A Literature Survey on Criminal Identification, Crime Pattern Detection, and Prediction in India using Data Mining

Chithra Shaji Thomas
Assistant Professor
Department of Computer Science and Engineering
Mount Zion Institute of Science and Technology, Alappuzha, India

Abstract— Data mining is one of the most powerful ways of knowledge extraction, especially in case of large datasets. It is one of the best approaches to detect underlying relationships among data with the help of machine learning and artificial intelligence techniques. Crime Detection is one of the most important topic in data mining where different patterns of crime are identified. It includes variety of steps, starting from criminal identification, till detection of crime pattern, and prediction. For this purpose, various crime detection techniques are used and they have been analyzed in this review.

Keywords: Data Mining, Knowledge Extraction Crime Detection, Crime Pattern, Prediction

I. INTRODUCTION

Crimes in India are rising at an alarming rate because of the factors such as increase in poverty, migration, unemployment, frustration, illiteracy and corruption. Crime investigating agencies search the database of criminals manually or with some computer data analyst which is a tedious process and takes much more time. So to contribute toward combating crimes and to identify criminals, different technologies are proposed using data mining.

"Data mining refers to the extraction, discovery and analysis of meaningful patterns and rules from a very large amount of data". It is emerging as very useful tool for crime detection. Data mining is a very powerful tool to undermine the activities of criminals by analyzing the criminal's record and information and preventing the crimes in future. Data mining for crime detection is considered one of the most important research area despite data mining being a new and evolving field itself. Data mining is very helpful and accurate in understanding the crime trends as compared to Humans. Data mining techniques offers some predictive models that manipulate the hidden information and can predict the trends.

Hosseinkhani et al.,[13] has suggested some data mining techniques that can be used for crime detection. These techniques are clustering, association rule mining, deviation detection, classification and string comparator. The crime detection data mining techniques as presented by Hsinchun et al., are entity extraction, clustering, association rule mining, sequential pattern mining, deviation detection, classification, string comparator and social network analysis. Hossein Hassani et al., have later presented a review of some existing crime data mining techniques which included; entity extraction, cluster analysis, association rule, classification and social network analysis [14]. The aim of this study is to get an insight of the techniques that are followed for crime data mining.

In this survey, the existing methods and techniques of data mining for crime detection and investigation are thoroughly observed. Different data mining techniques are discussed and analyzed for their usage. They are then compared for their strengths and weaknesses.

The various crime data mining methods [13], [14] include:

1) Clustering techniques
2) Association rule mining
3) Classification
4) Sequential pattern mining
5) Deviation detection
6) Social network analysis
7) Entity extraction
8) String comparator

II. EXISTING CRIME DATA MINING TECHNIQUES

A. *Predict a Future Crime using Data Mining:*

The method[7] look at the use of missing value and clustering algorithm for a data mining approach to help predict the crimes patterns and fast up the process of solving crime. MV algorithm and Apriori algorithm with some enhancements is used to aid in the process of filling the missing value and identification of crime patterns. These techniques are applied to real crime data. Also semi supervised learning technique is used for knowledge discovery from the crime records and to help increase the predictive accuracy.

B. *Detect Patterns of Crime:*

The aim is to automatically detect patterns of crime[24]. Among a large set of crimes that happen every year in a major city, it is challenging, time-consuming, and labor-intensive for crime analysts to determine which ones may have been committed by the same individual. If automated, data-driven tools for crime pattern detection are made available to assist analysts, these tools could help police to better understand patterns of crime, leading to more precise attribution of past crimes, and the apprehension of suspects. To do this, a pattern detection algorithm is proposed called *Series Finder* that grows a pattern of discovered crimes from within a database, starting from a "seed" of a few crimes. Series Finder incorporates both the common characteristics of all patterns and the unique aspects of each specific pattern.

C. *Area-Specific Crime Prediction Models:*

Another method is area-specific crime prediction models[25] based on hierarchical and multi-task statistical learning. These models will lessen sparseness by sharing information across ZIP codes, and retain the advantages of localized models in addressing non-homogeneous crime patterns. Out-of-sample testing on real crime data indicates predictive advantages over multiple state-of-the-art global models.

D. *Prediction of future crime locations:*

Spatial choice analysis can be used to discover the distribution of people's behaviors in space and time. Two adjusted spatial choice models that include models of

All rights reserved by www.ijsrd.com

decision making processes are presented. Results show that adjusted spatial choice models provide accurate predictions of future crime patterns and they can be used as the basis for a decision support system.

E. Grid-Based Crime Prediction Using Geographical Features:

Machine learning is useful for grid-based crime prediction[26]. Many studies have examined factors including time, space, and type of crime, but the geographic characteristics of the grid are rarely analyzed, leaving prediction models unable to predict crime displacement. The method incorporates the concept of a criminal environment in grid-based crime prediction modeling, and establishes a range of spatial-temporal features based on 84 types of geographic information by applying the Google Places API to theft data for a city in Taiwan.

F. Country crime analysis using the self-organizing map:

The self-organizing map (SOM) [21] is one of the widely used neural network algorithms. The purpose of this method is to apply the SOM to mapping countries with different situations of crime. A total of 56 countries and 28 variables considered. Results shows that some roughly definite patterns of crime situation is identified in traditionally homogeneous countries. In different countries, positive correlation on crime in some countries may have negative correlation in other countries. Overall, correlation of some factors on crime can still be concluded in most groups. SOM can be a new tool for mapping criminal phenomena through processing of large amounts of crime data.

G. Outlier detection and data association:

Outlier detection [4] has been extensively studied in the field of statistics, and a number of discordancy tests have been developed. In data mining, outliers are "meaningful input signals". In some cases, outliers represent unique characteristics of the objects, which are important to an organization. Law enforcement is one area where outlier detection is critically important. In law enforcement, we want to associate criminal incidents caused by the same person or group and detect outliers from this behavior. The purpose of the method is two-folded: to describe an outlier detection technique and to propose a data association method based upon this technique. Analysis is focused on categorical data since these data are typically found in crime analysis. First, an outlier score function is developed, and then a data association method is applied based on the outlier score function.

H. Crime Detection Technique Using Data Mining and K-Means:

The main objective of this method [5, 20] is to classify clustered crimes based on occurrence frequency during different years. Data mining is used extensively in terms of analysis, investigation and discovery of patterns for occurrence of different crimes. A theoretical model based on data mining techniques such as clustering and classification is applied to real crime dataset. Weights are assigned to the features in order to improve the quality of the model and remove low value of them. The Genetic Algorithm is used for optimizing of Outlier Detection operator parameters using Rapid Miner tool.

I. Crime Prediction and Forecasting using Clustering:

K-Means clustering, Agglomerative clustering and Density Based Spatial Clustering with Noise (DBSCAN) algorithms are used to cluster crime activities based on some predefined cases and the results of these clustering are compared to find the best suitable clustering algorithm for crime detection. The result of K-Means clustering algorithm is visualized using Google Map for interactive and easy understanding. The K-Nearest Neighbor (KNN) classification is used for crime prediction. The performance of each clustering algorithms are evaluated using the metrics such as precision, recall and Fmeasure, and the results are compared.

J. Crime Detection, Analysis & Prediction:

The system[27] is trained by feeding previous years record of crimes taken from legitimate online portal of India listing various crimes such as murder, kidnapping and abduction, dacoits, robbery, burglary, rape and other such crimes. As per data of Indian statistics, which gives data of various crime of past years, a regression model is created and the crime rate for the following years in various states can be predicted.

III. Related Works

Malathi and Baboo[7] used a classification technique (decision trees) to predict crime trends (out of four options). They also describe the prediction of the numbers of crimes for a particular year using data from the previous 8 years, but the method used for numeric prediction is not clearly mentioned. Saltos[18] used numerical prediction models and the data used is described in detail, both in terms of features and number of records. Another approach by Oatley and Ewart focused on the prediction of likelihood of repeated burglary for a particular property. For this purpose, they used a Bayesian belief network, using the following features or attributes: offender features; modus operandi features; property stolen; premise crime history; prevalence, incidence and concentration, which are numeric indicators of the distribution of crimes over an area. Saltos[18] approach focuses on the evaluation of prediction models, both in terms of their predictive performance, as well as their complexity, as an important practical aspect that is relevant for large volumes of data. Xue and Brown developed an approach for the prediction of future crime locations based on discrete choice theory and clustering. A classification approach has been used by Yu et al.[6] to classify areas into hot spots and cold spots, and to predict if an area will be a hot spot for residential burglary. They defined a hot spot as an area with at least 1 crime. They experimented with different levels of aggregation of historical data, and a variety of classification techniques: k-Nearest Neighbor (k-NN), Decision trees (J48 algorithm), Support Vector Machines (SVM), Neural Network, Naive Bayes and ensemble learning. They found that the best results were obtained with the 1-nearest neighbor and the neural network algorithms. Unlike previous research, they focused on the prediction of crime frequency as a numeric value rather than as a label (hot/cold spot), because the definition of a hot spot may vary. Some of the authors have discussed primary clustering (Chen et al.[6]; Kulis and

All rights reserved by www.ijsrd.com

Jordan[20]) and classification (Okonkwo and Enem 2011) techniques for crime detection, criminal identification theoretically; however, none of them provides a sound implementation for the same. Although some papers (Nath[3]; Malathi and Baboo[7]) discuss application of k-means for crime detection, but these and other works (Ehlers[19] ; Hussain et al. 2012) are deficient in integration among crime detection, criminal identification and prediction, and crime verification. Malathi et al.[8] work with crime attributes—number of crimes of a particular crime type, e.g., murder and burglary, versus years. Nath [3] tries to detect crime suspects based on their races, age and sex. Mande et al.[22] states that criminal identification is based on Gaussian mixture models. Jin et al. in their proposal define the position of crime events with longitude and latitude using k-means. Okonkwo and Enem[23] confer about terrorism attack as a type of crime. They focus on KNN's theoretical details, but there is no implementation provided. Li and Juhola[21] says that crime research is an area that can benefit from better visualization and DMT.

IV. Comparison

Technique	Method Used	Advantages	Disadvantages
Crime trend prediction	k-means, decision trees	Predict no.of crimes for a particular year	Method used for numeric prediction not specified
Discover associations between crimes	Association rule mining	Discover associations	Focused more on evaluation
Prediction of future crime locations	Clustering and discrete choice theory	Analyze and predict spatial choice of criminals	Focus mainly on residential crimes
Classify areas into hot spots and cold spots	KNN, decision trees, Naive Bayes	Prediction of crime frequency as a numeric value	More time is required to build and test models
Crime detection, criminal identification	k-means, decision trees	Clustering of crime data by crime type	Doesn't provide a sound implementation
k-means for crime detection	k-means, hierarchical clustering	Deceptive identity detection	Deficient in integration among crime detection, criminal identification, and prediction
Detecting crime suspects	k-means clustering	Six types of crime patterns are derived	Only race and age features are considered
Position of crime events with longitude and latitude	k-means	Efficient positioning	Only focused on plotting
Combating crime and terrorism	KNN	Largely focused on KNN's theoretical details	No implementation provided
Association of incidents for identification of crimes committed by same individual	Clustering and out lier-based approach	Identifies crimes committed by same individual	Robbery data is only considered
Predictive modeling	Clustering, social network analysis	Gives an overview of crime dataset predictive modeling	No implementation provided
Country crime analysis	Rule extraction, self-organizing map	Data analysis to support decision making	More focused on map implementation
Libyan crime data analysis	k-means, association rules	Efficient analysis of crime data	Less number of records
Patterns in house breaking crimes	Series finder	51 patterns identified	More time required

V. Conclusion

The analysis is helpful for studying and identifying crime data mining techniques. There are some limitations of each technique including computational efforts, structured data and rules. Based on the strength and weakness identified, it is concluded that each technique is specific to crime detection data mining scenario and has significant contribution to the field of crime detection. One of the facts identified is that crime data mining require data mining experts and data analysts equipped with sufficient knowledge of data mining and they need to collaborate with detectives in early phases of crime detection.

References

[1] K. Dahbur and T. Muscarello, Classification system for serial criminal patterns, Artificial Intelligence and Law 11(4) (2006) 251–269.
[2] Y. Peng, G. Kou, Y. Shi and Z. Chen, A descriptive framework for the field of data mining and knowledge discovery, International Journal of Information Technology & Decision Making 7(4) (2006) 639–682.
[3] S. V. Nath, Crime pattern detection using data mining, in Web Intelligence and Intelligent Agent Technology Workshops, 2006 IEEE/WIC/ACM International Conference on (IEEE, 2008), pp. 41–44.
[4] S. Lin and D. E. Brown, An outlier-based data association method for linking criminal incidents, Decision Support Systems 41(3) (2008),http://www.sciencedirect.com/science/article/pii/S0167923604001344

All rights reserved by www.ijsrd.com

[5] Chen H, Chung W, Xu JJ, Wang G, Qin Y, Chau M (2008) Crime data mining: a general framework and some examples. Comput IEEE 37(4):50–56.
[6] Yu G, Shao S, Luo B (2008) Mining crime data by using new similarity measure. In: 2nd International conference genetic and evolutionary computing (WGEC), IEEE, Washington, USA, pp 389–392.
[7] Malathi A, Baboo SS (2011) Evolving data mining algorithms on the prevailing crime trend—an intelligent crime prediction model. Int J Sci Eng Res 2(6).
[8] Malathi A, Babboo SS, Anbarasi A (2011) An intelligent analysis of a city crime data using data mining. In: International conference information electronic engineering, vol 6. IACSIT Press, Singapore, pp 130–134.
[9] J. Hosseinkhani, M. Koochakzaei, S. Keikhaee, and J. H. Naniz, "Detecting suspicion information on the Web using crime data mining techniques," International Journal of Advanced Computer Science and Information Technology, vol. 3, pp. 32-41, 2014.
[10] H. Hassani, X. Huang, E. S. Silva, and M. Ghodsi, "A review of data mining applications in crime," Statistical Analysis and Data Mining: The ASA Data Science Journal, vol. 9, pp. 139-154, 2016.
[11] S. Baluja, V. O. Mittal, and R. Sukthankar, "Applying Machine Learning for High-Performance Named-Entity Extraction," Computational Intelligence, vol. 16, pp. 586-595, 2000.
[12] A. Borthwick, J. Sterling, E. Agichtein, and R. Grishman, "Exploiting diverse knowledge sources via maximum entropy in named entity recognition," in Proc. of the Sixth Workshop on Very Large Corpora, 1998.
[13] S. Miller, M. Crystal, H. Fox, L. Ramshaw, R. Schwartz, R. Stone, et al., "Algorithms that learn to extract information: Bbn: Tipster phase iii," in Proceedings of a workshop on held at Baltimore, Maryland: October 13-15, 1998, 1998, pp. 75-89.
[14] I. H. Witten, Z. Bray, M. Mahoui, and W. J. Teahan, "Using language models for generic entity extraction," in Proceedings of the ICML Workshop on Text Mining, 1999.
[15] R. V. Hauck, H. Atabakhsb, P. Ongvasith, H. Gupta, and H. Chen, "Using Coplink to analyze criminal-justice data," Computer, vol. 35, pp. 30-37, 2002.
[16] R. T. Ng and J. Han, "Effcient and Effective Clustering Methods for Spatial Data Mining," in Proc. of, 1994, pp. 144-155.
[17] H. Yun, D. Ha, B. Hwang, and K. H. Ryu, "Mining association rules on significant rare data using relative support," Journal of Systems and Software, vol.67, pp. 181-191, 2003. University of Sindh Journal of Information and Communication Technology (USJICT), Vol.2 (1), pg.: 1-6.
[18] Saltos and Cocea, An exploration of crime prediction using data mining on open data, International journal of Information Technology and Decision making (2017), pp. 40–53.
[19] Ehlers D (1998) Predicting crime: a statistical glimpse of the future? Nedbank ISS Crime Index. Halfway House: Inst for Security Stud 2(2).
[20] Kulis B, Jordan MI (2011) Revisiting k-means: new algorithms via bayesian nonparametrics. In: 29th International conference machine learning. Omnipress, Edinburgh, pp 513–520.
[21] Li X, Juhola M (2014) Country crime analysis using self-organizing map, with special regard to demographic factors. AI Soc 29(1): 53–68.
[22] Mande U, Srinivas Y, Murthy JVR (2012a) Feature specific criminal mapping using data mining techniques and generalized Gaussian mixture model. Int J Comput Sci Commun Netw 2(3):375–379.
[23] Okonkwo RO, Enem FO (2011) Combating crime and terrorism using data mining techniques. In: 10th International conference IT people centred development, Nigeria Computer Society, Nigeria.
[24] Wang, G.; Chen, H.; and Atabakhsh, H. 2004. Automatically detecting deceptive criminal identities. *Communications of the ACM* 47(3):70–76.
[25] M. Al Boni and M. S. Gerber. Predicting crime with routine activity patterns inferred from social media. *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics*, 2016.
[26] Ying-Lung Lin, Meng-Feng Yen, Liang-Chih Yu. Grid-Based Crime Prediction Using Geographical Features. International Journal of Geo-Information, 2018.
[27] Sunil Yadav, Meet Timbadia, Ajit Yadav, Rohit Vishwakarma and Nikhilesh Yadav. Crime Pattern Detection, Analysis & Prediction. International Conference on Electronics, Communication and Aerospace Technology (ICECA), 2017.

All rights reserved by *www.ijsrd.com*