JCSE International Journal of Computer Sciences and Engineering Open Access

Research Paper Volume-5, Issue-8 E-ISSN: 2347-2693

Crime Data Mining an Indian Perspective

N. Narwal

Dept. of Computer Science, Maharaja Surajmal Institute, GGSIP University, New Delhi, India

**Corresponding Author: neetunarwal@gmail.com*

Available online at: www.ijcseonline.org

Received: 19/Jul/2017, Revised: 26/Jul/2017, Accepted: 20/Aug/2017, Published: 30/Aug/2017

Abstract— Today Information Technology is used in every domain of life. The traditional age-old system of intelligence and criminal record maintenance are no longer used in the current crime scenario. With the availability of spatial data related to crime can be used to integrate it with latest GPS system to highlight the location of crime and such indication can be used by people to be cautious and alert. This latest technology can be used as an aid to warn and guide people about such happening in the crime prone areas. Manual processes neither provide accurate, reliable and comprehensive data round the clock nor does it help in trend prediction and decision support. It also results in lower productivity and ineffective utilization of manpower. The solution to this ever-increasing problem lies in the effective use of Information Technology. This paper suggests methodology that combines the Geographical Information System with crime data to provide crime data mining.

Keywords—Crime Mining, Geographical Information System, Hot Spots.

I. Introduction

India is a vast country with more than one billion populations, and has a police force of 1.5 million. Police is a critical component of civil administration in India. In 1986, the Government of India created National Crime Record Bureau (NCRB). To give right impetus to the National Crime Record Bureau, State Crime Record Bureaux (SCRBx) at States and District Crime Record Bureaux (DCRBx) at Districts followed. There were neither a staffs nor time for entering data in records manually.

In order to make use of the information technology, Government of India approved the design, development and implementation of a 'Government to Government (G2G), model called the Crime Criminal Information System (CCIS). The CCIS was designed to create computerized storage, analysis and retrieval of crime criminal records. The Crime Criminal Information System today is in operation in all the States. In CCIS, the information is collected at district level not at basic unit of police administration i.e. the police station. Common Integrated Police Application (CIPA) was developed with objective of automation the processes (workflow) at police station and to provide inputs for building CCIS. Till date, CCIS is only collecting the information and creating a huge crime database and there is no analytical tool for analyzing huge building database. Absence of crime analysis tool made it somewhat 'standalone' system. Analysis of crime data is very challenging task, as lots of work needs to be done to mine the huge amount of data is available.

The data available from CCIS systems from different location need to be integrated and transformed according to the need of application. This data can be combined with geographic spatial data to produce valuable crime maps, so that special patterns can be formulated and analyzed by criminologist.

In order to analyze the crime data, two disciplines needs to be combined Data Mining and Geographical Information System. Data mining refers to extracting or mining knowledge from large amount of data. Many people treat data mining as a synonym for another popular term knowledge discovery in Databases (KDD), other view data mining as simply as essential step in the process of knowledge discovery in databases.

Knowledge discovery consist of an iterative sequence of the following steps:

1. *Data Cleaning* –This step remove noise and inconsistent data.
2. *Data Integration* –In this step multiple data sources may be combined.
3. *Data Selection* – Data relevant to analysis task are retrieved from the database
4. *Data Transformation* – In this step data are transformed or consolidated into forms appropriate for mining by performing summing or aggregation operation.

5. *Data mining* – This is an essential process where intelligent methods are applied in order to extract data patterns
6. *Pattern Evaluation* – This step identify the truly interesting pattern representing knowledge based on some interesting measures.
7. *Knowledge presentation* - In this step visualization and knowledge representation technique are used to present the mined knowledge to the user.

Data mining functionalities are used to specify the kind of patterns to be found in data mining tasks. Data mining task can be classified into two categories, descriptive and predictive:

Descriptive mining task characterize the general properties of the data in the database.

Predictive mining task performs inference on current data in order to make predictions.

Some of data mining functionalities are:

Associative Analysis

Associative analysis is the discovery of association rules showing attribute value conditions that occurs frequently together in a given set of data. Association rules are widely used for market basket or transaction data analysis.

Association rules are of the forms X=> Y that is A1 ^ A_2 ….^ Am-> B1 ^ B_2 ^ ... ^ Bn where Ai (for iЄ {1,..m}) and Bj (for j Є {1,..,n}) are attribute value pairs.

The association rule X=>Y is interpreted as database tuples that satisfy the condition in X are also likely to satisfy the conditions in Y.

Classification and Prediction

Classification is the process of finding a set of models (or functions) that describe and distinguish data classes or concepts for the purpose of being able to use the model to predict the class of objects whose class label is unknown. The derived model is based on the analysis of a set of training data. The derived model may be represented in various forms, such as classification rules, decision trees, mathematical formulae or neural networks. The case when the predicted value is numerical data and is referred as prediction.

Cluster Analysis

Cluster Analyzes data objects without consulting a known class label. The objects are clustered or grouped based on the principle of maximizing the intraclass similarity and minimizing the interclass similarity. Cluster can also facilitate taxonomy formation that is the organization of observations into hierarchy of classes that group similar events together.

Outlier Analysis

A database may contain data objects that do not comply with the general behavior or model of the data. These data objects are outliers. Most data mining methods discard outlier as noise or exception. In some application such as fraud detection, the rare event can be more interesting that the more interesting than the more regularly occurring ones. The analysis of outlier data is referred to as outlier mining.

Evolution Analysis

Data Evolution analysis describes and models regularities or terms for objects whose behavior changes over time. This may include characterization, discrimination, association, classification or clustering of time related data, distinct features of such an analysis include time-series data analysis, sequence or periodicity pattern matching and similarity based data analysis.

A GIS is an information system designed to work with data referenced by spatial / geographical coordinates. Burrough in 1986 defined GIS as, "Set of tools for collecting, storing, retrieving at will, transforming and displaying spatial data from the real world for a particular set of purposes". Arnoff in 1989 defines GIS as, "a computer based system that provides four sets of capabilities to handle geo-referenced data: data input, data management (data storage and retrieval), manipulation and analysis, data output.

In other words, GIS is both a database system with specific capabilities for spatially referenced data as well as a set of operations for working with the data. It may also be considered as a higher order map.

GIS technology integrates common database operations such as query and statistical analysis with the unique visualization and geographic analysis benefits offered by maps. These abilities distinguish GIS from other information systems and make it valuable to a wide range of public and private enterprises for explaining events, predicting outcomes, and planning strategies (ESRI). It is used to digitally reproduce and analyze the feature present on the importance of a system which can represent the given data geographically. Earth surface and the events that take place on it. In the light of the fact that almost 70% of the data has geographical reference.GIS is looked upon as a tool to assist in decision-making and management of attributes that needs to be analyzed spatially.

Data mining is a young discipline with wide and diverse applications, there is still a non trivial gap between general principles of data mining and domain specific, effective data mining tools for particular applications.

Various application domains are Data mining for Biomedical and DNA Data analysis, Data mining for financial data analysis, Data mining for retail industry, Data mining for telecommunication industry, Data mining for weather forecasting, Data mining for crime analysis etc.

The traditional and age-old system of intelligence and criminal record maintenance has failed to live up to the requirements of the existing crime scenario. Manual processes neither provide accurate, reliable and comprehensive data round the clock nor does it help in trend prediction and decision support. It also results in lower productivity and ineffective utilization of manpower. The solution to this ever-increasing problem lies in the effective use of Information Technology.

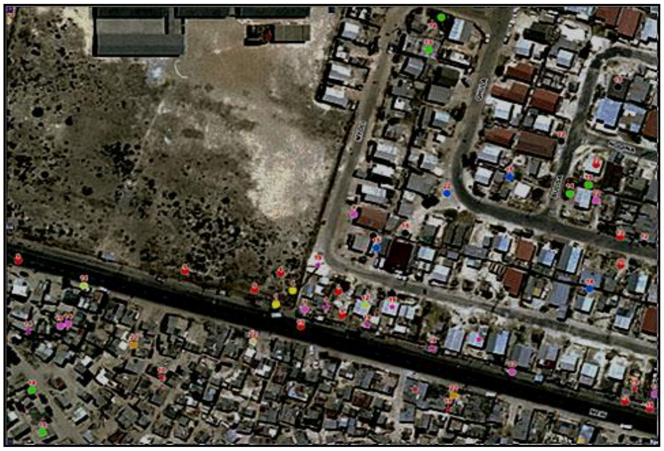


Figure 1. Hotspot Visualization

Result of Baseline Survey (VPUU Violence Protection through urban upgrading, Khayelitsha) Hotspot: A cluster of spots showing murder and robbery around a particular road intersection. A Crime Map was drawn up as a result of the Baseline Survey. A total of 11 priority crimes were identified, with robbery, murder and rape being the top priorities overall. In addition, places of particular concern, called hotspots were identified.

Each hotspot was examined to see the prevalent reasons for crime occurrence in that area.

Rest of the paper is organized as follows, Section II contains the related work done in the area of crime data mining, Section III contains methodology adopted in the research, Section IV describes Data analysis and Interpretation techniques used in the research, Section V discusses the experiments and observation, Section VI concludes the research work with future directions.

II. Related Work

Various research have been done in past, relating to crime pattern prediction and prevention. Brown [1] has constructed a software framework for mining data in order to catch professional criminals. They proposed an information system that can be used to catch the criminals in their own regions. The software can be used to turn data into useful information with two technologies, data fusion and data mining. Data fusion manages fuses and interprets information from multiple sources, and it overcomes confusion from conflicting reports and cluttered or noisy backgrounds. Data mining is concerned with the automatic discovery of patterns and relationships in large databases. His software was called ReCAP (Regional Crime Analysis Program), which was built to provide crime analysis with both technologies.

Abraham et. al. [2] proposed a method to employ computer log files as history data to search some relationships by using the frequency occurrence of incidents. Then, they analyze the result to produce profiles. The profiles could be used to perceive the behavior of criminal. It should be noted that the types of crime could be changed over time influenced by the variations of globalization and technology. Therefore, if we want to prevent the crime efficiently, the behavior must be used with another kind of knowledge.

Bruin et. al. [3] introduced a new distance measure for comparing all individuals based on their profiles and then clustering them accordingly. This method yields a visual clustering of criminal careers and enables the identification of classes of criminals. They demonstrated the applicability of data mining in the field of criminal career analysis. Four important factors play a role in the analysis of criminal careers: crime nature, frequency, duration and severity. They also develop a specific distance measure to combine this profile difference with crime frequency and change of criminal behavior over time.

Nath [4] proposed the use of clustering algorithms for data mining approach to help detect the crime patterns and speed up the process of solving crime. He proposed that this framework can be implements with geospatial plot of crime and helps to improve the productivity of detectives and other law enforcement officers.

Song Lin and Donald E Brown[5] proposed an outlier-based data association method. An outlier score function is defined to measure the extremeness of an observation, and a data association method is developed based upon the outlier score function. He applied this method to the robbery data in

Richmond, Virginia, and compared the result with a similarity-based association method. The results show that the outlier-based data association method is promising.

C.P. Johnson [8], proposed GIS can be used from crime analysis; Geographic Information System (GIS) uses geography and computer-generated maps as an interface for integrating and accessing massive amounts of location-based information. GIS allows police personnel to plan effectively for emergency response, determine mitigation priorities, analyze historical events, and predict future events. It is used world over by police departments, both large and small, to provide mapping solutions for crime analysis, criminal tracking, traffic safety, community policing, Intranet/Internet mapping, and numerous other tasks. Guangzhu et. al. [9] proposed a new similarities measure method, segmented Multiple Metric Similarity Measure (SMMSM) to improve the accuracy of similarity measures. Attributes are divided into different groups according to their importance to similarity.

Leong et. al.[10] proposed a new crime pattern analysis model, STEM(Space Time Event Model), this model allows user to investigate the spatio-temporal patterns of events..

III. Methodology

The objective of the study is to build tool to map crime that can help police to protect citizen more effectively. Crime Maps may prove valuable in solving crimes.

a) Display spatial patterns of events

b) Produce thematic maps

It also aims at performing crime analysis to provide timely and pertinent information relative to crime patterns and trends that will assist administrative personnel in planning to prevent and suppression of criminal activities. And to perform outlier based data association method, to find extreme observation from available data.

The methodology used in the research comprises of six steps:

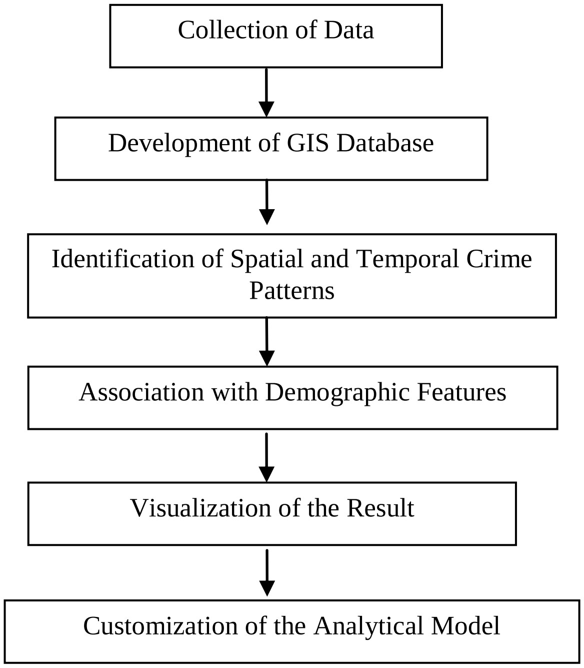

Collection of Data
Development of GIS Database
Identification of Spatial and Temporal Crime Patterns
Association with Demographic Features
Visualization of the Result
Customization of the Analytical Model

Figure 2. Methodology used

1. *Collection of Data*: First step, is collection of data, since crime data is available at different places, data needs to be gathered and combined to meet the requirement for research study.

2. *Development of GIS Database*: Second Step, The crime data is compiled by matching the addresses where the crimes occurred with their geographic locations. We will get a processed data which can be used for building crime maps.

3. *Identification of Spatial and Temporal Crime Patterns*: Third Step, after developing GIS database, the data can be used to find pattern based on spatial and temporal parameters i.e., to find occurrence of specific crime in a given area during a given period.

4. *Association with Demographic Features*: A combination of the crime and specific location will be used to link demographic characteristics with the location of the crimes. Population traits such as median income, people religion, and other demographic features are collected to suggest the association between the demographic characteristics of the areas and the location of the crimes.

5. *Visualization of the Result*: A series of maps will be generated to graphically display the location of the crime and the demographic characteristics of the specified location in which the crimes were committed. In addition, statistics can be compiled to show the findings of the analysis.

6. *Customization of the Analytical Model*: Final step is to build a model, which can be integrated with CCIS, to provide aiding tool in the working process of criminologist.

IV. DATA ANALYSIS AND INTERPRETATION

The crime data can be analyzed using any of the data mining and representation techniques:

1. *Outlier-based data association* method can be used in crime data analysis to find the extreme observation from the crime data. Outlier helps in finding patterns which are useful, when we are looking for data that does not follow the normal trend.

2. *Crime mapping*: Maps offer crime analysts graphic representations of crime-related issues. An understanding of where and why crimes occur can improve attempts to fight crime. Mapping crime can help police protect citizens more effectively. Simple maps that display the locations where crimes or concentrations of crimes have occurred can be used to help direct police to places they are most needed. Policy makers in police departments might use more complex maps to observe trends in criminal activity, and maps may prove invaluable in solving criminal cases.

Display spatial patterns of events: Digital maps are the quickest means of visualizing the entire crime scenario. The locations of crime events, arrests, etc. can be routinely displayed on maps. This provides an easy method of viewing activities in an area rather than searching through a listing of events. Maps can also be used to convey more than one type of information at a time. Crime locations can be symbolized according to the day of week, type of crime, modus operandi (a particular suspect's method of operation when committing a crime) or frequency.

Integrate community characteristics: Community characteristics (for e.g., slums, markets, colleges, parks, alcohol permit locations, red light area, etc.) can be routinely displayed on maps while analyzing crime patterns to interpret relationship between these characteristics and the crime. For example, the locations of aggravated assaults, robberies and alcohol permits can be displayed to see if crime is clustering around locations that sell alcohol. Other mapping data such as bus routes and public housing can also be displayed at the same time to analyze relationships between neighborhood characteristics and crime.

Produce thematic maps: Maps can be produced at any geographic level (e.g. Police stations, divisions, or zones) to aid in the analysis of crime patterns. Each response area can be shaded to represent the number of crimes that occurred in that area during a specific time frame. The darker shade depicts the frequency of crimes occurrence within the response area. These thematic maps can also be used to show the change in an area's crime rate. The percent change in the number of crime incidents can be displayed by shading each area according to whether there was an increase, decrease or no change.

3. *Crime Analysis:* It is a set of systematic, analytical processes directed at providing timely and pertinent information relative to crime patterns and trend correlations to assist the operational and administrative personnel in planning the deployment of resources for the prevention and suppression of criminal activities, aiding the investigative process, and increasing apprehensions and the clearance of cases. It supports a number of department functions including patrol deployment, special operations, and tactical units, investigations, planning and research, crime prevention, and administrative services.

Crime analysis can be divided into three categories:

Tactical: An analytical process that provides information used to assist operations personnel in identifying specific and immediate crime trends, patterns, series, sprees and hotspots, providing investigative leads and clearing cases. Analysis includes associating criminal activity by method of the crime, time, date, location, suspect, vehicle, and other types of information.

Strategic: Concerned with long-range problems and projections of long-term increases or decreases in crime (crime trends). Strategic analysis also includes the preparation of crime statistical summaries, resource acquisition, and allocation studies.

Administrative: Focuses on provision of economic, geographic, or social information to administration.

4. *Clusters of events* (hotspots): GIS identifies areas that contain dense clusters of events (hotspots). These high concentration areas usually demand special police attention. For example, GIS allows an analyst to identify all of the areas in a police station area where at least 5 robberies occurred within a 1km radius. These areas are then outlined on the map. Using GIS to identify hotspots provides a consistent method to measure concentrations of criminal events over time. Hotspots of violent crime, robbery, residential burglary, commercial burglary, auto theft, rape, etc. can be calculated every month for each police station area. Compare locations of hotspots across time: Crime hotspots that have been identified over several months can be displayed at the same time. This allows for the identification of areas with chronic problems and indicates the direction in which a particular crime may be shifting. These types of

maps can also be used to solicit resources for an area from other public and/or private agencies.

Compare hotspots of different crime types: Hotspots of different offence types can be displayed to identify where they overlap. For instance, residential burglary hotspots can be displayed along with robbery hotspots to discover where they overlap. A more detailed analysis of these intersecting areas can then be performed.

Shade grid cell maps: While multiple crimes (or events) at the same locations are not visible on a pin map, they are clearly accounted for in a grid cell map. For instance, ten thefts occurring at a single location would be represented by a single point on a pin map. Using grid cell mapping, all ten thefts would be counted toward the total for the grid cell that would determine the shade of the cell. The darker the shade, the higher the number of incidents occurred at that particular location.

5. *Crime investigation*: Historically, the causes and origins of crime have been the subject of investigation. Some factors known to affect the volume and type of crime occurring from place to place are:

- Population density and degree of urbanization.
- Variations in composition of the population, particularly youth concentration.
- Stability of population with respect to residents' mobility, commuting patterns, and transient factors.
- Modes of transportation and highway system.
- Economic conditions, including median income, poverty level, and job availability.
- Cultural factors and educational, recreational, and religious characteristics.
- Family conditions with respect to divorce and family cohesiveness.
- Climate.
- Effective strength of law enforcement agencies.
- Administrative and investigative emphases of law enforcement.
- Citizens' attitudes toward crime.
- Crime reporting practices of the citizen.

V. Experiment and Observation

For conducting experimental study we picked Delhi as study area. It lies between Latitude 28.38 N , Longitude 77.12 E. The size of study area is 1483 Sq. K.M. having density 9340 per Sq. K.M. Using Delhi Crime Dataset, where the Data is arranged in the form Table under the heads Crime Date, Location of Incident, Crime Type, Longitude, and Latitude.

Table 1. Sample Crime Dataset

Sno	Crimedt	location	Crime Type	Long	Lat
1	3-Jan-08	subhash vihar	robbery	77.2688	28.6975
2	3-Jan-08	gorav nagar	accident	77.0547	28.6998
3	3-Jan-08	west st nagar	accident	77.1925	28.7476
4	10-Jan-08	vasundara enc	robbery	77.3163	28.5999
5	10-Jan-08	lado sarai	robbery	77.1986	28.5255
6	6-Feb-08	pochan pur	robbery	77.0465	28.5626

The crime records present in the database and are mapped using GIS tools. Boundary map of Delhi was obtained from Survey of India, and exported to Google earth, where an outline map was created and imported in MapInfo Professional 10.0 software, and crime data was plotted on the map.

The following outcomes were obtained:
A map displaying types of crimes in different legends

Table 2. Legends and their description used in Figure.3

Legends	Crime Type
★	Chain Snatching
★	Robbery
★	Theft
★	Theft Accident

The map displays the location and the type of crime committed in that area. The concentration of crime is visualized through map. It further can be analyzed for demographic factors, to find the reason of crime concentration in that area.

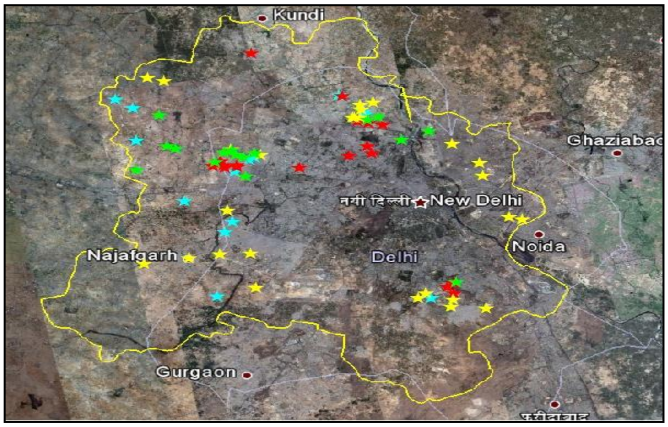


Figure 3: Delhi Map showing crime occurrence hotspots.

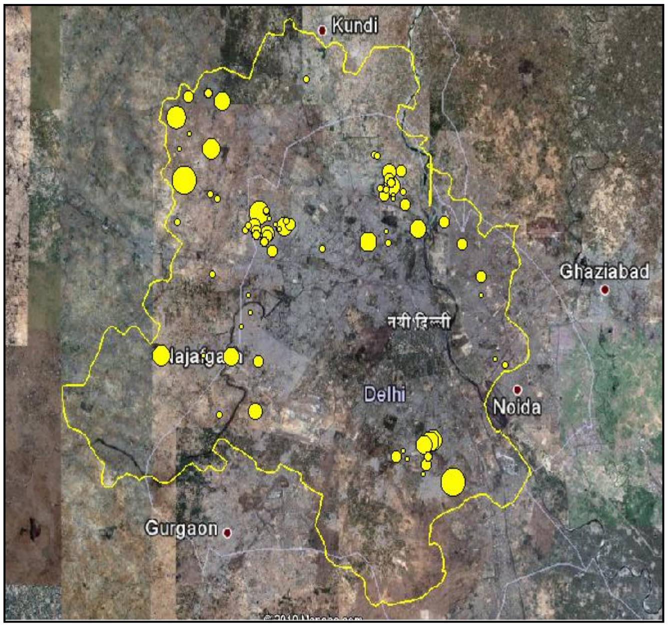


Figure 4: Delhi Map showing intensity of crime at specific area, thematic map

Map shown in Figure 4 shows another depiction of crime concentration in the form of bubble. The bigger bubble depicts the higher number of incidents occurrence in the specific area.

Table 3. Legends description used in Figure 4

Bubble Type	Incident Occurrence
●	Crime more than 20 incidents
●	Crime incidents less than 5

Longitude and Latitude of a crime record is added by finding the location from Google map. Information obtained is in the format Degree, Minute and Second, it needs to be converted to Decimal Degree to be plotted on the map.

Degrees Minutes and Seconds to Degrees Minutes.m (GPS) is computed as

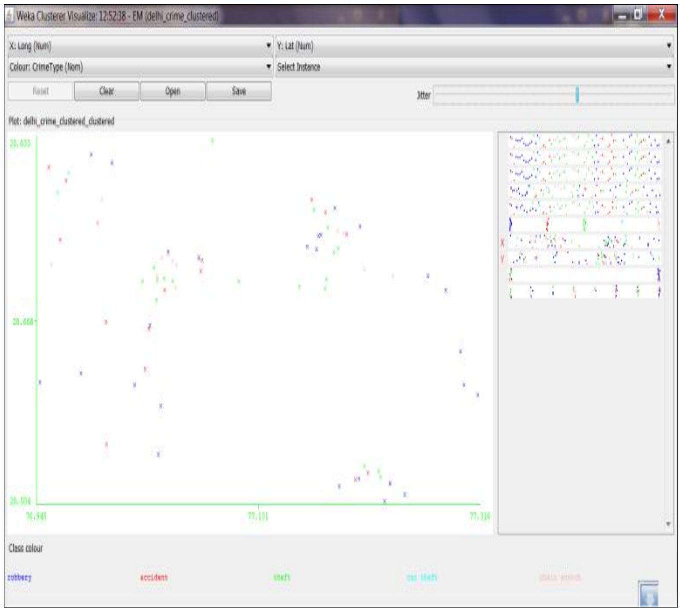
$$Minutes.m = Minutes + \frac{Seconds}{60}$$

Degrees Minutes.m to Decimal Degrees .d is computed as

$$.d = \frac{Minutes.m}{60}$$

Decimal Degree is computed as

$$Decimal\ Degrees = Degrees + .d$$


Weka Clusterer Visualize: 12:52:38 - EM (delhi_crime_clustered)

X: Long (Num)

Y: Lat (Num)

Colour: CrimeType (Nom)

Select Instance

Plot: delhi_crime_clustered_clustered

Figure 5: Cluster Analysis performed in Weka Software

The data then is analyzed in Weka Software to find clusters of crime according to longitude and latitude. Figure 5 shows Longitude in X axis, Latitude in Y axis, crime is mapped in the graph. Figure 3 and Figure 4 shows concentration of crime at specific areas.

VI. Conclusion

In this paper a Geo-statistical Spatial Clustering Technique and Mapping Techniques are presented. This technique mainly uses the geographical information related to crime data to plot the maps and provide clustering. These techniques are of great importance for future crime prediction and control. It can provide valuable information for government authorities for combating the specific crime situation and making the area safer for citizen. The paper suggested the use of thematic maps, hotspots visualization techniques for better crime management and use of clustering technique for crime concentration depiction.

References

[1] D.E. Brown, "*The regional crime analysis program (RECAP): A frame work for mining data to catch criminals*", In the Proceedings of 1998 IEEE International Conference on Systems, Man, and Cybernetics, Vol. 3, pp. 2848-2853, 1998.

[2] T. Abraham and O. de Vel, "*Investigative profiling with computer forensic log data and association rules*,", In the Proceedings of the 2002 IEEE International Conference on Data Mining (ICDM'02), pp. 11 – 18, 2002.

[3] J.S. De Bruin, T.K. Cocx, W.A. Kosters, J. Laros and J.N. Kok, "*Data mining approaches to criminal career analysis*" in the Proceedings of the Sixth International Conference on Data Mining (ICDM'06), pp. 171-177, 2006.

[4] S.V. Nath, "*Crime pattern detection using data mining*," in the Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, pp. 41-44, 2006.

[5] S. Lin, D. E. Brown, "*An Outlier-based Data Association Method For Linking Criminal Incidents*", Decision Support System, Vol 41, Issue 3, pp. 604-615, 2006.

[6] J. Han and M. Kamber, "*Data Mining: Concepts and Techniques*," Morgan Kaufmann publications,India, pp. 1-39, 2006.

[7] M. Gupta , B. Chandra and M. P. Gupta, "*Crime Data Mining for Indian Police Information System"*, Computer Society of India, pp. 389-397, 2008.

[8] C.P. Johnson, "*Crime Mapping and Analysis Using GIS*", In the Conference of Geomatics in Electronics Governamce, 2000.

[9] G. Yu, S. Shao, B. Luo, "Mining Crime Data by using New Similarity Measure", IEEE, 2008. In the Proceedings of the 2008, Second International Conference on Genetic and Evolutionary Computing, pp. 389-392, 2008.

[10] K. Leong, S. Chan, V. Ng, S. Shiu, "*Introduction of STEM: Space Time Event Model for Crime Pattern Analysis*", Asian Journal of Information Technology,Vol 7 Issue 12, pp.516-523, 2008.

Authors Profile

Dr. Neetu Narwal is Doctorate in Computer Science from Banasthali Vidyapith, Rajsathan, India. She is working as Assistant Professor in Department of Computer Science, Maharaja Sruajmal Institute. She has 16 years of Teaching experience and five years on Industry Experience. Her research areas include web content mining, social media mining and crime mining.