network is a circuit composed of a very large numbers of processing elements that know as Neuron. Each element works only on local information. Furthermore each element operates non-parellerliy, thus there is no system clock [7]. The neural network have high strength when modeling a complex system. It gives higher accuracy when increasing the data but sudden changes in new data might give low results [4]. It takes long running time [6].

5) K- Nearest Neighbor

The k-Nearest Neighbors algorithm (or k-NN for short) is a non-parametric method used for classification and regression [8]. In this technique the classification is done by comparing feature vectors of the different points. Nearest neighbor classifier makes their predictions based on local information, whereas decision tree and rule based classifiers attempts to find a global model that fits the entire input space. Because the classification decision are made locally. This classifier can produce wrong predictions unless the approximate proximity measure and data preprocessing steps are taken [6].

In the next section, observation of survey are presented.

III. COMPARATIVE ANALYSIS

Parameters	Support vector machine	Decision tree	Naïve Bayes
Summary	Supervised learning model. Analyze data recognize patterns [9].	Flowchart like structure in which each internal node represents a "test" on attribute, branch as an outcome of taste and leaf as a class label [1].	Supervised learning algorithm. It is a statistical model that predicts class membership probabilities based on Bayes' theorem [4].
Accuracy			
Crime dataset US(Northeast)[4]	70% to 80%[4]	60% to 70%[4]	70% to 80%[4]
Crime data set of Denver[4]	–	42%[4]	51%[4]
Crime data set of Los Angeles[4]		43%[4]	54%[4]
F1-measure			
Crime dataset US(Northeast)[4]	70% to 80%[4]	70% to 80%[4]	Above 80%[4]

Table 1. Comparative analysis for data mining techniques.

Table 1 shows the comparative analysis of different data mining techniques. All techniques are compared with its prediction accuracy and f1-measure. The support vector machine provide good accuracy but it is computationally expensive thus runs slow. The complexity of decision tree is high. Small variations in data might results in completely different tree being generated. The Naïve bays is highly scalable and easy to implement. Good results obtained in most cases. The results are also depends on which type of data is given. Here the naïve bays gives high accuracy and f1-score for different crime database compared to support vector machine and decision tree.

IV. CONCLUSION

The data mining techniques for predicting crime hotspots are discuss in this paper. This techniques are capable to enhance the prediction accuracy, performance and speed. After analyzing different results from various paper we can conclude that Naïve Bayes gives efficient results for crime hotspot prediction.

V. AKNOWLWDGEMENT

I take the opportunity to express my gratitude and regards to my guide Prof.Shivani Vora, CE & IT Dept., CGPIT, Bardoli for her suggestions and encouragements.

REFERENCES

[1] Somayeh Shojaee,Aida Mustapha, Fatimah Sidi, Marzanah A.Jabar, "A study on Classification Learning algorithms to predict crime status", International Journal of Digital Content Technology and its Applications, vol 7, 2013
[2] Shiju Sathyadevan, Devan M.S and Surya Gangadharan. S," Crime Analysis and Prediction Using Data Mining", First International Conference on Networks & Soft Computing,IEEE
[3] Chung-Hsien Yu, Max W. Ward, Melissa Morabito and Wei Ding(2011),"Crime Forecasting Using Data Mining Techniques", Department of Computer Science, 2Department of Sociology, University of Massachusetts Boston, 100 Morrissey Blvd., Boston, MA 02125
[4] Tahani Almanie, Rsha Mirza and Elizabeth Lor," Crime Prediction based on Crime Types and using Spitial and Temporal Criminal Hotspots", International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.5, No.4, July 2015
[5] https://en.wikipedia.org/wiki/Bayes%27_theorem
[6] Michael Steinbach, Vipin Kumar, Pang-Ning Tan, "Classification: Alternative Techniques" in Introduction to Data Mining 3rd edition, 2006
[7] Nikhil Dubey, Setu Kumar Chaturvedi,ph.d, "A Survey paper on Crime Prediction Technique using Data Mining", International journal of Engineering research and Applications, vol 1, 2014
[8] https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm
[9] Keivan Kianmehr and Reda Alhajj, "Crime Hot-Spots Prediction Using Support Vector Machine" pp. 952-960 IEEE 2006.
[10] https://en.wikipedia.org/wiki/Support_vector_machine.

