International Journal of Engineering Technologies and Management Research
A Knowledge Repository
IJETMR

AN APPROACH TO CRIME DATA ANALYSIS: A SYSTEMATIC REVIEW

Deepika Tyagi [*1], Dr. Sanjiv Sharma [*2]
[*1] Department of CSE, Madhav Institute of Technology and Science, Gwalior, India
[*2] Department of CSE / IT, Madhav Institute of Technology and Science, Gwalior, India

Abstract:
In the current era, number of crimes occurs in the society and this criminal rate increase day by day. There is tremendous growth of criminal data. Crime has negatively influenced the societies. Crime control is essential for the welfare, stability and development of society. Law enforcement agencies are seeking for the system to target crime structure efficiently. The intelligent crime data analysis provides the best understanding of the dynamics of unlawful activities, discovering patterns of criminal behavior that will be useful to understand where, when and why crimes can occur. There is a need for the advancements in the data storage collection, analysis and algorithm that can handle data and yield high accuracy. This paper demonstrates the data mining technologies which are used in criminal investigation. The contribution of this paper is to highlight the methodology used in crime data analytics. This paper summarizes the challenges arising during the analysis process, which should be removed to get the desired result.

Keywords: *Crime Data Analysis; Data Mining; Machine Learning; Big Data.*

Cite This Article: Deepika Tyagi, and Dr. Sanjiv Sharma. (2018). "AN APPROACH TO CRIME DATA ANALYSIS: A SYSTEMATIC REVIEW." *International Journal of Engineering Technologies and Management Research,* 5(2:SE), 67-74. DOI: 10.5281/zenodo.1197513.

1. Introduction

Crime is one of the concerning aspect of the society. Crimes affect our society in different ways. Crime investigation plays an important role in police system in the country. Criminal analysis and investigation is the process to explore and detect crime and criminals relationship [1]. There are lots of data related to the crime in police station records, information related to the particular crime or the essential information which is directly or indirectly related to crime should be extracted. So there is need of such technology, which separate all these data from huge content. On the basis of previously known (historical) crime and criminals relationship record, the criminal investigation team can extract useful information so that they can identify the facts related to the committed crime and minimize the future crime possibilities [2]. Criminal investigation acts on criminal cases like murder cases, child abuse, threats, hacking, financial crime detection like money laundering, terrorism funding, fraud, etc. So the criminal investigation team should use techniques so that they can predict the future crime trends on the basis of available historical criminal data and in this way the future crime rate will decrease. The need of criminal investigation is to identify and apprehend the criminal if a crime has been committed and provide the evidence to support a conviction in court.

The criminal investigation is the process to seek the methods, motives and identities of criminals and prove the guilt of a criminal. Crime investigate refers to the process to discover important information relevant to the crime. Investigation can be done by Evidence preservation, interviewing, record collection, electronic discovery, forensic anthropology, investigation and search warrants, email trace, criminal forensics, intelligence gathering, etc.

Big data analytics is the process to examine the huge amount of data to find hidden information patterns and trends [3]. It helps in cost reduction, faster and better decision making. It provides a framework for storing and analyzing huge amounts of unstructured criminal data in real time. An analytical system can cope up with predicting crimes. Investigation analytics system can deal with many information like text data, audio, video, DNA. Combining with security intelligence sources, it provides the information about the latest vulnerabilities and identify outliers and anomalies in security data. Big data security analytics can minimize flows of raw security events to a manageable number of alerts. It provides details to the investigator about the incident and its relationship with historical anomalies. Big data analytics can be used for analyzing the financial transaction, log files to identify suspicious activities.

Various technologies such as association, classification, clustering are used in criminal investigations in data mining. Crime investigation is done by using artificial intelligence methods. For predicting and matching crime incidences neural network, Bayesian networks and genetic algorithm are used. NLP approach is used in criminal investigation. Lots of work has been done in this field like mining criminal database to find investigation clues in the case of financial crime detection stolen automobiles. Integrative OSINT cyber crime investigation framework has been developed. In this paper different technique are described which improves the existing system which makes the criminal investigation process efficient, reduce the complexity & consume less time.

2. Literature Review

The historical review covers the beginning of a criminal investigation and its types. The motivation of this review is to gather knowledge about the work done in the past decades in the field of criminal investigation data mining. Reviewers concentrate not only on the reduction and prevention of crime but also enhancing the quality of criminal investigation. The authors recommend the continuation and further development in the concerned field.

- Brahan JW., Leung W. et al. (1998) [4] describes three specific applications as related to the human machine integration aspects of spatial temporal predictive modeling of the crime-map kiosk, of neural network learning for mug shot matching for facial sketch, and of fuzzy expert systems for money laundering detection.
- Zhou, F., B. Yang, L. Li and Z. Chen (2008) [5] Overview of the new types of intelligent decision support system.

- Li, S.T., S.C. Kuo and F.C. Tsai (2010) [6] design a decision support model using FSOM and rule extraction for crime prevention.
- The usability of ESOM and MDS as text exploration instruments in police investigations compared by Poelmans J, Viaene S, et al. (2011) [7]. Combine them with traditional classification instruments such as the SVM and Naïve Bayes. The possibilities offered by the ESOM and MDS are compared for iteratively enriching feature set, discovering confusing situations, faulty case labeling and significantly improving the classification accuracy. It demonstrate the use of MDS and ESOM for automating the detection of domestic violence from the unstructured texts comprising the police reports.
- Phillips, P. and I. Lee (2012) [8] figure out crime datasets to discover co-distribution patterns that may add to the formulation of the crime and proposed a graph based data illustration that allows to extract patterns from heterogeneous areal aggregated datasets and visualize the resulting patterns efficiently.
- Noor, N.M.M, et al. (2013) [9] proposed autoregressive Integrated Moving Average (ARIMA) model and fuzzy alpha-cut for crime forecasting. This combination is expected to generate more accurate forecasting result with minimum error. It will aid the decision maker's in making the right decision in crime prevention strategies.
- Maarten van Banerveld Nhien-An Le-Khac M Tahar Kechadi 2014 [10] uses Natural Language Processing (NLP) techniques to help criminal investigator handle large amount of textual information in a more efficient and faster way. It focuses on the evaluation its performance in terms of speed, smarter and easier for investigators.
- Wu, J. and D. Wang (2014) [11] The prior distribution of the incidence of crime is based on a large-scale of criminal investigation of historical data on their psychology problems and the new sample data is based on the measured people's sampled data of investigation on their psychological problems. The incidence of crime of the measured people is the posterior distribution of the measured that need to be predicted. With the application of Bayesian statistical methods, compute the incidence of the crime of the measured and provide a basis to judge whether the suspect is a criminal.
- Li, X. and M. Juhola (2014) [12] apply SOM to map countries with different situations of crime. In different countries, positive correlation on crime in some countries may have negative correlation in other countries. It proved that the SOM can be a tool for mapping criminal phenomena through processing of large amount of crime data.
- Noor Maizura Mohamad Noor, et al. (2015) [13] design architecture of Decision Support System for crime visualization. Due to difficulty in decision making for crime prevention, Decision Support System (DSS) and data mining approach can be used to resolve the problem. As a result, the architecture of DSS using visualization technique is proposed because it can represent the crime data into a more comprehensible presentation.
- Tayal, D.K., A. Jain, et al. (2015) [14] "Crime detection and criminal identification in India using data mining techniques".
- Morgan Burcher, Chad Whelan (2017) [15] describes the Social network analysis (SNA) as a tool for criminal intelligence. SNA is capable of revealing significant insights into the dynamics of dark networks, the identification of critical nodes, which then can be targeted by law enforcement and security agencies for disruption. The primary contribution of this is to call attention to the organizational characteristics of law enforcement agencies which can influence the capacity of criminal intelligence analysts

to successfully apply SNA as much as the often cited characteristics of criminal networks.

3. Types of Crime

Crime is divided into different categories such as traffic crime (While driving being alcoholic, property and lives destroyed by accident), fraud (identity deception, forgery, embezzlement, transactional money laundering), violent crime (attacker having weapons during robbery, criminal homicide, assaults, terrorism) and Cyber crime (illegal trading, theft of confidential information, internet fraud).

4. Data Sources

Law enforcement agencies collect data from different sources like telephone records, location based social networks like facebook, twitter, blogs, surveillance record, police records, financial transaction data for the crime investigation process.

5. Data Mining Techniques

Rate of criminal activities is increasing day by day, so there is need of data mining technologies used by law enforcement agencies [25]. The criminals develop networks in which they form groups or teams to carry out various illegal activities. The combination of data mining techniques is used to obtain more accuracy [16].

- Entity extraction: Entity extraction is the process of identifying the particular patterns so that they provide basic information for the crime analysis [17]. Entity extraction is used to extract valuable information of person address, time, vehicle, gender, crime type, personal property, suspect description relevant to particular cases automatically. It is the process to identify the potential suspects.
- Link analysis: Link analysis is used to analyze the criminal incident & forms a network of the suspect. Social network analysis is used to analyze the associated elements of criminals in the criminal network for disrupting the network [18]. Link analysis is used to find strong association between criminal objects [30].
- Classification: Classification technique is a supervised machine learning method. Classification divides the dataset based on some predefined condition. Classification is the process to specify the class of the object to which it belongs. Objects have characteristics, on the basis of which objects are classified into different categories. It is helpful in predicting previously known class of the upcoming objects in the future. In mail spamming classification is used. Algorithm such as C4.5, CART is used for detection of specific activities of the criminals in large sized data sets, classify the crime activities into different categories and predict crime hotspots.
- The K-nearest neighbor algorithm (K-NN) is a classification algorithm used for classifying the objects. It is the simplest machine learning algorithm. It is used to determine similarity between train and test record.
- Artificial neural network is the interconnection network of processing elements known as neurons. ANN mimics the cognitive, neurological functions of the human brain. Inputs

are multiplied by weights and produces outputs as labels. Neural network techniques are used for entity extraction from the criminal data records. Its prediction accuracy is high. ANN is used for pattern recognization, decision problem and prediction tasks. It is used to identify the crime hot spots of high level.

- The decision tree is tree like structure to demonstrates the flow of data where testing over the attributes, is performed at each node and on the basis of condition correctly label the objects. The decision tree is used to detect suspicious mails and provide accuracy in classifying emails.
- Support vector machine (SVM) is a supervised machine learning algorithm that is used for classification problems by separating hyperplane. SVM uses the Kernel function.
- The Naïve Bayes Belief network is a probabilistic model that is used for the classification [22]. Bayesian theorem is also used in crime analysis. Its accuracy is good. It demonstrates the variable using directed acyclic graph.
- Clustering: Clustering is unsupervised machine learning algorithm. Clustering algorithm is used to group the records of similar type and dissimilar type of objects are grouped in different groups [19]. Clustering refers to the process of grouping the objects into unknown label classes. Clustering provides efficiency in identifying crime zones and trends and in this way crimes can be controlled. Self organizing map [12], link analysis technique such as (Shortest path algorithm) [20], hierarchical clustering, DB Scan, K-means clustering to detect hotspots, to automatically identify the association. SOM is a type of ANN that uses unsupervised learning that transform the attributes and produces results.
- K-mean clustering is used to partition the data into k- clusters based on their mean. Agglomerative algorithm and partitional algorithm are used for hierarchical clustering.
- Association rule mining: Association rule mining is unsupervised learning method. Rules are generated by association rule mining, based on the frequent occurrence of crime patterns from criminal dataset to help decision makers to take decision for the prevention of the society from criminal activities. Apriori algorithm, Outlier score function, Frequent pattern growth, Temporal association rule are used to link crime incident, possible suspect, provide informative association between crime identities and discover crime patterns. Fuzzy based system is also used in crime domain for the knowledge discovery. Its accuracy is very high. It is based on fuzzy logic. Its performance is better in time space domain.
- An Intelligent agent is a computing agent which performs tasks autonomously. Agents monitor & identify the real time response and generate alert messages through emails & instant messages [2]. When deformities are encountered, then agents deliver messages to alert the criminal investigators. It increases the efficiency & accuracy in criminal investigation.
- Text mining: Text mining is used to extract information from textual dataset [23]. Natural language processing is used to identify the relevant entities. It compares phrases or sentences to extract associations within the criminal network.

6. Methodology

CRISP-DM methodology stands for Cross industry standard process for data mining. It consists of business understanding, data understanding, data preparation, modeling, evaluation. Law enforcement data are collected from heterogeneous sources. Understand the data available in the crime record on which the processing task is to be performed to identify criminal trends, type of crime and crime zone in order to predict the hot spots of future criminal activities so that crime rate can be reduced. As the collected law enforcement data are in many formats, data preprocessing is performed to improve the quality of dataset to obtain the desired accurate result efficiently. The model is to be designed to perform the operation such as feature selection, clustering, analysis, prediction and then evaluation are performed and visualization of results is done by graphs using visualization tools. Block diagram of criminal data analysis is represented in fig 1.

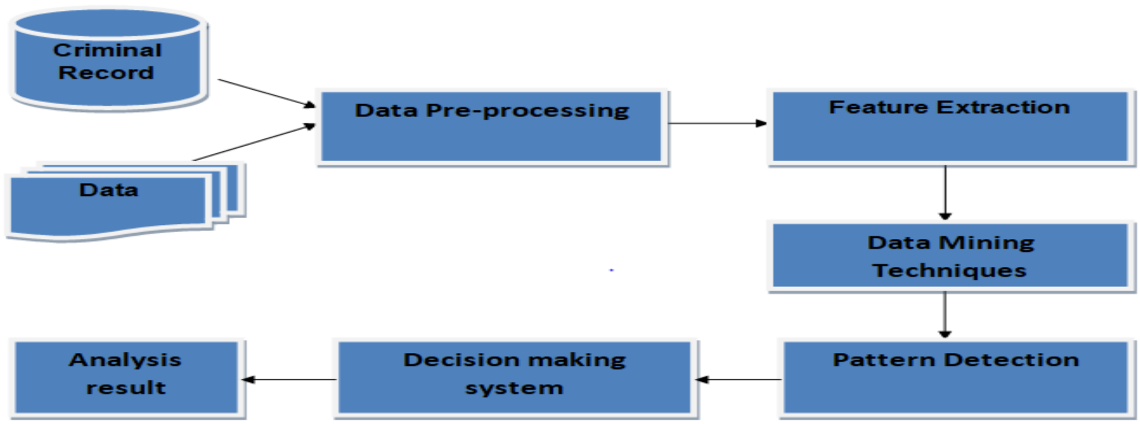

Figure 1: Block diagram of crime data Analysis

7. Comparison Study

SVM is used for identifying digital evidences related to computer crime. ANN provides higher accuracy than logistic regression when logistic applied to identify smuggling vessels. The SVM approach provides better accuracy as compared to multilayer perceptron neural network. Artificial neural networks (ANNs), decision trees and logistic regression are used for uncovering lies from statements of different types of crimes. These data mining techniques are used for auto-insurance fraud.

Nearest Neighbor, Decision tree (J48), Support Vector Machine (SVM) Naïve Bayes and Neural network are applied in [24]. Neural network performs better as compared to J48 decision tree and SVM. Neural network provides accuracy.

8. Challenges

- Criminal data are available in different formats, thus tackling the variety of data formats from multiple data sources and transforming the data into a desirable form so that result can be obtained, is also a challenging task.

- There is another challenging aspect of storing the law enforcement data which are in large amount of size. As the size of the data is huge, so there is need of storage devices which have large capability to store the data.
- There are different analytical models are available, but appropriate analytical model is to be selected for data analyzing purpose and this is a challenging task.
- Some other challenging factors are also exists, such as matching data mining technique and methodology, exploring proper integration methods to tackle complicated investigation problem.
- Complexity is another great challenge as the data reduction techniques are prior to the analyzing task.

9. Conclusion

There are a number of factors responsible for the rising of crimes at an alarming rate in India like illiteracy, poverty, unemployment, migration, frustration & corruption. Intelligence agencies search the database manually, which is a tedious task and consume more time. New advanced technologies & tools are used for combating crimes and to identify criminals. This paper reviewed the use of data mining techniques and tools for identifying crime patterns. New methodologies and analytical techniques should be explored to address the fundamental challenges of criminal data, and to leverage big data to facilitate criminal investigations. Big data analytics has the potential to transform the way that law enforcement and security intelligence agencies extract vital knowledge (e.g., criminal networks) from multiple data sources in real-time to support their investigations.

References

[1] xu j, chen h., " criminal network analysis and visualization," community acm, 2005, 48:100-107.
[2] wang t, rudin c, wagner d, sevieri r., "learning to detect patterns of crime." Joint european conference on machine learning and knowledge discovery in databases, springer, 2013 515-530.
[3] m.i.pramanik, raymond y.k. lau, wei t.yue, yunming ye and chunping li., "big data analytics for security and criminal investigations" wiley interdisciplinary reviews-data mining and knowledge discovery 2017 vol.7 no.4,1-19.
[4] brahan jw, lam kp, chan h, leung w. "aicams: artificial intelligence crime analysis and management system," knowledge system, 1998, 11: 355-361.
[5] zhou, f., b. Yang, l. Li and z. Chen "overview of the new types of intelligent decision support system," international conference on innovative computing information and control ieee 2008 267-272.
[6] li, s.t., s.c. kuo and f.c. tsai. "an intelligent decision-support model using fsom and rule extraction for crime prevention" expert system with applications, 2010, 37:7108-7119.
[7] poelmans j, van hulle, mm, viaene, s, elzinga p, dedene g, "text mining with emergent selforganizing maps and multi-dimensional scaling: a comparative study on domestic violence," applied soft computing 2011 11:3870-3876.
[8] phillips, p. And i. Lee, "mining co-distribution patterns for large crime datasets," expert system with applications 2012 39:11556-11563.
[9] noor, n.m.m, retnowardhani a, m.l. abd, saman mmy "crime forecasting using arima model and fuzzy alpha-cut," journal of applied sciences, 2013, 13:167-172.

[10] maarten v banerveld, nhien-an le-khac, m-tahar kechadi, "performance evaluation of a natural language processing approach applied in white collar crime investigation," international conference on future data and security engineering, springer, 2014, 19-21.
[11] wu, j. And d. Wang "the research based on bayesian behavior recognition technology," journal of applied sciences, applied mechanics and materials, 2014, 543-547 2167-2170.
[12] li, x. And m. Juhola "country crime analysis using the self-organizing map, with special regard to demographic factors," artificial intelligence and society, springer, 2014, 29(1) 53-68
[13] noor maizura mohamad noor, siti haslini ab hamid, rosmayati mohemad, masita muhammad suzuri hitam. "Architecture of decision support system for crime visualization" journal of applied sciences 2015 15:1176-1183.
[14] tayal d.k., a. Jain, surbhi arora, surbhi agarwal, tushar gupta "crime detection and criminal identification in india using data mining techniques" ai & society springer, 2015, 30(1):117-127.
[15] morgan burcher, chad whelan, "social network analysis as a tool for criminal intelligence: understanding its potential from the perspectives of intelligence analysts," springer 2017 1084-4791 1936-4830.
[16] chen h, chung w, xu jj, wang g, qin y, chau m. "crime data mining: a general framework and some examples," computer, ieee, 2004 37:50-56.
[17] ku ch, iriberri a, leroy g., "crime information extraction from police and witness narrative reports", ieee international conference on technologies for homeland security, ieee, 2008, 193-198.
[18] lu, y, luo x, polgar m, cao y , "social network analysis of a criminal hacker community," journal computing information system, 2010 51: 31-41.
[19] brown, de, gunderson, lf, "using clustering to discover the preferences of computer criminals," ieee transaction system man cybern a syst hum, 2001, 31:311-318.
[20] xu, jennifer j, chen h, "fighting organized crimes: using shortest- path algorithms to identify associations in criminal networks," decision support system, 2004, 38:473-487.
[21] nath, sv, "crime pattern detection using data mining," ieee/wic/acm international conference on web intelligence and intelligent agent technology workshop, 2006, 41-44.
[22] baumgartner k, ferrari s, palermo g, "constructing bayesian networks for criminal profiling from limited data," knowledge system, 2008, 21:563-572.
[23] tseng th, ho, zp, yang, ks, chen, cc., "mining term networks from text collections for crime investigation," expert system with applications 2012 39:10082-10090.
[24] chung-hsien yu, max w. Ward, melissa morabito, wei ding, "crime forecasting using data mining techniques," international conference on data mining workshops, ieee, 2011, 978-0-7695-4409-0.

*Corresponding author.
E-mail address: tyagideepika39@ gmail.com