|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Method** | **Input** | **Dataset Used** | **Pre-Processing** | **Feature Extraction** | **Classification Clustering** | **Strength** | **Weakness** | **Outcome** |
| **Text/NLP-Based Methods** | E-mail message | Real and open emails sent by terrorists and some dummy emails | **Nil** | Selection of a subset of the original of 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  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 | Device sensors, Security  camera information, Messages, Audio feeds, Social network posts and messages | Structuring collective data into {Time, Final Movement, Frequency rate, Video, Images, Audio} | Similarity matching for sensory images using sliding window. Text semantic Analysis of the text information performed using Lexical processing, Natural Language Processing (NLP). | A trained classification model is used to predict the similarity of a given input to the suspicious item or location | Consideration of location feeds and mobile usage information | Not giving a clear view of the processing and comparison of criminal behavior | Suspicious behavior to three levels such as “High”, “Medium” and “Low” |
| Crime patterns and Evidenc e-based method | Crime evidences including many attributes like crime scene, day, month, offense, resources used, time, role in crime, transportatio n 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 patterns and Evidence-based method** | 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 nonoptimized 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” “highway robbery”, “snatching” | 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 crime |
| 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 |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Crime patterns and Evidence-based method** | 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” “highway robbery”, “snatching” | 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 crime |
| 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 Geolocation based method** | Geolocation 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 OpenStreetMa p in San Francisco Bay area and Criminal pattern discovery according to the context of user activity and locationbased social networks. Predict crime frequency and find which crime is to be more difficult or easier to be predicted |
| **Commu nication based methods** | Flow of communicat ions/informa tion links between two criminals (e.g., phone call records, messages, etc.), names of criminals/su spects, the type of crime, location and date of the crime | Real-world communicat ion 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 lowerlevel 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 experimentall y 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  Charlottesvil le 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. |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Communication based methods** | Flow of communicat ions/informa tion links between two criminals (e.g., phone call records, messages, etc.), names of criminals/su spects, the type of crime, location and date of the crime | Real-world communicat ion 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 experiment all y with Crime Net Explorer and Log Analysis | Nil | System can identify the influential members of a criminal organization and the immediate leaders of a given list of lower-level criminals |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Spatial and Geolocation based method** | Geolocation and Crime Type | SNAP Gowalla dataset, Data SF 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 OpenStreetMa p 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 |