Survey Paper on Crime Prediction using Ensemble Approach

Ayisheshim Almaw1
Department of CS/IT, Symbiosis International University, Pune, India
Ayisheshim.bogale@sitpune.edu.in

Kalyani Kadam2
Assistant professor, Department of CS/IT, Symbiosis International University, Pune, India
kalyanik@sitpune.edu.in

Abstract—Crime is a foremost problem where the top priority has been concerned by individual, the community and government. This paper investigates a number of data mining algorithms and ensemble learning which are applied on crime data mining. This survey paper describes a summary of the methods and techniques which are implemented in crime data analysis and prediction. Crime forecasting is a way of trying to mining out and decreasing the upcoming crimes by forecasting the future crime that will occur. Crime prediction practices historical data and after examining data, predict the upcoming crime with respect to location, time, day, season and year. In present crime cases rapidly increases so it is an inspiring task to foresee upcoming crimes closely with better accuracy. Data mining methods are too important to resolving crime problem with investigating hidden crime patterns. The objective of this study could be analyzing and discussing various methods which are applied on crime prediction and analysis. This paper delivers reasonable investigation of Data mining Techniques and ensemble classification techniques for discovery and prediction of upcoming crime.

Keywords: crime prediction, Naïve Bayes, J48, Artificial network.

I. INTRODUCTION

Crimes are a community bother and charges the community tremendously in different means. Crime data mining has the capability of mining or extracting relevant information and unknown crime patterns from the huge crime database. It becomes obligatory to develop an organized review focuses on crime data mining methods and technologies used in previous studies.

By studying the crime models forecasting can be done and the effect of the crime in different locations can be stopped. There are different data development and procurement tools available for this purpose [2-7]. In this paper all types of crime data mining techniques are studied and comprehensive evaluation is presented.

Different classification algorithms in problem resolving are nominated based upon the suitable requirements in crime data prediction. Each technique provides a diverse accuracy and estimated outcome. One technique may provide better accuracy values than different techniques which are nominated for solving the particular problem. Some papers introduced combined different models to achieve better performance which overcomes the individual models called ensemble learning. Ensemble learning is a way of combining different diversified data mining algorithms to achieve better accuracy to solve a particular problem [12-16] [30-32].

II. LITERATURE SURVEY

Researchers have proposed a variety of data mining techniques to provide crime data analysis, crime prediction, criminal identification and crime hotspot area identification. Some of the papers are discussed here.

Mehmet Sait, and Mustafa Gök presented [1] the criminal prediction for finding the most probable criminal of a particular offense incident when the suspected list of offenders are provided with the criminal data which is generated synthetically using Gaussian Mixture Model. The authors used Naïve Bayes Classifier and Decision tree for offender prediction method and compared its performance. As
a result of the comparison the authors achieved that the Naïve Bayes Classifier consumed less execution time and performs better with 78.05% accuracy. Agarwal A., analyzed [3] various offenses done by offenders and predict the chance of each offense that can again be performed by that offenders. The authors used Apriori practice for frequent item set generation that can be done by the offenders.

Ahishakiye, E., Anisha, and C.Dhanashree applied [2] J48 base model to predict crime category or level in certain location that will occur in the future. Sivaranjani, S., S. Sivakumari, and M. AashaPresented [4] crime analysis for six cities of Tamilnadu, India by using clustering practices k-means, DBSCAN and Agglomerative clustering for grouping the similar patterns to recognize offenses and the authors conclude that DBSCAN clustering performs better with precision 0.95, recall 0.91 and F measure 0.93 for grouping the similar patterns to identify crimes in for six cities of Tamilnadu, India. The authors used KNN practice to extract and predicts future offenses that will occur in the future in six cities of Tamilnadu, India which have possibility of low, medium and high offense occurrence by visualizing on google map. Emmanuel A., et al. [4] Analyzed crime data by using support vector machine, naïve bayes, neural network and J48 and contrasts the techniques by using accuracy and execution time for predicting offense level as ‘Low’, ‘Medium’, and ‘High’. As a result of the contrast the authors conclude that the decision tree (J48) consumed less execution time with 0.06 seconds and performs better with 100% accuracy for crime forecasting. Yerpude P., et al. applied [7] data mining practices from crime data for foreseeing features that affect the high or low crime rate in certain region. The authors used Random Forest, Naïve Bayes and Linear Regression for recognizing factors that affect the high crime rate and compared its performance. As a result of the comparison the authors conclude that the Random Forest performs better with 81.35% accuracy.

Nafiz M., et al. introduced [6] CRIMECAST, a mathematical simulation tool that analyzes past crime trends, patterns, features affect crime, Crime occurrence frequency, crime taken place, crime happened time, type of crime and victims from past crime data up to 30 years to forecast future crime. Tahani A., et al. [8] analyzed two different crime data using Decision Tree and Naïve Bayesian classifier to locate the most probable crime locations and their frequent occurrence time using Apriori Algorithm. The authors introduced what kind of offense might happen next in a specific place within a certain time and combining crimes’ dataset with its demographics information to capture the issues that might disturb the safety of neighborhoods. As a result of the comparison the authors conclude that the Naïve Bayes performs better with 51% accuracy for Denver and 54% for Los Angeles for crime prediction.

Thongsataponiwatana U., studied 0, [22] different researches which are used data mining techniques for crime data analysis and foreseeing. The author identifies the research gap and challenges from different studies and recommends different data mining techniques for finding the patterns and trends in crime data to help the researcher on crime data.

Rasoul K et al., [10] analyzed the crime data using the data mining techniques such as K-means Algorithm for grouping the similar crime patterns for identifying crime in different years based on amount of crime occurrence during different years and recognizing the crime patterns and trends to suggest this way can be used to decrease and avoid crime for the coming future years. The author presented the effect of parameters such as effect of outlier in data preprocessing and introduced GA for outlier detection in data preprocessing stage. The following table describes the remaining various papers which are done on crime data mining and ensemble learning.

<table>
<thead>
<tr>
<th>No</th>
<th>Authors</th>
<th>Journal and publication year</th>
<th>Methodology &amp; tools</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Sathyadevan, Shiju, and Surya Gangadharan [17]</td>
<td>IEEE 2014</td>
<td>Naïve Bayes Decision Tree, Apriori Algorithm, Mongo DB, GraphDB</td>
<td>It forecasts districts which have high likelihood of crime happening in districts in particular day and Naïve Bayes predict with 90% accuracy for crime data analysis and foreseeing. Apriori Algorithm is used to find crime patterns that occur repeatedly in a particular districts. The paper visualized high crime prone districts by GraphDB &amp; data stored in mongoDB.</td>
</tr>
<tr>
<td>2</td>
<td>Andrey B et al. [18]</td>
<td>ACM 2014</td>
<td>Random Forest, support vector machines, neural networks, decision trees</td>
<td>They used smartstep datasets a data derived from digital telecommunication of individual mobile system activity and demographic data or co-localised Open Data to predict the crime category which will occur in upcoming months in London metropolitan area by dividing into different cells. The authors used Random Forest, support vector machines, neural networks and decision trees for predicting crime category by integrating to different data sets and random forest performs better</td>
</tr>
</tbody>
</table>
The authors used LDA for feature reduction and KNN for prediction of high crime intensity regions.

4. Tayebi MA, Ester M, Glässer U, Brantingham PL [20]. IEEE/ACM 2014 CrimeTracer model (Personalized Random Walk Model) -Mohammad A. et el. presented CrimeTracer (which consists of a criminal daily lives, road network and whole places where criminal carry out a crime), random walk approach for crime location forecasting for urban area. The CrimeTracer is based on the probabilistic framework to model the features of known individual criminal within urban areas they are aware with, known as activity spaces.

5. Cesario, Cesario E, Catlett C, Talia D [22]. IEEE 2014 Naïve Bayes classifier Neural network - They analyzed two various data sets for prediction of crime level category collected from different sources. The authors analyzed & compare Naïve Bayes & Neural network to predict crime level that will occur in particular location and Naïve Bayes classifier performs better with 90.2207% of accuracy.

6. Retnowardhani, Retnowardhani A, Triana YS [23]. IEEE 2016 Crime Prevention Decision Support System (CreP-DSS) php Astari R. et el. Developed CreP-DSS web-based decision support system in PHP that supports law enforcement agency to control and support decision-making tasks for crime prevention. It splits the interval range for predicted crime trend, based on a level of offense for better decision-making to prevent crime.

7. Yu CH, Ward MW, Morabito M, Ding W [24]. IEEE Chung 2011 Support Vector Machine, Neural Network, Naïve Bayes, 1-NN (One- Nearest Neighbor), J48 and Ensemble learning (Voting) Chung-Hsien Y. et el. Presented predictive technique for areas that have a high density of residential burglaries and the probability that will happen in future. Primarily, a proper technique of organizing data is considered to store temporal/spatial information as well as combined numbers of offenses and offense-related actions by the police officer. Secondly, an ensemble supervised data mining algorithm is applied to achieve residential burglary prediction with better accuracy.

<table>
<thead>
<tr>
<th>No</th>
<th>Authors</th>
<th>Journal &amp; publication year</th>
<th>Methodology and tools</th>
<th>Data source and Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>Iqbal R, Murad MA, Mustapha A, Panahy PH, Khanahmadiravi N [25].</td>
<td>Indian Journal of Science and Technology 2013</td>
<td>Naïve Bayes, Decision Tree</td>
<td>Dataset: crime and communities from the UCI machine learning repository with 128 attribute and 1994 observations. Result: Predicts crime category in different states of USA. Decision Tree &gt; Naïve Bayes with an accuracy of 83.95%.</td>
</tr>
</tbody>
</table>

11. Dubey N, Chaturvedi SK [28]. IERA 2014 SVM, BayesNet Result: presented theoretical compression techniques for future crime prediction based on the idea of each technique used, predictive accuracy and disadvantages of each techniques.

12. Gupta A, Mohammad A, Syed A, Halgamuge MN [29]. International Journal of Advanced Computer Science and Applications 2016 (Naive Bayes, OneR, Decision Table, J48, JRip, BayesNet Dataset: it contains Denver city, USA crime and accident data Result: to identify the best classification algorithm and compares using time taken to execute to predict the crime that will occur in certain season, JRip > all the other techniques with accuracy of 73.71%.

<table>
<thead>
<tr>
<th>No</th>
<th>Authors</th>
<th>Journal &amp; publication year</th>
<th>Methodology &amp; tools</th>
<th>Source of dataset and Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>14.</td>
<td>Rojarath, Arittayapron, WararatSongpan, and Chakrit Pongtwong[12].</td>
<td>IEEE 2016</td>
<td>3-ensemble models (Naive Bayes, Decision Tree, &amp; Multilayer perceptron 4-ensemble (Naive Bayes, Decision Tree, Multilayer perceptron and K-Nearest Neighbor)</td>
<td>Dataset: UCI Harberman, Urban, Chronic_kidney, Mammographic, Phoneme &amp; Pima. Result: Presented M-ensemble learning approach which is divided into two, 3-ensemble model (contains odd number of base models) and 4-ensemble model (number of base models) to enhance the performance of base model classifiers through majority voting. 3-ensemble models are better than 4-ensemble models to enhance the performance with an accuracy of 83.13% in order to forecasting unknown data.</td>
</tr>
<tr>
<td>15.</td>
<td>Hassan MP, Abdel-Qader I [16].</td>
<td>IEEE 2015</td>
<td>Computer simulation tool</td>
<td>Result: Developed a theoretical technique to study the performance of the majority voting integrator by considering classifiers’ outcomes are dependent. The author concludes the accuracy of ensemble model degrades when each classifiers output are correlated.</td>
</tr>
<tr>
<td>17.</td>
<td>Joshi, Nikita, and Shweta Srivastava [31].</td>
<td>IJCSMC 2014</td>
<td>BF: Tree, J48, Decision stump and CART Bagging</td>
<td>Dataset: UCI Anneal, Credit, Iris, Wine, Zoo, Vowel and Dermatology datasets Result: applied bagging ensemble learning approach to enhance the performance of individual decision tree base J48, Decision stump and CART classifiers. The author performs experiment by applying different data set and comparative analysis to know the effectiveness. As a result, bagging ensemble model with decision tree classifiers performs better.</td>
</tr>
<tr>
<td>18.</td>
<td>Wan S, Yang H [13].</td>
<td>IEEE 2013</td>
<td>bagging, boosting, stacking &amp; random forest</td>
<td>Dataset: UCI database 31 different data sets Result: The authors compared bootstrap bagging, boosting, random forest and stacked generalization ensemble techniques to attain better performance. As a result, the shows that random forest performs better.</td>
</tr>
</tbody>
</table>
Crime is a serious problem which should be undertaken and controlled by community as well as the whole world. Huge number of peoples, society regions and world is affected with crime. Crime prediction and finding relevant information from large amount of crime data is very important but challenging. If the advanced prediction about the problem can be made then crime may be stopped. If not stopped then can be reduced. Large amount of work is being done toward this area. But still some work can be done to improve the forecasting system. A survey is conducted so that Crime forecasting can be improved by the use of efficient data collection and data mining strategies.

Crime trends and patterns identification and Predict what kind of offenses might occur next in a particular district within a specific period of time and season. So, combined techniques are required to build a better crime prediction by integrating multiple models to solve simple problem for improving performance prediction of single classifier that helps to predict what kind of crime might occur next in a particular district within a certain period of time and identifies the Season and time factor at which crimes are occur more frequently happening of crime. Hence crime prediction helps people stay away from the districts at a certain time of the day, month and season along with saving living style. In addition, having this kind of knowledge would help people to improve their living and travelling place choices.

REFERENCES


