CRIME ANALYSIS AND PREDICTION USING DATA MINING

TANAY BAGAYATKAR 2019230065

SHIVAM PAWAR 2019230068

VISHAL SALVI 2019230069

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 .
- 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 .
- 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.

- 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.

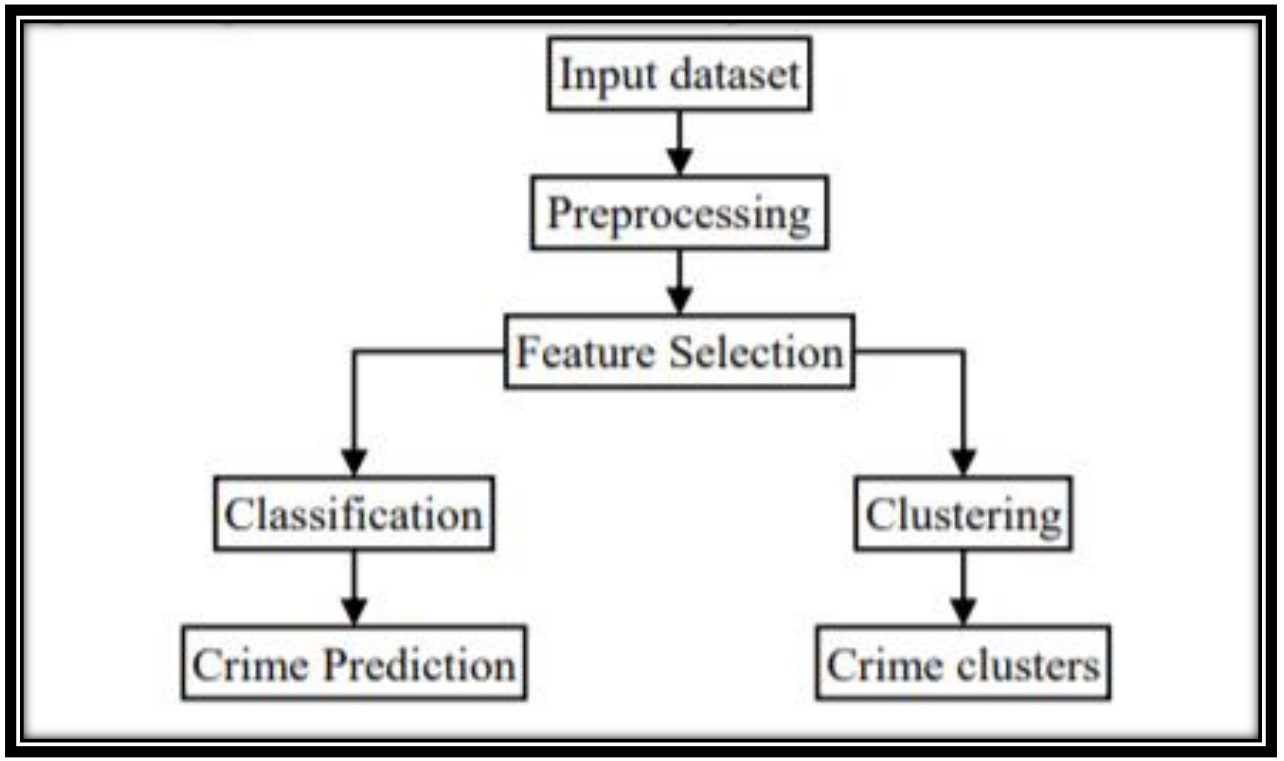
- **BIG DATA ANALYSIS:**

1. Big data analytics is the process to examine the huge amount of data to find hidden information patterns and trends.
2. 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.
3. An analytical system can cope up with predicting crimes. Investigation analytics system can deal with many information like text data, audio, video, DNA.
4. Combining with security intelligence sources, it provides the information about the latest vulnerabilities and identify outliers and anomalies in security data.
5. 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.
6. 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.

CRIME ANALYSIS PROCEDURE

1. The criminals when leaving the crime scene does leave some traces which can be used as a clue to identify the criminals.
2. The crime sequence and the patterns which several criminals follow when committing a crime make it easy for analyzing the crime.
3. 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.
4. 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.
5. 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 above figure 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 preprocessor 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.
- 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.

CRIMINAL ANALYSIS METHODS

METHOD		INPUT	DATASET USED	PRE PROCESSING	FEATURE EXTRACTION	CLASSIFICATION/ CLUSTERING	STRENGTH	WEAKNESS	OUTCOME
Text/ NLP-based methods		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
		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		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"
		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
		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
		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"				
		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		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, Neighbour 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		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		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 Region Ten Community Services Board.	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 charges", "traffic violations", "driving while intoxicated",	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 categorized into two categories "referred." and "not-referred."	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.

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.

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.

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.

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

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.

Crime Analysis Techniques

- **Classification**
- Classification is a supervised machine learning technique that categorizes data into groups (classes) and is used for making the future predictions. In other words, classification is used to predict the classes of a given data based on certain attributes known as predictors [7].
- A survey has been made on the detection and analysis of different crime detection mechanism using machine learning [6]. Sathyadevan et al. [8] have applied classification methods on location, date, and type of crime data extracted from Websites, Blogs, Social Media, RSS feeds to predict areas of high probability for crime occurrence.
- **They have used the following classification methods:**

1. Naive Bayes classification method to create a model by training crime data related to vandalism, murder, robbery, burglary, sex abuse, gang rape, arson, armed robbery, highway robbery, snatching, etc.
2. Apriori algorithm to identify crime patterns that occur frequently.
3. Decision Tree to predict the crimes areas. The techniques used are easy to apply and interpret, but they can only predict the crime spot without predicting the crime occurrence time.

- K. Bogahawatte and S. Adikari [9] developed a system called Intelligent Crime Investigation System (ICIS), which can predict a possible criminal based on the evidence and observations collected from the crime scene.
- Colombo crimes and criminal database records is trained using Naive Bayes classification algorithm to distinguish potential suspects.

- ✓ Decision tree that uses an improved feature A. Chandrasekar carried out an attempt to predict and classify the crimes in San Francesco city. The dataset of crimes that took place in San Francisco city over twelve years.
- ✓ First, they preprocessed the data by removing some features, replacing some indexed features with a number, and decomposed timestamp feature into five features.
- ✓ The dataset is enriched the feature by adding features from United States Census data. Finally, they performed multiple classification techniques in order to classify the crimes into 39 categories (original dataset classes).
- ✓ The study results in high testing error, which indicated a high variance, so they merge their dataset classes into two classes and re-applied the same classification techniques i.e. Naive Bayes, Random Forests, Support Vector Machines, and Gradient Boosted Decision Trees.
- ✓ High accuracy and precision were achieved using Gradient Boosted Trees and Support Vector Machines [10].

Clustering

- Clustering is an unsupervised data mining technique used to split a group of items and data into clusters based on certain characteristics, each cluster contains a group of similar data.
- Clustering is used in undefined and unfixed classes and without supervision when it comes to grouping the objects [7]. R. et al. [1] have used diverse clustering methods to analyze crimes in the dataset State Crime Records Bureau (SCRB) of Tamilnadu, India.
- The data has been obtained which has approximately 38 several cities and districts crime data. The clustering method that has been used for the analysis was DBSCAN and K-means. To resolve the efficiency of the two clustering approaches, they used Silhouette coefficient. The study concluded DBSCAN clustering with a high accuracy and resulted in more accurate clusters.
- P. Vrushali et al. [16] focused on the frequency rate during diverse years to classify clustered crimes and the study was conducted on real data which offers an innovative framework for clustering and predicting crimes. To analyze the crime data, in addition, to categorize crimes by grouping the related patterns, several clustering methods were applied.
- In conclusion, several crime categorization and prediction methods were discussed that were applied to different crimes datasets. All the discussed papers have applied more than one machine learning technique to achieve the expected results. Therefore, we will need to apply many techniques to our dataset to analyse the data behavior, in order to create an efficient module to predict the crime category with the optimal accuracy.

Description of the Proposed Techniques

This section presents the description of the machine learning techniques that is used in our study to predict the crime category based on behavioral data. In our study we used supervised learning approach and chose Naïve Bayes and Decision Tree algorithms, which fall under the classification technique.

Naïve Bayes Technique

Naïve Bayes is a simple supervised classification algorithm. It uses the concept of Bayes Theorem, which classifies tuples to a class label related to a dataset based on the calculations of the conditional probability for each class label and attribute. The Naïve Bayes algorithm works by the given probability rule.

$$P(H \mid X) = \frac{P(X \mid H)\, P(H)}{P(X)}$$

Where X is a data sample of dimensions x = < $x1$, $x2$, $x3$ … . xn > with an unknown class label, and X belongs to H hypothesis class label. Where $P(H)$ is the prior probability (initial probability) associated with hypothesis H [17].

Decision Tree Technique (J48)

- Decision Tree is a supervised learning technique, which recursively partitions the instance space. Its main use is to predict the class labels.
- The Decision Tree is composed of internal nodes, which represents set of predictors (attributes), edges, which represent a specific value or range of values of the input predictors (attributes), and leaf nodes, which represent the class labels.
- The internal nodes along with their edges split the instance space into two or more partitions and each terminal node (leaf node) of the tree is a class label [18].
- J48 classifier is a simple C4.5 Decision Tree for classification. In this technique, a binary tree is constructed to model the classification process.
- Once the tree is built, it is applied to each tuple in the database and results in a classification for that tuple. J48 splits the data into range based on the attribute (predictors) values.
- J48 allows classification by Decision Tree or the rules generated from them [17].

The Basic Steps in J48 classifier:

1. For the instances of the same class, the tree represents a leaf labeled with the class.
2. By applying a test on each attribute, the potential information is calculated for all of them.
3. Then, a test is applying on the attribute to calculate the information gain.
4. Then, based on the selection criterion, the best attribute is selected and used for branching.

The “Entropy” is used in this process, which is a measure of the data disorder. The Entropy of $\vec{y}$ is calculated by

$$Entropy\ (\vec{y}) = -\sum_{j=1}^{n} \frac{|y_i|}{|\vec{y}|}\ log\ (\frac{|y_i|}{|\vec{y}|})$$

$$Entropy\ (j|\vec{y}) = \frac{|y_i|}{|\vec{y}|}\ log\ (\frac{|y_j|}{|\vec{y}|})$$

And Gain is

$$Gain\ (\vec{y}, j) = Entropy\ (\vec{y} - Entropy\ (j|\vec{y}))$$

The objective is to maximize the Gain, dividing by overall entropy due to the split argument by value j.

Comparison of papers

References