A COMPARATIVE STUDY ON PREDICTION OF CRIME PATTERNS

*R.Sujatha, SITE, **Dr.Ezhilmaran, SAS
VIT University, Vellore - 632 014.

Email: *r.sujatha@vit.ac.in

Received on 10-11-2016
Accepted on: 28-11-2016

Abstract:

In the daily life, crime keeps increasing and threatens the lives of the people in public. The accuracy and time of tracing are robust while data mining technique is indulged. The hurdle in the process starts from selecting the related variable for analysis and their sensitiveness. The research work carried on the crime is potential area that requires optimization. The quantum of the data and volatility makes the field challenging. In this paper summarized the various existing techniques.

Keywords: Crime pattern, framework, forecast, clustering, association rule mining and visualization

1. Introduction

Data is the vital thing that decides about the decision of various aspects in all fields from financial, medical, marketing, demographic and scientific, the list goes on. The way of analysing the data requires tremendous amount of time and effort. The retrieval of the informative knowledge is quite interesting, from the huge data and it is also a challenging task. Data mining is the powerful technology to analyse the data skillfully from different perspectives and summarizes it to useful information. The finding of the cold spot and hot spot is the challenging task in crime analysis. Crime is the act that harms the public, increases the violence, demolishes the assets and denies the respect to people. Distribution of Crime is not even across the globe. The huge available data and hands-on experience helps for investigation. But this data has arisen in a drastic manner which paves the way for the data mining to take care of investigation [1].

2.1 Crime Analysis Techniques

2.1.1 Statistical Approach

The imperfect information from an extensive range of data sources and a complex structure is the barrier for the crime exploration and prevention. The warehouse of crime profiles of recent and historical along with the socio-economic factors like population, poverty, unemployment are taken into consideration along with the multiple...
statistical analysis like correlation and regression, make the process soother. The work reveals the relationship of the crime pattern and to build a process administration tactic [2].

Police started to forecast the crime on the basis of criminality and modus operandi of the sequence criminals. The reason for not conceivable crime forecasting is due to the fact that the preferred scale for inspection is miniscule for the estimation of dependable model. The study conducted in jurisdictions of United Kingdom and United States use univariate and multivariate methods catering to widespread horizons. The forecast models are designed depending on the distinctive attributes of the problem domain and data. Police must identify the crime in regions as small as feasible for tactical purposes, at the patrol district level or smaller [3].

The two main arguments are is it possible to forecast accurately the selected crime in advance, in a small geographical area. On the accuracy level comparison between the two approaches namely model-based forecasting and present practices of police are made. The univariate method along with multivariate method helps in the short term crime forecasting. The hotspot and the criminality tracing of certain places can be attained with the help of forecasting process. Explore on cause and result must deal with the essential recognition problem that happens when trying to forecast outcomes. Investigating some of the pragmatic matters involve in anticipating city level crime rates by means of a general panel dataset in this light [4].

Harries worked on modelling along with forecasting logged assets crimes in the area of Welsh (England and Wales) is the provoking effort that weakened in the domestic break-in. The quantity of recorded offences of housebreaking is termed as residential burglary. The young men proportion, economy level and trend level are the key factors that have impact on the model. In particular, spotlight on the problem of forecasting future crime rates from examined data, not the problem of forecasting how dissimilar policy levers impact crime. Certainty, proportionality and celerity should be exhibited by criminal justice system to effectively prevent individuals from making offence [4] [5].

The task on huge crime datasets to extract the co-distribution patterns was carried out by Phillips and Lee. For the task of crime mining they utilized combination of socio-economic and socio-demographic issues. For the experimental purpose MATLAB program is used to generate synthetic data using the multivariate Gaussians. Based on the base map the conversion to real date is carried out. Then ten datasets are produced and the average of output is considered. In detail the framework starts with normalizing using min-max technique then weighted, directed multi-graph for each dataset is created. Using global minimum edge the graphs are pruned to extract influential edges. The correlation analysis is done to find similar co-distribution. To find correlation, person’s correlation coefficient or Jaccard’s index is involved [6].
The model permits agent-based representations of offense based on the geographical location and also aids in finding the individual victims characterisation. The population Reconstruction Model established with the idea of utilizing the mixture comprising the census of small zone and the sample of anonymised data to afford the synthetic list pertaining to the city or region population in the country [7].

2.1.2 Expert Knowledge Approach

The DNA and fingerprint from the crime scenes are used in the investigation model developed by incorporating the data mining techniques. Crime scene investigators are classified into three levels based on the skill set of training they underwent and the capability to examine the crime scenes. The activity of investigators for each crime scene is recorded with the information about time, day, date happened along with the complete and also forensic samples and results. The cross industry standard procedure for data mining is utilized in this investigation model. Data is prepared and let into the modelling with the domain expert knowledge to bring out the effective results of the investigation. The expert classifies the collected data of fingerprint from the crime scene into insufficient, eliminated, matched and outstanding. Even for the DNA the classification is done. Along with this two flag fields created helps in the matching purpose [8].

2.1.3 Collective Data Mining Techniques

Profiling the offender information then applying the data mining paves way to understand the behavioural pattern of the criminals. The rules mined from the profiles provide the preview of the relationship. The association rule mining is considered as the successful technique. Concept hierarchies are constructed which in turn generates rules. It is of significance, that the incorporation of dissimilar statistics, such as Pearson’s sample coefficient correlation and multiple correlations, and an incomplete correlation coefficient, improves viable benefits in the view of obscured circumstances and final support [9].

Anticipating has been a substantial part of the criminal justice practice from the early 20th century. Providing the decision support system for policing with the help of data mining is achieved using the soft forensic proof like modus operandi, temporal and geographical crime features. The key things of the system includes forensic computing, extracting practical lessons relating to the purpose of computer science and to set of conclusions that includes need for multidisciplinary contribution to guide. The use of various data mining technology utilized in the investigation of crime like geographical information systems displays, link analysis algorithms, clustering, and association along with more complex process as per requirements. Contemplation is focused on the intelligence based on conceptual knowledge and reasonable logics of different kinds of statistical models related with the calculation of a vast quantity
of data that mimic illegal incident and their correlations [10]. Data mining technique classification and clustering applied over Felonious dataset to determine the hotspot and to calculate the crimes and felons. The classification is done in three ways based on crime place, types and time to produce crime hotspot or crime cold spot depending on the incidents. The clustering makes use of the weighted attributes along with threshold to generate the cluster [11].
Tayal proposed the crime detection and criminal identification using data mining perspective for the cities of India. The work is carried on in six phases beginning from data extraction, pre-processing, clustering, representation using Google map followed by classification and implementation in WEKA. From various sources data is retrieved for 12 years during 2000-2012. The KNN is used for the classification purpose to identify the criminal. K-means clustering is used and GMAPI is utilized for the visualization purpose. Comparison of the clustering results is done with the WEKA and prediction rate of around 93 % is the output [13].
Coplink – intelligent analysis and knowledge management was designed based on the concept analysis. With the help of this tool intelligence is derived from data at the faster rate and helps in investigation [13]. With the experience from the COPLINK project Chen proposed the Crime data mining framework is the great boon for the law enforcement agencies to analyse the voluminous crime data. The local and federal authority devised the eight categories of crime type based on the increasing harm to the public. The crime types are traffic violation, sex crime, theft, fraud, arson, drug offences, violent crime and cybercrime. The various data mining techniques are entity extraction, clustering techniques, classifications, association rule mining, deviation detection, sequential pattern mining, string comparator and social network analysis. The crime data mining frameworks produces the result between the crime types and data mining methods like classification, clustering, association and trend visualization clearly depicts the intensity of the crime types. Based on the investigators the usage of single or combined data mining techniques can be made [14].
The researcher discussed analysis of crime domain in the light of data mining. The crime control and criminal suppression are considered as two significant rudiments by the law enforcement in the data analysis. Former uses information to control and later uses history stored in mining to seize a scandalous person. Narrated about the various works carried out and allows the discussion on the finalizing of the attributes and assessment requires quality work. Machine learning methods make the greater impact in the effective and efficient analysis [15].
The data is gathered from various city of United states and data pre-processing is performed to ensure the better performance in the classification purpose. The work is carried out to classify the crime area to find hotspot area. The preparation of datasets from crime records. The prepared datasets comprise of various types of crimes and related
incident as characterized by the police department of the city. Spatial and temporal information is embedded into the prepared datasets. Various classification techniques are utilized to analyze the hotspot area. The cumulative work provides the pattern of crime and also information based on space and time. The data grid is made and empty grids are removed followed by the feature construction. The feature construction depends on leveraging temporal knowledge and maximizing spatial knowledge. The work compares the performance on the various classification processes [16].

Ding and his colleagues designed the PerpSearch that is based on the Law Enforcement Tactical System. The said system holds various cross-jurisdictional records about arrest, sentencing, vehicle registration etc. The PerpSearch process flows with the four component Geographic Profiling, Social Network Analysis, Physical Match and Crime Pattern along with the Score Engine. Based on the values of Score Engine the output is retrieved for the provided input. The planning is to include the historical result that will optimize the system and also enriches the investigation work [17].

Le Khac worked on the money laundering as a part of the modern crime. Architecture was built for the anti-money laundering to track the suspicious cases. The framework constitutes the data pre-processing followed by data mining process like classification and clustering that provides the knowledge. The sorting is performed to categorize the customers into pre-defined types of risk. The centre based clustering is used to group like transaction into groups and generate the suspicious profiles. Continuous process strengthens the profile and helps to tackle the money laundering in the investment bank [18].

The work is the combination of the data mining technique with spatial clustering for visualizing the crime over the Washington Area. Followed by classification models to classify based on closest training examples in feature space. Then association rules with minimum support and confidence. They are able to provide information about safety place with the help of statistical methods with data mining essence [19].

Bruin proposed the tool that works on the national crime records database to construct the criminal profile. Criminal career analyzer is a multiphase procedure that works on the national crime record database. The database is so reliable and up-to-date in nature. The process begins with the extraction of features like crime nature, frequency, seriousness and duration. The criminal crime profile is created from the extracted data based on per offender. Comparison procedure is performed on the overall collected set. Distance is calculated based on the profile, crime severity and crimes count. Cumulated over the time and human centered multi-dimensional clustering methods is utilized for
clustering purpose. The system is strengthened by normalizing the careers based on time so that comparison will yield better result [20].

2.1.4 Clustering Approach

In the very early stage itself Brown suggested the method called ReCAP, Regional Crime analysis program is to deliver crime analyst. It is integrated system that makes searching with respect to time, geography and location. According to him data fusion and data mining provides the useful information. Where former is used to combine the data from multiple sources and later is used to discover the pattern and also meant for automation to get the relationship among various attributes that have impact on the crime. The work is carried out on the spatial data mining. K-means and the nearest neighbourhood clustering are performed [21].

On the view of tactical operation support for the law enforcement agencies on the weekly basis multivariate prediction model for hotspots was introduced. For prediction based on space-time event a point-pattern based transition density model was used and it depends on observed past offence preference. The components of the model include Geo-Space Feature density, Spatial Transition density and temporal transition density. The team worked for the time scale of the Richmond, VA area considering the counts of criminal activities, demographic features, consumer expenditure features and distance features. The outcome shows via the probability density estimates over time and space [22].

The model used the tactical police resources to identify the high crime spots. Predicting the crime future is the process of crime analysis and along with space and time in combination of people’s behaviours is termed as spatial choice analysis. The approach depends on the clustered results of the happened offence activities with respect to time and location. Decision is derived with the two proposed models, space adjusted and key featured adjusted model. They employed spatial analysis that performs point density analysis, cluster analysis, cluster allocation and bounding. Sequential cluster modelling is performed consisting of gamma test and model implementation for crime prediction purpose [23].

The crime prediction model considers the real time data about crime information along with the police strength of both civil and armed. Then care is taken to get rid of missing data on careful analysis. The required algorithms are used in combination to ensure the quality data is used for further clustering process. Hybrid of K-means with DBScan (Density-Based Spatial Clustering Application with Noise) is used to cluster into four clusters stating crime is steady; it is rising; generally increasing and always is in flux. For knowledge discovery semi-supervised learning was utilized and this will boost the analytical precision. The ultimate aim of the work is to set the tool development for Indian
The city level crime data is worked with conventional k-means and proposed the weighted k-means along with Radial Basis Function classification in both the methods. The accuracy is higher when the fine tuning done with the help of weights. The weighted k-means works on the basis logical difference between the clusters inspected for the variations and between smallest and largest weight. In iterations changes in the smallest and largest weight occurs and finally cluster in the acceptable format is found. The accuracy is higher in the proposed technique of the weighted K-means along with the Radial Basis Function [26].

The behaviour of violent criminals is taken into consideration in simulation model along with the data mining techniques. The three types of criminals are classified as Violent Psychopaths, Persons with Antisocial Personality disorder and Persons with Intermittent Explosive Disorder. The pertinent challenges are increase in crime data, inconsistent in data, mapping the criminal behaviour to the crime. In this simulation model clustering, classification and outlier detection is applied. The process of cleaning and retaining the required attributes are performed. Clustering performs grouping followed by tracing common properties in Classification and analysis done based on specificity. The properties used in the model are based on the criminal behaviour literature. The probabilistic and timing factors can be inserted to get informative output [27].

The model proposed by Shyam worked on the basis of semi supervised technique. The k-means clustering is utilized to group the available crime data base. The attributes selection in the available huge crime data base is challenging and that increases the uses of the clustering. The weighing concept in introduced based on different crime types. Dynamically the different weights are assigned to attributes and cluster is formed. Geo-spatial is merged with this to retrieve based on the location. The crime patterns will act as the resource for the detective agencies to take decisions. The profiling data base in interaction makes the matching crime with culprit is achieved. Crime reports comprises of type of crime, date/time, location, suspect, victim and witness etc. were considered along with k-Means Clustering. The crime clusters are plotted with legend for significant attributes for that crime pattern [28].

Gupta worked on the National crime record holding the data of all the states and union territories of India. The multivariate time series clustering with dynamic time wrapping method is applied on the districts crime dataset. The multivariate time series data and the weight matrix based on domain are taken into consideration. The distance matrix is computed using the parametric minkowski model. Dynamic programming technique used for wrapping followed
by dynamic time wrapping for each series is computed. To predict similar trends of crime hierarchical clustering is done. The comparison is done with Euclidean and Minkowski to obtain the crime trends that are similar [29].

2.1.5 Association Rule Mining and Frequent Itemset

Knowledge extraction is attainable with the frequent itemset mining. The developed system concentrates on analysing the frequent pattern and then clusters it. The processing of crime data is performed to generate crime rates, trends, hot spots etc., Apriori algorithm is utilized to generate the frequent itemset to retrieve pattern [30].

Wang proposed the model named series finder to detect crime patterns. Learning is accomplished from a seed of a few crimes by growing a pattern of discovered crimes from within a databank. The pattern discovery process was made on crime data collected by the crime analysis unit of Cambridge police department and revealed promising patterns. Supervised learning approach with Pattern Generic Coefficients and Pattern Specific Coefficients are used for crime pattern detection. Modus operandi of offender and the way of happening acts as the input and the size of the data helps to get the more realistic and informative pattern. The process goes in sequential manner with the candidate crimes and tracks the similar crimes. The process comes to the end when no more related things exist in the database. Thus the sequential pattern building yields several patterns [31].

Brown proposed the data association methods in the law enforcement. The work is based on the records association in related database on the other hand they are not precise matches. In the crime analysis manually based on the incidents tracking with similar will occur. The cumulative task is done to minimize the time incurring in the manual tracking system. Numbers of automated tools were developed to aid the activities pertaining to the law enforcement. The example of work is categorized into i) expert system, ii) investigative support system and iii) non-automated crime analysis methods. The Armed Eidetic Suspect Typing system is constructed holding the 13 suspect traits of witness and confidence level was assigned. Level of certainty that matches the suspect with crime is observed, then rules are mined.

Data association methods used for suspect and incidents, incidence ranking and done on real data for crime analysis. Many investigative support systems were designed and commercial products in market. Considering all this based on similarity approaches work was carried. Total similarity measure is computed and based on the similarity of trait between records. Traits can be quantitative or qualitative in the database. Then based on the expert similarity measure and total similarity measure work is performed. Then in the angle of optimization total similarity measure is performed along with the weights. On comparison with methods data association rules possess high accuracy [32].
2.1.6 Machine Learning Approach

The motivation of this developed system relies on the problem faced by the police in tackling voluminous crime. The lack of software based on the police need, paved way to the OVER project, that aimed at managing various records of the past to address the issues. The work started with the Microsoft Access and analyzed using Structured Query Language. Followed by developing system that afford the mapping and visualization tools for the existing data along with predicting perspective. The final part of the work focused on the crime profiling that help to solve the crimes based on victimization and possible offenders of the same. In order to match crimes with the criminal kohonen neural network is used and to predict the re-victimization, Bayesian belief network were used which provided the graphical overview [33]. Li team proposed the Fuzzy self-organizing map (FSOM) network. They worked on temporal crime activity data of the non-western real world. The objective of the work was to find the crime trend patterns and rule generation to reveal hidden causal –effect knowledge. The data from national police agency in Taiwan is obtained and data standardization and fuzzification is applied in pre-processing phase. The FSOM phase works on the processed data by setting the parameter for number of clustering, confidence and support for the rule extraction. The output reveals the crime trends, offences, locations and side around effects. The analysis provides the information about typical, gradual increase, sharp increase and winter time crime patterns. The statistical tests supports the above said crime patterns were true [34].

Alruily worked on the Arabic documents to categorize into type of crime. The theft, fraud, drug and alcohol smuggling, magic and sorcery, sex and violent crime are the various crimes taken into the consideration. The system works on the document. The process starts with normalization that helps in resolving the difference that exists for the same word. Followed by that, information extraction is performed which is combined with the stemming process. The stemming eliminates the affixes and works based on unique word concept along with that removes the stop word. The clustering and visualization is invoked to cluster the related documents with the help of Self Organising Map technique. The process begins with the initialization of the weight randomly and neighbourhood ratio. Until the convergence criterion is met, the input pattern is set, Euclidean distance is calculated, winner neuron is found and updating occurs [35].

The work is concentrated on deriving patterns based on different aspects in crime and aspects refer to the data sources that are versatile. A pertinent thing is to disclose the trends with the help of exploring the data at various stages based on granularity and also aspect criteria. Growing self-organising maps - a method of hierarchical clustering is deployed. The process involved is initialization phase, growing phase and smoothing phase. The further work based
on the concept of hierarchy is called granularity based multi-modal data, the model is constructed. The architecture made the identification on the basis of horizontal and vertical pattern identification with global concept hierarchy. The same provides the abstraction and granularity at different levels [36].

Iqbal proposed the work of classification technique on UCI Communities and Crime dataset. The work was carried on with twelve attributes. The prediction is done using the Decision tree and Naïve Bayesian algorithm with the help of WEKA tool. The model was built after removing the missing values attribute and then multiple linear regressions on the UCI Communities and Crime dataset. The work was also performed on with wine quality dataset, spam base dataset along with various techniques to bring out the performance. The prediction rate on the crime prediction over crime data is observed as 83% [37].

In the range of crime analysis Keyvanpour suggested the framework for detection and investigation purpose. Crime variable and matching are the two pillars according to the proposed method. Irrespective of the crime spatio-temporal crime variables, crime natural specification and offender profiles are taken into consideration for crime variables. The component in Crime matching holds, starting with the entity extracting, clustering and finally shows the matching. The entity extracting starts with feature extraction turn, created crime dictionary along with lexical lookup engine, the accuracy is guaranteed. Then SOM neural network clustering is performed to the selected entities where the features map is retrieved using the self-organizing neural network and k-means is applied to the same [38].

### 2.2 Crime Analysis for Visualization

Visualizing is the main area of interest after analysing the data. The data mining, data management, human computer interaction, perception and cognition act as the key elements in scientific and information visualization. The challenges in the visualization process are volume of data, dimensions of data, data quality, and graphical representation, level of detail, display devices, evaluation and infrastructure. The careful analysis and clear idea about the need of the end user is the vital thing to provide highly informative and interesting visualization [39].

The work was carried out for the Indian Police Department. The initial works that paved way to the proposed model are Crime Criminal Information System (CCIS) and Common Integrated Police Application (CIPA). The CCIS serves only the purpose of collecting the data and analysis is not carried out. The district level information is only done in the CCIS and to integrate all the districts the CIPA was done. The connection is established with National Crime Record Bureau (NCRB). In order to improve the entire system, that connected all the district and state was modelled. The feeding of the state, year with crime type gives the result of hotspot [40].
Bhat mentioned firmly as the information and communication field advance the committed crimes are also becoming
intensive technically. Digital devices are utilized for crimes and the necessity of usage of trending techniques become
important. Digital forensic investigation framework impacts the flow of information among the technical and non-
technical members of the team. The practical working encompasses of six stages beginning from preparation,
collection and preservation of digital device, Data extraction and pre-processing, data examination along with
analyses, reporting combined with documentation and finally presentation in the court of law. In order to get
visualization of the preserved data, simple K-means clustering and classification technique is implemented. (Bhat et
al., 2010).

Mapping or visualizing graphically the crime occurrence probability at any given time with the available police
START and END crime times. Aoristic analysis is combining temporal and spatial to ensure higher level of
knowledge grabbing. It calculates the probability of event occurred in limit of temporal parameters followed by
cumulating the probabilities to get the weight of the area. The nine categories taken into consideration are assaults,
break and entry in residential and non-residential, malicious damage, robbery, stealing, motor vehicle crime, street
offences and drugs. By noticing the routine activities of the crime, better understanding of the forthcoming crimes can
be tracked [41].

3. Conclusion and Future Work

The purpose of data mining is vital in all the area and from the above works it is evident that each researcher worked
on different perspective to ensure the safety for the nation perspective. The data utilized for the purpose is varied in
nature and it is important to create a common ground for the storing of the crime related data. The usage of the soft
computing algorithms will enhance the rate of finding the highly critical data in the shorter span of time. The data
mining techniques like classification, clustering and association rule mining is well serving the purpose of handing
the versatile data. The effective management of crime related data along with the optimized algorithms is the
potential area of research.

References:

2. Dale Dzemydiene, Vitalija Rudzkiene., Multiple Regression Analysis in Crime Pattern Warehouse for Decision


34. Sheng-Tun Li et al., An intelligent decision-support model using FSOM and rule extraction for crime prevention, Expert Systems with Applications 37, 2010, pp. 7108–7119.


