SURVEY ON CRIME ANALYSIS AND PREDICTION USING DATA MINING TECHNIQUES

Article · April 2017

DOI: 10.21917/ijsc.2017.0202

2 authors, including:

Benjamin Fredrick David. H
K.R. College of Arts and Science
11 PUBLICATIONS **15** CITATIONS

SEE PROFILE

ISSN: 2229-6959 (ONLINE)
DOI: 10.21917/ijsc.2017.0202

SURVEY ON CRIME ANALYSIS AND PREDICTION USING DATA MINING TECHNIQUES

H. Benjamin Fredrick David[1] and A. Suruliandi[2]
Department of Computer Science and Engineering, Manonmaniam Sundaranar University, India

Abstract

Data Mining is the procedure which includes evaluating and examining large pre-existing databases in order to generate new information which may be essential to the organization. The extraction of new information is predicted using the existing datasets. Many approaches for analysis and prediction in data mining had been performed. But, many few efforts has made in the criminology field. Many few have taken efforts for comparing the information all these approaches produce. The police stations and other similar criminal justice agencies hold many large databases of information which can be used to predict or analyze the criminal movements and criminal activity involvement in the society. The criminals can also be predicted based on the crime data. The main aim of this work is to perform a survey on the supervised learning and unsupervised learning techniques that has been applied towards criminal identification. This paper presents the survey on the Crime analysis and crime prediction using several Data Mining techniques.

Keywords:

Criminology, Crime Analysis, Crime Prediction, Data Mining

1. INTRODUCTION

Historically solving crimes has been the right of the criminal justice and law enforcement specialists. With the increase in the use of the computerized systems to track crimes and trace criminals, computer data analysts have started lending their hands in helping the law enforcement officers and detectives to speed up the process of solving crimes. Criminology is process that is used to identify crime and criminal characteristics. The criminals and the crime occurrence possibility can be assessed with the help of criminology techniques. The criminology aids the police department, the detective agencies and crime branches in identifying the true characteristics of a criminal. The criminology department has been used in the proceedings of crime tracking ever since 1800. Crimes are a social nuisance and cost our society dearly in several ways. Even, the Indian Government has taken steps to develop applications and software for the use of State and Central Police in relation with the National Crime Records Bureau (NCRB) [27]. Any research that can help in solving crimes faster will pay for itself. About 10% of the criminals commit about 50% of the crimes [15]. People who study criminology will be able to identify the criminals based on the traces, characteristics and methods of crime which can be collected from the crime scene. In the middle of 1990s, data mining came into existence as a strong tool to extract useful information from large datasets and find the relationship between the attributes of the data [11]. Data mining originally came from statistics and machine learning as an interdisciplinary field, but then it was grown a lot that in 2001 it was considered as one of the top 10 leading technologies which will change the world [12]. According to many researchers such as Nath [23], solving crimes is a difficult and time consuming task that requires human intelligence and experience and data mining is one technique that can help us with crime detection problems. For solving crimes faster we have to develop a data mining paradigm that performs an interdisciplinary approach between computer science and criminal justice. As said earlier, the Criminology is a process that aims to identify crime characteristics and it is one of the most important fields for applying data mining. By using this, data mining algorithms will be able to produce crime reports and help in the identification of criminals much faster than any human could. Because of this remarkable feature, there is a growing demand for data mining in criminology. Actually, Crime analysis is a process which includes exploring the behavior of the crimes, detecting crimes and their relationships with criminals. The huge volume of crime and criminal datasets and the complexity of relationships between these kinds of information have made criminology an appropriate field for applying data mining techniques. Identifying crime characteristics is the first step for proceeding with any further analysis. The quality of data analysis depends greatly on background knowledge of analyst. A criminal can range from civil infractions such as illegal driving to terrorism mass murder such as the 9/11 attacks, therefore it is difficult to model the perfect algorithm to cover all of them [21]. The knowledge that is gained from Data Mining approaches is a very useful and this can help and support, the police. More specifically, we can use classification and clustering based models to help in identification of crime patterns and criminals. The wide range of data mining applications in the criminology has made it an important field of research. Data mining systems have played as a key role in assisting humans in this forensic domain and criminology domain. This makes it one of the most challenging decision-making environments for research.

The motivation for proceeding with this survey work is to aid a helping hand to the young researchers who are performing their research in criminal analysis and crime prediction areas. The paper is organized in such a manner to provide insights about the crime analysis procedure and then produce different types of crime analysis operations and those which can be applied together for producing an end user product which can be applied to the crime analysis in any police stations and detective agencies. This work will be a valuable reference to those who precede their research work in the crime analysis and Crime prediction using data mining techniques.

This survey paper is organized in such a manner for easy understanding of the concepts. The general crime analysis procedure is discussed in section 2. The Criminal analysis methods are discussed in the section 3 which will include all the different types of methods grouped under their own categories. Finally, section 4 gives the Qualitative analysis of the Crime Analysis and Prediction techniques and section 5 gives the

Quantitative analysis of the Crime Analysis and Prediction techniques.

2. CRIME ANALYSIS PROCEDURE

Usually, the crime analysis tasks can be a tedious process for the police or the investigation team to work with. The criminals when leaving the crime scene does leave some traces which can be used as a clue to identify the criminals. The crime sequence and the patterns which several criminals follow when committing a crime make it easy for analyzing the crime. This process includes several procedures to be followed in order to identify the criminals and getting more information based only on the clues or information given by the local people. The criminal can be analyzed based on the information from the crime scene which is tested against the previous crime patterns and judging by the method which is implied to test and proceed with the information that can affect the prediction results. The prediction can be further made useful for detecting the crimes in advance or by adding more cops to the sensitive areas which are identified by the system .The police stations can put up special force when there are chances for crime ahead of time. This type of the system will ensure there are peace and prosperity among the citizens.

The crime analysis can be performed procedure which is similar to figure Fig.1 which specifies each module which is used for machine learning to predict the crime or form group of clusters of criminals according to crime records. The criminals can hold certain properties and their crime characteristics and crime careers may vary from one criminal to another. Such a type of information can be taken as the input dataset. The input dataset is given to a pre-processor which performs the preprocessing based on the requirements. Once the pre processing is completed the features or attributes from those information are extracted which may be in the form of text content from emails, the crime factors for a day, criminal characteristics, geo-location of the criminal, etc,. The pre processed result is further given to the classification algorithm or the clustering algorithm based on the requirements. The requirements may be anything from selecting the crime prone areas to predicting the criminal based on the previous crime records.

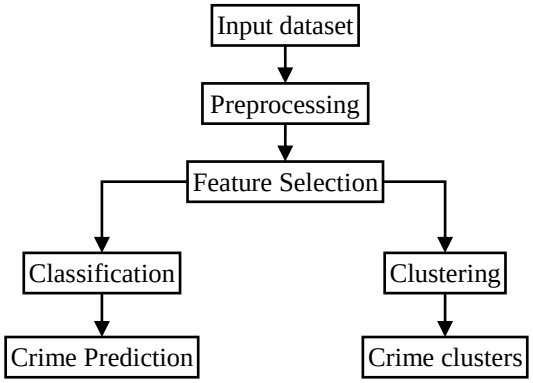

Fig.1. Crime Prediction and Crime clustering based on the input dataset

The classification algorithm works in a supervised learning manner in which the training and testing phase is required in order to train the classifier to identify the new unknown crime record. This is known as prediction. Whereas the clustering algorithm works in an un-supervised learning manner which automatically separates the crime records based on the number of groups to be created. The groups created in such a manner are known as clusters. Such a type of design can be a general template for applying crime prediction and crime analysis based on data mining algorithms.

3. CRIMINAL ANALYSIS METHODS

3.1 TEXT, CONTENT AND NLP-BASED METHODS

Sharma [1] proposed a concept which depicts zero crime in the society. For detecting the suspicious criminal activities, he has concentrated on the importance of data mining technology and designed a proactive application for that purpose. In his paper, he proposed a tool which applies an enhanced Decision Tree Algorithm to detect the suspicious e-mails about the criminal activities. An improved ID3 Algorithm with an enhanced feature selection method and attribute-importance factor is applied to produce a better and faster Decision Tree based on the information entropy which is explicitly derived from a series of training data sets from several classes. He proposed a new algorithm which is a combination of Advanced ID3 classification algorithm and enhanced feature selection method for the better efficiency of the algorithm.

Hamdy et al. [8] described an approach based on the people's interaction with social networks and mobile usage such as location markers and call logs. Their work also introduced a model for detecting suspicious behavior based on social network feeds and it not only describes a new method using the social interaction of people but, their work proposes a new system to help crime analysis create faster and precise decisions. The suspicious movement of the entity can be determined using the sequence of inference rules. Their constructed model is able to predict and characterize human behavior from reality data sources

3.2 CRIME PATTERNS AND EVIDENCE-BASED METHODS

Bogahawatte and Adikari [2] proposed an approach in which they highlighted the usage of data mining techniques, clustering and classification for effective investigation of crimes and criminal identification by developing a system named Intelligent Crime Investigation System (ICSIS) that could identify a criminal based up on the evidence collected from the crime location. They used clustering to identify the crime patterns which are used to commit crimes knowing the fact that each crime has certain patterns. The database is trained with a supervised learning algorithm, Naïve Bayes to predict possible suspects from the criminal records. His approach includes developing a multi-agent for crime pattern identification. There are agents for the place, time, role trademark and substance of criminals which separates the role of the criminals in components. The system is a multi-agent system and made with managed Java Beans. It makes it easy to encapsulate the requested entities in the work into objects and returns it to the bean for exposing properties. Classifying the criminals/ suspects is based on the Naïve Bayes classifier for identifying most possible suspects from crime data. Clustering the criminals is based on the model to help to identify patterns of committing crimes.

Agarwal et al. [3] used the rapid miner tool for analyzing the crime rates and anticipation of crime rate using different data

mining techniques. Their work done is for crime analysis using the K-Means Clustering algorithm. The main objective of their crime analysis work is to extract the crime patterns, predict the crime based on the spatial distribution of existing data and detection of crime. Their analysis includes the tracking homicide crime rates from one year to the next

Kiani et al. [4] performed a crime analysis work based on the clustering and classification techniques. Their work includes the extraction of crime patterns by crime analysis based on available criminal information, prediction of crimes based on the spatial distribution of existing data and crime recognition. They proposed a model in which the analysis and prediction of crimes are done through the optimization of outlier detection operator parameters which is performed through the Genetic Algorithm. The features are weighted in this model and the low-value features were deleted through selecting a suitable threshold. After which the clusters are clustered by the k-means clustering algorithm for classification of crime dataset.

Satyadevan et al. [5] has done a work which will display high probability for crime occurrence and can visualize crime prone areas. Instead of just focusing on the crime occurrences, they are focusing mainly on the crime factors of each day. They used the Naïve Bayes, Logistic Regression and SVM classifiers for classification of crime patterns and crime factors of each day. Their method consists of a pattern identification phase which can identify the trends and patterns in crime using the Apriori Algorithm. The prediction of crime spots is done with the help of Decision Tree algorithm which will detect the crime possible areas and their patterns.

Bruin et al. [7] proposed a technique which is used to determine the clustering of criminals based on the criminal careers. The criminal profile per offense per year is extracted from the database and a profile distance is calculated. After that, the distance matrix in profile per year is created. The distance matrix including the frequency value is made to form clusters by using naïve clustering algorithm. They made a criminal profile which is established in a way of representing the crime profile of an offender for a single year. With this information, the large group of criminals is easily analyzed and they predicted the future behavior of individual suspects. It will be useful for establishing the clear picture on different existing types of criminal careers They tested the tool on actual Dutch National Criminal Record Database for extracting the factors for identifying the criminal careers of a person.

3.3 SPATIAL AND GEO-LOCATION BASED METHODS

Huang et al. [6] focused on a different approach for criminal activity prediction based on mining location based Social Network interactions. By using these interactions, they can collect information using the geographical interactions and data collections from the people. They devised a working procedure in which a series of features are categorized from the Foursquare and Gowalla used in the San Francisco Bay area. The crime patterns and the crime occurrences are tracked with the geographical features which are extracted from the map and they are analyzed to detect the urban areas with high crime activities. Their work aims at exploiting the location-based social network data to investigate the criminal activities in urban areas. By using the Haversine formula the distance between the two points i.e. the crime location and venue location is calculated and shown in the Google Maps API and OpenStreetMap.

Chen [19] have presented a general framework for crime data mining that draws on experience gained with the Coplink project with the researchers at Arizona and their work mainly focuses on showing the relationships between crime types and the link between the criminal organizations. They used a concept space approach which will extract criminal from the incident summaries.

Yu [20] have discussed the preliminary results of a crime forecasting model developed in collaboration with the police department of a United States city in the Northeast. Their approach is to architect datasets from original crime records. The datasets contain aggregated counts of crime and crime-related events categorized by the police department. The location and time of these events is embedded in the data. Additional spatial and temporal features are harvested from the raw data set. Second, an ensemble of data mining classification techniques is employed to perform the crime forecasting. Then they analyzed a variety of classification methods to determine which is best for predicting crime “hotspots”. They even investigated classification on increase or emergence. Last, they have proposed the best forecasting approach which is aimed at achieving the most stable outcomes.

Rizwan et al. [22] have performed classification of crime dataset to predict Crime Category for different states of the United States of America. The crime dataset that they used in this research is real in nature. That is, it was collected from socio-economic data from 1990 US Census, law enforcement data from the 1990 US LEMAS survey, and crime data from the 1995 FBI UCR. Their work compared the two different classification algorithms namely, Naïve Bayesian and Decision Tree for predicting Crime Category for different states in USA. The results from their experiment showed that, Decision Tree algorithm out performed Naïve Bayesian algorithm and achieved 83.9519% Accuracy in predicting Crime Category for different states of USA.

Donald [24] have proposed a system for Crime Analysis which was named by them as The Regional Crime Analysis Program (ReCAP) system. It was designed by them as a computer application designed to aid local police forces (e.g. University of Virginia (UVA), City of Charlottesville, and Albemarle County) in the analysis and prevention of crime. ReCAP works in cooperation with the Pistol 2000 records management system, which aggregated and housed all of the crime information from a region. Their research and development was primarily focused on the individual components of the system which includes a database, geographic information system (GIS), and data mining tools which consisted of data mining alrogithms which produced spatial mining results over the crime hotspots. Their system consists of the seamless integration of all the components in the system.

3.4 PRISONER BASED METHODS

Sheehy et al. [10] came up with a research idea which was geared towards the treatment of the mentally ill people inside the prison. According to their work, the mentally ill criminals are identified using their Social Security Number (SSN) with all the criminal personal records and their crime career records attached. As the outcome, the Criminals are classified into “high”, “medium” and “low” levels of recidivism risk potential according

to their mental health. Their objective was to describe and classify the criminals into a misdemeanor and a felony which can be referred and not referred based on the mental health of the criminals. Their ill activities are monitored and data collection is continuous. By these, the criminals can be separated from other criminals who are hazardous and those who can cause damage to other inmates along with them. Further, their study also involves the classification of the mental health of the criminals into two categories i.e. "referred" and "not-referred". This helps the guards to identify the prisoners who are referred for the mental health check-up. The research work they had undergone will provide a summary of the inmates who are seriously mentally ill and those who are to be separated from the other inmates.

3.5 COMMUNICATION BASED METHODS

Taha et al. [9] has developed a forensic investigation tool for identifying the influential members who create an impact in a criminal organization. The immediate leaders can also be identified in a criminal organization. Removing these influential members can weaken the strength of the criminal organization. Their work is based on this methodology. They proposed a new work which is known as SIIMCO which first constructs the graph representing the criminal group or organization as a network from either mobile communication data of the criminal organization or based on the crime records. The system works on the basis of the created networks. These networks represent the criminal organization or crime incident reports. The vertex represents the individual criminals and the link represents the relationships or communication link between those two criminals. They employed certain formulas that quantify the degree of influence/ importance of each vertex in the network relative to all other vertices i.e. criminals in the graph. Based on this their system identifies the immediate leaders with the weighted graph which connects the criminals and identify them for further processing.

4. QUALITATIVE ANALYSIS OF CRIME ANALYSIS AND PREDICTION APPROACHES

The prediction can be made based on the Textual information or the Geospatial information or even the prisoner records which were manually recorded. By using the real open data such as internet, social feeds and messages the researcher can use the text processing or NLP techniques to mine information from the data and categorize the e-mails, messages or posts into a suspicious or a non-suspecious record [1]. Whereas in the Spatial mining area, the extraction of features from SNAP Gowalla dataset, DataSF criminal dataset up to February 2015 provides the way to plot the crime occurences on the Google Map which is interpreted easily. The communication based methods describe the identification of the leaders in a criminal organization may be a tedious process. Kamal Taha et al. [9] produced an approach through the phone calls and other communication data such as call logs and records, the influential members on a crime organization can be tracked. Kevin Sheehy, Thomas Rehbreger, Andrew O'Shea, William Hammond, Charlotte Blais, Michael Smith K., Preston White, Jr., Neal Goodloe [10] introduced an approach to categorize and identify the mentally ill prisoners among the prisoners and keep them separate from other prisoners to avoid conflict and injuries between them. Even though there are many methods for analyzing the crimes , this paper concludes many results based on the qualitative analysis When considering the Text/NLP based methods, Hamdy et al. [8] overcame the defects from the work of Sharma [1] based on many factors such as implementation of preprocessing for the data and extraction of relevant features. Both the paper labels the outcome based on suspicious activity. Mugdha Sharma [1] used an enhanced ID3 algorithm whereas the work produced by Ehab Hamdy, Ammar adl, Aboul Ella Hassanien, Osman Hegazy and Tai-Hoon Kim. [8] does not specifies the classification algorithm. The weakned of this paper is mostly about not giving the clear view of the pre processing and classification alogorithm. When considering crime patterns and evidence based methods, there are clustering and classification based papers. Bogahawatte and Adikari [2] concentrated on using the Naïve Bayes for finding out most possible suspect. Jyoti Agarwal et al. [3] on the other hand focused on crime analysis by implementing the K-Means clustering algorithm on crime dataset using rapid miner tool and the author had performed the crime analysis by considering the homicide crimes and plotting it with respect to year. Kiani et al. [4] concentrated on using the Genetic Algorithm to optimize the distance operator parameter of the decision tree using GINI index. The clustering of the criminal careers has been effectively done in the work [7]. Whereas, Shiju Satyadevan, et al. [5] have performed a comparison of the Naïve Bayes, SVM, Logistic Regresion and Decision Tree. This paper presents the crime prone regions and represented as heatmaps which indicate the level of heat. When considering the Spatial and Geolocation based methods, all these methods are analysed based on qualitative manner and the analysis information is described in the below mentioned table Table.1.

Table.1. Qualitative Analysis of Crime Analysis and Prediction

METHOD		INPUT	DATASET USED	PRE PROCESSING	FEATURE EXTRACTION	CLASSIFICATION/ CLUSTERING	STRENGTH	WEAKNESS	OUTCOME
Text/ NLP-based methods	[1]	E-mail messages	Real and open emails sent by terrorists and some are dummy emails	Nil	Selection of a subset of the original text containing "kill", "death", "bomb", "guns", "blasts"	Enhanced ID3 Decision Tree algorithm	Introducing attribute importance as a factor before information gain in the decision tree	Nil	Labeling email as Suspicious, Non-suspicious, and May be suspicious
	[8]	Crime history, age, previous	Device sensors, Security	Structuring collective data into {Time,	Similarity matching for sensory images	A trained classification model is used to predict the	Consideration of location feeds and	Not giving a clear view of the processing	Suspicious behavior to three levels

		arrests, Modus Operandi, countries visited, place of birth, Average use of ATM, Types of crimes, Entrance with respect to Time of Day, Crime areas, Victims' mistakes	camera information, Messages, Audio feeds, Social network posts and messages	Final Movement, Frequency rate, Video, Images, Audio }	using sliding window. Text semantic Analysis of the text information performed using Lexical processing, Natural Language Processing (NLP).	similarity of a given input to the suspicious item or location.	mobile usage information	and comparison of criminal behavior.	such as "High", "Medium" and "Low"
Crime patterns and Evidence-based methods	[2]	Crime evidences including many attributes like crime scene, day, month, offense, resources used, time, role in crime, transportation etc.,	Colombo crime and criminal records	Nil	Extraction of evidence	Clustering based model to identify patterns of committing crimes. Naïve Bayes classifier applied to find most possible suspect	Uses Naïve Bayes so this can be even suitable for small datasets.	No clear view of clustering method and Prisoner verification	Finding Categories as robbery, burglary, and theft Classifying person as "suspect" and after judgment "criminal"
	[3]	Homicide crimes and their occurrences	Crime dataset for crime analysis by polices in England and Wales from 1990 – 2011-12	Nil	Extraction of crime patterns based on the available crime and criminal data	K-means clustering algorithm	Produces year wise clusters of homicide crimes committed	Concentration is only on clustering of homicide crimes	Year and analysis of variation in clusters formed
	[4]	Burglary, Robbery, and Homicide	Crime dataset for crime analysis by polices in England and Wales from 1990 – 2011	Nil	Filtering of dataset, Outlier detection using distance operator (k-NN), Genetic Algorithm used for optimizing of outlier detection operator parameters	Classification was done using Decision Tree using GINI index and the testing and training done using Sample Stratified	Use of GA to optimize the distance operator parameters in Clustering and Predict the cluster's members based on classification using Decision Tree	The number of clusters in the clustering process needs to be optimized and further optimization of the technique needs to be done	The results for the optimized and non-optimized parameters were compared to show the difference in quality and effectiveness
	[5]	location, date, type of crime data extracted from Websites, Blogs, Social Media, RSS Feeds	Websites, Spatial Information, and date about crimes	Nil	Extraction of the following crime data related to "vandalism", "murder", "robbery", "burglary", "sex abuse", "gang rape", "arson", "armed robbery",	Naïve Bayes, SVM, Logistic regression Crime prediction was done using decision tree which is done using sample police complaints	Comparison of Naïve Bayes with SVM. Decision Tree is easy to interpret and understand for crime spot identification.	Not predicting the time in which the crime is happening.	The crime-prone areas (regions) are graphically represented using a heat map which indicates the level of crimes

					"highway robbery", "snatching"				
	[7]	Crime database and criminal information	National Crime Record Database	Nil	Crime nature, frequency, duration, severity	Crime profile of offender for single year is determined for comparison and he	Development of new distance measures with combination of profile distance with crime frequency of criminals	The runtime of the chosen approach is not optimal	Clustering of criminal careers based on the nature. One time criminal, severe criminals and minor career criminals
Spatial and Geo-location based methods	[6]	Geo-location and Crime Type	SNAP Gowalla dataset, DataSF criminal dataset up to February 2015	Extraction of crime type like Assault, Robbery, Theft, Vandalism, Drug	Geographical features, Popularity, Location category, Neighbor entropy, Social Tightness density, crime location, venue from Foursquare	Random Forest(RF), Linear Regression (LR) and Support Vector Machine (SVM)	Random Split method utilized with 80% for training and 20% for testing in classification	Nil	Crime Areas plotted using Google Map API and OpenStreetMap in San Francisco Bay area and Criminal pattern discovery according to the context of user activity and location-based social networks. Predict crime frequency and find which crime is to be more difficult or easier to be predicted
Communication based methods	[9]	Flow of communications/information links between two criminals (e.g., phone call records, messages, etc.), names of criminals/suspects, the type of crime, location and date of the crime.	Real-world communication records (DBLP, Enron email dataset, Nodobo mobile phone records dataset)	Creating the graph based on the data and then assigning weight to a vertex based on its number of communication attempts in the criminal graph	The immediate leaders of lower-level criminals and the lower-level criminals themselves are extracted.	Evaluation of the accuracy of the three systems by measuring their Recall, Precision, and Euclidean Distance.	Evaluated SIIMCO by comparing it experimentally with CrimeNet Explorer and LogAnalysis	Nil	System can identify the influential members of a criminal organization and the immediate leaders of a given list of lower-level criminals
Prisoner based methods	[10]	The Social Security Number (SSN) with all the criminal personal and crime career records.	Albemarle-Charlottesville Regional Jail (ACRJ), Jefferson Area Community Corrections (JACC) and	A combination which includes the Social Security Number (SSN) and date was used to link the databases together.	age, criminal history, employment history, crime type := "assault", "larceny", "supervision violations", "narcotics	Offenders are classified into three classes namely "high", "medium", and "low" as levels of recidivism risk potential. Further, the mental health status of the inmates is	Analysis for the identification of the mentally ill felony.	Statistical classification of criminals missing. Could have taken more features	"Referred" individuals can be made to have a longer stay in jail longer than "not-referred" individuals.

			Region Ten Community Services Board.		charges", "traffic violations", "driving while intoxicated",	categorized into two categories "referred." and "not-referred."			

5. QUANTITATIVE ANALYSIS OF CRIME ANALYSIS AND PREDICTION APPROACHES

For performing the quantitative analysis of the methods taken, the performance metric value needed to be computed and they are to be compared with the other. Hence, for performing the calculations of the performance metric there are a few formulas which can be utilized for achieving the performance value from the dataset. The formulae for the calculation of the performance metrics are given below in Table.2.

Table.2. Metrics and their formula

METRIC	FORMULA
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$
Error Rate	100 - Accuracy
Precision	$\frac{TP}{TP+FP}$
Recall	$\frac{TP}{TP+FN}$
F-value	$\frac{2}{\left(\frac{1}{P}+\frac{1}{R}\right)}$ where, P is the Precision and R is the Recall

Although many papers were studied in the literature review all the papers were irrevalent to the crime prediction and criminal analysis domain. Hence a few papers in the crime analysis and prediction domain has been taken and their results have been reproduced as given originally in the reference papers. The below given table Table 3 provides quantitative analysis of the three tools and the Decision Tre algorithm which is suppored with the Genetic Algorithm for optimization of the parameters. When the parameters are optimized, the classification accuracy of the Decision Tree is increased a bit further. This shows that although the Decision tree performs well, when it used with the Genetic Algorithm for optimization of the decision tree paramters, the results shown show significant improvement in the accuracy and further more the tools given below have the metric value, which is purely based on the dataset and the records and the performance values are taken as it is in the reference paper. The quantitative analysis produced results which show the increase in the accuracy level of classification because of using the GA to optimize the parameters. This occurs because of the ability of the GA to learn the optimal values and then it is applied to set the parameter to optimal value when performing calculation. Also, the Precision, Recall and F-value varies from the dataset and the system. This shows the SIIMCO performing well when defined in terms of the metrics.

Table.3. Quantitative Analysis of Crime Analysis & Prediction

	NAME OF THE METHOD/ SYSTEM	PERFORMANCE METRIC	PERFORMANCE VALUE	
[4]	Decision Tree classification with GA for optimizing the the parameters	Accuracy of Prediction	Optimized parameter	91.64%
			Non-Optimized parameter	85.74%
		Clasification Error	Optimized parameter	8.36%
			Non-Optimized parameter	13.26%
		Fitness Function	Optimized parameter	72.28%
			Non-Optimized parameter	72.48%
[9]	SIIMCO	Recall	0.62	
		Precision	0.56	
		F-Value	0.59	
	CrimeNet Explorer	Recall	0.36	
		Precision	0.41	
		F-value	0.38	
	Log Analysis	Recall	0.53	
		Precision	0.51	
		F-value	0.52	

6. CONCLUSION

In this paper, we have studied some known approaches for crime analysis and prediction concerned with data mining. Although many papers have been studied, only those papers with background in the crme prediction and criminal identification papers are compared with a theoretical study. Each paper has their own advantages and disadvantages. Each paper has its own individual approach for solving the crimes and criminal prediction. This is a theoretical study for several methods in identification of crime and criminals which includes Text/ NLP based methods, crime patterns and crime evidence based methods, spatial and geo location based methods, communication based methods and finally Prisoner based methods. The data mining techniques studied from this survey can be applied for identifying the criminals in the society and also for providing a better future to live in.

REFERENCES

[1] Mugdha Sharma, "Z-Crime: A Data Mining Tool for the Detection of Suspicious Criminal Activities based on the Decision Tree", *International Conference on Data Mining and Intelligent Computing*, pp. 1-6, 2014.

[2] Kaumalee Bogahawatte and Shalinda Adikari, "Intelligent Criminal Identification System", *Proceedings of 8th IEEE International Conference on Computer Science and Education*, pp. 633-638, 2013.

[3] Jyoti Agarwal, Renuka Nagpal and Rajni Sehgal, "Crime Analysis using K-Means Clustering", *International Journal of Computer Applications*, Vol. 83, No. 4, pp. 1-4, 2013.

[4] Rasoul Kiani, Siamak Mahdavi and Amin Keshavarzi, "Analysis and Prediction of Crimes by Clustering and Classification", *International Journal of Advanced Research in Artificial Intelligence*, Vol. 4, No. 8, pp. 11-17, 2015.

[5] Shiju Sathyadevan, M.S. Devan and S. Surya Gangadharan, "Crime Analysis and Prediction using Data Mining", *Proceedings of IEEE 1st International Conference on Networks and Soft Computing*, pp. 406-412, 2014.

[6] Yu-Yueh Huang, Cheng-Te Li and Shyh-Kang Jeng, "Mining Location-based Social Networks for Criminal Activity Prediction", *Proceedings of 24th IEEE International Conference on Wireless and Optical Communication*, pp. 185-190, 2015.

[7] Jeroen S. De Bruin, Tim K. Cocx, Walter A. Kosters, Jeroen F. J. Laros and Joost N. Kok, "Data Mining Approaches to Criminal Career Analysis", *Proceedings of 6th IEEE International Conference on Data Mining*, pp. 1-7, 2006.

[8] Ehab Hamdy, Ammar Adl, Aboul Ella Hassanien, Osman Hegazy and Tai-Hoon Kim, "Criminal Act Detection and Identification Model", *Proceedings of 7th International Conference on Advanced Communication and Networking*, pp. 79-83, 2015.

[9] Kamal Taha and Paul D. Yoo, "SIIMCO: A Forensic Investigation Tool for Identifying the Influential Members of a Criminal Organization", *IEEE Transactions on Information Forensics and Security*, Vol. 11, No. 4, pp. 811-822, 2016.

[10] Kevin Sheehy et al., "Evidence-based Analysis of Mentally 111 Individuals in the Criminal Justice System", *Proceedings of IEEE Systems and Information Engineering Design Symposium*, pp. 250-254, 2016.

[11] David J. Hand, Heikki Mannila and Padhraic Smyth, "*Principles of Data Mining*", MIT Press, 2001.

[12] 10 Emerging Technologies That Will Change Your World, Available at: http://www.rle.mit.edu/thz/documents/10_emerging_tech.pdf.

[13] U. Fayyad, G. Piatetsky-Shapiro and P. Smyth, "The KDD Process for Extracting Useful Knowledge from Volumes of Data", *Communications of the ACM*, Vol. 39, No. 11, pp. 27-34, 1996.

[14] Illhoi Yoo, Patricia Alafaireet, Miroslav Marinov, Keila Pena-Hernandez, Rajitha Gopidi, Jia-Fu Chang and Lei Hua, "Data Mining in Healthcare and Biomedicine: A Survey of the Literature", *Journal of Medical Systems*, Vol. 36, No. 4, pp. 2431-2448, 2011.

[15] Shyam Varan Nath, "Crime Pattern Detection using Data Mining", *Proceedings of IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology Workshops*, pp. 1-4, 2006.

[16] Hsinchun Chen, Wingyan Chung, Yi Qin, Michael Chau, Jennifer Jie Xu, Gang Wang, Rong Zheng and Homa Atabakhsh, "Crime Data Mining: An Overview and Case Studies", *Proceedings National Conference on Digital Government Research*, pp. 1-5, 2003.

[17] Tong Wang et al., "Learning to Detect Patterns of Crime", *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 515-530, 2013.

[18] Karl F. Schuessler and Donald R. Cressey, "Personality Characteristics of Criminals", *American Journal of Sociology*, Vol. 55, No. 5, pp. 476-484, 1950.

[19] H. Chen, W. Chung, J.J. Xu, G. Wang, Y. Qin and M. Chau, "Crime Data Mining: a General Framework and Some Examples", *Computer*, Vol. 37, No. 4, pp. 50-56, 2004.

[20] Chung-Hsien Yu, Max W. Ward, Melissa Morabito and Wei Ding, "Crime Forecasting using Data Mining Techniques", *Proceedings of 11th IEEE International Conference on Data Mining Workshops*, pp. 779-786, 2011.

[21] P. Thongtae and S. Srisuk, "An Analysis of Data Mining Applications in Crime Domain", *Proceedings of IEEE 8th International Conference on Computer and Information Technology Workshops*, pp. 122-126, 2008.

[22] Rizwan Iqbal, Masrah Azrifah Azmi Murad, Aida Mustapha, Payam Hassany Shariat Panahy and Nasim Khanahmadliravi, "An Experimental Study of Classification Algorithms for Crime Prediction", *Indian Journal of Science and Technology*, Vol. 6, No. 3, pp. 4219-4225, 2013.

[23] Shyam Varan Nath, "Crime Data Mining", *Proceedings of Advances and Innovations in Systems, Computing Sciences and Software Engineering*, pp. 405-409, 2007.

[24] Donald E.Brown, "The Regional Crime Analysis Program (RECAP): A Framework for Mining Data to Catch Criminals", *Proceedings of IEEE International Conference on Systems, Man, and Cybernetics*, pp. 2848-2853, 1998.

[25] Colleen McCue, "*Data Mining and Predictive Analysis: Intelligence Gathering and Crime Analysis*", Butterworth-Heinemann, 2014.

[26] Arunima S. Kumar and Raju K. Gopal, "Data Mining based Crime Investigation Systems: Taxonomy and Relevance", *Proceedings of Global Conference on IEEE Communication Technologies*, pp. 850-853, 2015.

[27] Manish Gupta, B. Chandra and M.P. Gupta, "Crime Data Mining for Indian Police Information System", *Journal of Crime*, Vol. 2, No. 6, pp. 43-54, 2006.