Using Data Mining Techniques to Analyze Crime patterns in Sri Lanka National Crime Data

K.P.S.D. Kumarapathirana

139169A

Dissertation submitted to the

Faculty of Information Technology, University of Moratuwa, Sri Lanka

for the partial fulfillment of the requirements of the

Degree of Master of Science in Information Technology.

March 2016

Declaration

We declare that this thesis is our own work and has not been submitted in any form for another degree or diploma at any university or other institution of tertiary education. Information derived from the published or unpublished work of others has been acknowledged in the text and a list of references is given.

Name of Student

K. P. S. D. Kumarapathirana

Signature of Student

.....

Date: 24.04.2016

Supervised by

Name of Supervisor

S. C. Premaratne

Signature of Supervisor

UOM Verified Signature

Date: 24/04/2016

Acknowledgements

I would like to express my gratitude to my supervisor, Mr. S. C. Premaratne, Senior Lecturer in University of Moratuwa, Sri Lanka whose expertise, understanding, and patience, added considerably to my research experience. I appreciate his vast knowledge and skill in many areas, and his assistance in writing reports.

I would like to thank the other lecturers of University of Moratuwa, Sri Lanka, especially, Prof. Asoka Karunananda for the knowledge and assistance they provided at all levels of the research project.

Moreover, a very special thank should go out to Mr. C. V. Millawithana, Criminologist, Crime Record Division, Department of Police, Sri Lanka and his staff members for assisting me in collecting data sets and giving me their valuable comments on the research goals and objectives.

I would also like to thank all the batch mates of the M.Sc. in IT degree program who gave their valuable feedbacks to improve the results of the research and my family for the support they provided me through my entire life and in particular. I must acknowledge my husband and best friend, Tharindu, without whose love, encouragement and editing assistance, I would not have finished this thesis.

Abstract

Crime is one of the dangerous factors for any country. Although crimes could occur everywhere, it is common that criminals work on crime opportunities they face in most familiar areas for them. The ultimate goal of crime analysis is to identify likely targets for police intervention and prevent crime or solve past crimes by making statistical predictions. Criminals and victims follow common life patterns and most of the time overlaps in those patterns indicate an increased likelihood of crime. Geographic and temporal features influence the where and when of those patterns.

Our proposed solution consists of four major modules namely; Hotspot Analysis Module, Offender Profiling Module, Victim Profiling Module and Predicting Suspects Module. Data related to fast crimes which is used for the analysis is collected from Department of Police, Sri Lanka. Hotspot analysis module identifies crime hotspots considering geographical data of past crime where victim profiling and suspect profiling modules identify the patterns or groups of victims who are most vulnerable and suspects who share same characteristics. First three modules are developed based on simple k-means clustering algorithm where the fourth module is based on simple k-means clustering and j48 algorithm to generate the classifier model which can be used to predict the cluster of suspects of a crime.

The results of this analysis can be used by law enforcers to find general and specific crime trends, patterns, and series in an ongoing, timely manner in order to take advantage of the abundance of information existing in law enforcement agencies, the criminal justice system, and public domain, to maximize the use of limited law enforcement resources, to have an objective means to access crime problems locally, regionally, nationally within and between law enforcement agencies, to be proactive in detecting and preventing crime, to meet the law enforcement needs of a changing society and to understand the criminal behaviors.

Table of Content

List of Figures

List of Tables

Chapter 1

Introduction to Crime Record Analysis in National Crime Records in Sri Lanka

1.1 Prolegomena

Widespread infrastructural and industrial development of many developing countries, developed countries, and third world countries and their cities is rapidly transforming the whole world into a single globalized village. The effects of this globalization have been accelerated the rate of crimes in major cities. Moreover, the concern about nationwide security in Sri Lanka has increased significantly since the major terrorist attacks turned out during last three decades. Security of the citizens is the major concern of any country with no matter it is a third world country, a developed country or a developing country. And this has become an issue which is continuing to grow in intensity and complexity. However, the birth and growth of crime in a community is based on many characteristics related to the community and society. These characteristics are different nationalities and religions in a society, different income levels, different categories of age, different family structures (single, divorced, married, and number of kids), different levels of education, the locality or the community where people live (cheap or expensive housing, size of houses, number of rooms), number of police officers allocated to a region, number of employed and unemployed people in the area, etc.

1.2 Background and Motivation

Crime occurs in a variety of forms. These forms has been recognized internationally as Traffic Violations (driving under influence (DUI), fatal/personal injury/property damage traffic accident, road rage), Sex Crime (Sexual offenses, sexual assaults, child molesting, organized prostitution), Theft (robbery, burglary, larceny, motor vehicle theft, stolen property, theft of national secrets or weapon information), Fraud (forgery and counterfeiting, frauds, embezzlement, identity deception, transnational money laundering,

identity fraud, transnational financial fraud), Arson (arson on buildings, apartments), Gang / drug offenses (narcotic drug offenses (sales or possession), transnational drug trafficking), Violent Crime (criminal homicide, armed robbery, aggravated assault, other assaults, terrorism (bioterrorism, bombing, hijacking, etc.), and Cyber Crime (internet frauds, illegal trading, network intrusion/hacking, virus spreading, hate crimes, cyber-piracy, cyber-pornography, cyber-terrorism, theft of confidential information) [14]. Out of them volume crimes such as burglary and shoplifting are happening island-wide. Sri Lanka Police Department has categorized grave crimes in to 21 classes: Abduction/Kidnapping, Arson, Mischief over Rs.5000/=, House Breaking & Theft, Grievous Hurt, Hurt by Knife etc, Homicide/Abetment to commit suicide, Attempted Homicide, Rape/ Incest, Riots, Robbery, Unnatural Offense/Grave sexual abuse, Extortions, Cheating/Misappropriation, C.B. trust over Rs. 100,000/=, Theft of property including praedial produce over Rs. 5,000 & Cycle & Cattle thefts irrespective of their value, Counterfeiting Currency, Offence against the state, Cruelty to children & sexual exploitation of children, Procuration/ Trafficking, Offenses under the offensive weapons act, Possession of Automatic or Repeater Shot Guns and Manufacture or any quantity Heroin, Cocaine, Morphine, trafficking, import or possession of dangerous Drugs of an above 2 gms of Heroin, 2 gms Of Cocaine, 3 gms of Morphine, 500 gms of Opium, 5 kgs of Cannabis and 1kg of Hashish [2].

Most of the crime data in Sri Lanka is manually recorded in criminal record books of individual police stations in the form of statements made by victims, eye witnesses, other witnesses and arrested suspected criminals, reports by informers and information gathered by the police themselves. In addition to that, police stations keep pin-up maps in order to maintain information on the locations of crime incidents.

Department of Police, Sri Lanka employ specialists in crime analysis, especially criminologists, people who have specialized training in a variety of disciplines including investigation techniques, criminal psychology and information technology. It is their task to assist investigating officers by analyzing crime trends and patterns, and identifying links between crimes. In the developed countries like the UK, a majority of crime

prevention forces use different types of relational database management systems (RDBMS) for recording and subsequent analysis of crime [1].

Present responsibilities of the Crimes Division are maintaining of crime statistics and its evaluation, compilation island wide crime statistics related to Grave Crimes, Reportable Crimes, Serious & Organized Crimes and Numerous other allied subjects, compilation of statistics pertaining to offenders/Criminals, preparing and maintaining Maps, Graphs charts etc. in respect of Crimes and other allied subject, preparation of returns, and reporting to IG Police & S/.DIG Crimes [16].

1.3 Crime and Criminology

"Legally, crime is the breaking or breaching of the criminal law (penal code) that governs a particular geographical area (jurisdiction) aimed at protecting the lives, property and the rights of citizens of belonging to that jurisdiction. Crime is an offence against a person, or his/her property or the State regulation" [2].

According to Webster Dictionary, crime is "an act or the commission of an act that is forbidden or the omission of a duty that is commanded by a public law and that makes the offender liable to punishment by that law".

"Criminology is an area that focuses the scientific study of crime and criminal behavior and law enforcement" [5]. The procedure aims to identify characteristics of crime incidents and the patterns of criminals. The towering quantity of crime datasets and also the complexity of relationships between these kinds of data have made criminology an appropriate field for applying data mining techniques where important results can be gained providing betterment for the society. "Crime analysis is a part of criminology that includes exploring and detecting crimes and their relationships with criminals" [5]. The need for a good crime analysis tool to catch criminals is rising with the technologies used in crimes happening in nowadays.

In the recent years, criminal justice authorities have turned to those criminologists to get some concrete explanations to the identified crime patterns and criminal patterns and

ultimately any crime analysis tool, same as a professional criminologist, must provide the solutions to five basic questions of crimes:

1. Where would the next crime occur (areas can be identified with crime hotspots, type of location)?
2. What type of crime is happening at the identified location in the previous question (e.g. drug sales, robbery, burglary)?
3. When is the identified crime happening (temporal details e.g. day of week, hour of day, season)?
4. Who is likely to commit the crime (repeat offender traits and locations)?
5. What is the motive for crime occurring?

1.4 Data Mining Approach

"Data mining guarantee easy, convenient, and practical way to explore very large databases for organizations and users" [3]. Many classic data mining techniques such as association rule mining, classification, and clustering have been successfully used for crime analysis in different researches.

1.5 Crime Analysis

The ultimate goal of any crime analysis system is to identify likely targets for police intervention and prevent crime or solve past crimes by making statistical predictions. Predictive data mining methods may allow police forces to work more proactively with limited amount of human resources attached to police forces. The objective of this research is to develop effective strategies that will prevent crime or make investigation efforts more effective. A policing strategy must produce tangible results to be considered as effective. "A perfect crime analysis tool should analyze and summarize collected data, identify the patterns of happening crimes quickly and predict for future crimes using the patterns generated" [10].

The complexity of crime and criminal related data has made data mining a rapidly growing field among criminologists, crime investigators and crime analysts. The growing

volumes of crime and criminal related data recorded in databases of Department of Police, Sri Lanka and also the complexity of the relationships between these kinds of data has forced the traditional and manual crime analysis methods to become outdated and difficult to perform. These methods require a considerable amount of time and human resources and they are not able to get all effective parameters/relationships involved due to their high amount of human interference.

Crime Division of department of Police, Sri Lanka has a separate division to keep records, i.e. Crime Records Division. Crime Record division keeps records on gravy crimes occurred in island wide police divisions. They keep crime record number, the police division and station to which the incident is reported, type of the offence (one of the above 21 categories), date and time of the offence (a day is divided into four time intervals: 00.00 Mid night to 06.00 AM, 06.00 AM to 12.00 Noon, 12.00 Noon to 06.00 PM and 06.00 PM to 00.00 Mid night), reported date, place of offence (three categories: house, shop and other), suspect's relation to the victim, age and gender of the victim, in case of a murder, type of the murder (murder / homicide) and number of victims, in case of a rape, type of rape (four categories: child rape, child abuse, adult rape and love affair), in case of a house break and theft, place (home/shop/other), items of theft (electronic equipment, money, etc), the relationship to the house owner, in case of a robbery, whether single person or a group is involved, type of robbery (vehicle, craft, etc), whether a weapon is used, identified robbery group (police/force/civil), in case of drug handling, type of the drug, offence type (manufacture, transport, keep, etc), in case of a kidnapping or an abduction, type of the person being kidnapped (child / adult), in case of keeping weapons, reason for keeping, in case of counterfeiting currency, type of counterfeiting (printing, keeping, etc) and amount, modus operand adopted (from window, from main door, breaking main door, from the roof, etc) and, finally, the motive for the offence.

The problem of detecting specific patterns of crime that are committed by the same offender or group can be solved using machine learning approach. The learning algorithm processes information similarly to how crime analysts process information manually. The algorithm searches through the database looking for patterns of crimes in a growing manner and in the rest of the database, and tries to identify the modus operandi (M.O.) or

the pattern of committing crimes of the particular offender. The modus operandi is defined as "the set of habits that the offender follows, and is a type of motif used to characterize the pattern" [4]. As more crimes are committed by the same suspect and they are recorded in the database, the modus operandi becomes more well-defined. Criminals are self-consistent in the way they commit crimes. Different criminals may have completely different modus operands. Some offenders like to work on weekdays while the residents are at work and some operate at night, while the residents are sleeping. Some offenders favor large apartment buildings, where they can break into multiple units in one day; others favor single-family houses, where they might be able to steal more valuable items. These patterns of committing crimes can be dynamic.

With situational awareness and anticipation of human behavior, police can identify and develop strategies to prevent criminal activity by repeat offenders against repeat victims.

Dispute of having dynamic patterns, statistically, crime is always predictable because criminals tend to operate according to their patterns. That is, they tend to commit the type of crimes that they have committed successfully in the past, generally close to the same time and location. Although this is not universally true, it occurs with sufficient frequency to make these methods work reasonably well.

Criminals and victims follow common life patterns and most of the time overlaps in those patterns indicate an increased likelihood of crime. Geographic and temporal features influence the where and when of those patterns. As they move within those patterns, criminals make "rational" decisions about whether to commit crimes, taking into account such factors as the area, the target's suitability, and the risk of getting caught.

1.6 Goal of the Research

The main requirement of this analysis is to identify the mining methods which cope with growing volumes of crime data. This analysis can be used to inform law enforcers about best methodologies to find general and specific crime trends, patterns, and series in an ongoing, timely manner in order to take advantage of the abundance of information existing in law enforcement agencies, the criminal justice system, and public domain, to

maximize the use of limited law enforcement resources, to have an objective means to access crime problems locally, regionally, nationally within and between law enforcement agencies, to be proactive in detecting and preventing crime, to meet the law enforcement needs of a changing society and to understand the criminal behaviors.

1.7 Objectives

Our study aims to find spatial and temporal criminal hotspots using a set of real-world datasets of crimes. We will try to locate the most likely crime locations and their frequent occurrence time. In addition, we will predict what type of crime might occur next in a specific location within a particular time. Finally, we intend to provide an analysis study by combining our findings of a particular crimes dataset with its demographics information.

Our main objectives can be outlined as follows.

- Identify the nature of crime and the crime prevention process
- Collect data
- To explore and choose among the various data mining software and algorithms that support clustering and association rule mining techniques to experiment with crime records.
- To select data cleaning algorithms that clean the crime dataset, by removing unwanted data and filling missing values in an efficient manner
- Feature selection using suitable methods
- Explore and enhance clustering algorithms to identify crime patterns from historical data
- Identify hotspots of crime
- Detecting crime patterns
- Explore and enhance classification algorithms to predict future crime behavior based on previous crime trends
- Compare the clustering and association rules data mining techniques and select the one which performs the best results.

- Provide information to formulate strategies for crime prevention and reduction based on the finding of the research
- Identify and analyze common crime patterns to reduce further occurrences of similar incidence

1.8 Acknowledgments

We are grateful for the assistance provided by the professional staff of several divisions in the Department of Police, Sri Lanka. Especially, Mr. C. V. Millavithana, the criminologist, Crime Records Division of Department of Police, Sri Lanka.

1.9 Overview of the Report

In this report, we write the progress of the research carried on. First chapter includes a summary and overview of the research problem and the solution is explained briefly. The second chapter includes the literature survey done based on the topic and it summarizes the technologies and the algorithms used in similar cases. Third chapter has summarized the technologies in to four major areas focusing the objectives of the research and the fourth chapter describes our approach to analyze crime records of Sri Lanka. Fifth chapter includes the design of the analysis. Chapter 06 covers the implementation details of the four modules including the algorithms and the techniques selected where chapter 07 evaluates them. Finally, chapter 08 discusses the results, limitations and future developments for the solution.

1.10 Summary

This chapter has introduced the research problem and the solution. Predictive policing is the application of analytical techniques, particularly quantitative techniques—to identify likely targets for police intervention and prevent crime or solve past crimes by making statistical predictions. Several predictive policing methods are currently in use in law enforcement agencies across the United States, and much has been written about their effectiveness. Another term used to describe the use of analytic techniques to identify likely targets is forecasting. Although there is a difference between prediction and

forecasting, for the purposes of this guide, we use them interchangeably. In the next chapter, we review some researches carried out related to our research problem.

Chapter 2

State of the Art of Exploring Current Findings in Crime Record Analysis

2.1 Introduction

In this chapter we discuss some of the findings recently done in crime record analysis. We conducted a literature search of academic papers, vendor tool presentations, and recent presentations at conferences, drawing lessons from similar predictive techniques used in related research. We have selected 15 research papers based on our main goal and the research question. All the technical papers are published after 2000. A large variation of techniques and algorithms are considered when selecting research papers for the literature review. Moreover, Google Scholar research articles were selected and they all are finally categorized according to the techniques and algorithms used in their researches.

2.2 Framework 1

A software framework called ReCAP (Regional Crime Analysis Program) was developed for analyzing crime records using data mining and data fusion technologies in order to grab professional criminals. Data fusion has been used to manage, fuse and interprets information from multiple sources of crime records. The main purpose of this framework was to overcome confusion from conflicting reports of crime analysis in cluttered or noisy backgrounds [4].

2.3 Framework 2

Another framework for crime trends was developed later using a new distance measure. This technique is used to compare all individual suspects or offenders based on their crime profiles and then cluster them accordingly to identify patterns. This method also provides a visual clustering of criminal careers and identification of classes of criminals

[6]. According to the findings, the most successful method which can be used to identify specific crime patterns includes review of crime reports daily basis and the comparison of the reports on past crimes [8]. But, this process can be time consuming in an extraordinary manner.

2.4 Framework 3

Another framework, which uses Exploratory Data Analysis (EDA) techniques is interactive and the results can be visualized, and there are many effective graphical presentation methods for data sets with moderately small less number of dimensions. When the number of variables used in the analysis is increased, the visualization of points gets difficult. The methods used in crime spatial data analysis can be classified into those concerned with visualization of data, those for exploratory data analysis and methods for the development of statistical models [9].

2.5 Framework 4

Spatial point patterns (SPP) which are based on the coordinates of locations of crime incidences are been used in this framework with the aim to detect whether there is a random, clustered or regular distribution in the point patterns. SPP is typically interpreted as analysis of clustering, especially using simple k-means algorithm. This method assesses clustering of crime incidences in detection of hot spots where time and space relationship analysis is required [15].

This framework uses three basic methods, namely, Knox's method, Mantel's Method and K-nearest neighbor method. All the methods require a distance matrix in order to identify the relationships between the spatial and as well as temporal related data of different crime incidences.

Knox's method requires critical distance in temporal data and, as well as, in spatial data defining the closeness with regard to the distance matrix and to determine these critical distances, subjective decision is required [15].

However, Mantel approach does not require a distance matrix to measure critical distances but it uses both temporal and spatial data for the analysis, in spite of being insensitive to non-linear relations [15].

The K-nearest neighbor is focusing the approximate randomization of the Mantel product statistic [15].

This research emphasizes on the increasing amount of crime data to very large quantities (into Giga Bytes) has raised the need for advanced and efficient techniques for analysis. Data mining was identified as an analysis and knowledge discovery tool which has immense potential for crime data analysis.

2.6 Framework 05

Malathi introduces a tool which is effective in terms of analysis speed, identifying common crime patterns in Indian society and future predictions for the crime incidents with a potential value in the current changing crime scenario [12].

2.7 Framework 6

Some other crime analysis tools use the geo spatial plots to indicate the locations of crime incidents. The manual pin maps used by police stations and other law enforcement authorities can be replaced by the use of Global Positioning System (GPS) technology [4].

2.8 Framework 7

COPLINK is another framework which has been an earlier projects in collaboration with Arizona University and the police department of US to extract entities from police narrative records. This allows law enforcement authorities to discover investigative case leads, visualize and analyze data on maps through time-sequence playback, centralize multiple data stores in one system and discover hidden value in existing information stores [5].

2.9 Framework 8

J. Agarwal, R. Nagpal and R. Sehgal have analyzed crime records and selected homicide crime incidents collected during the corresponding year and identified that the trend is descending from 1990 to 2011. They have selected the k-means clustering technique to identify the patterns in homicide crime records for extracting useful information from the crime dataset using RapidMiner tool. RapidMiner tool can be considered as a solid and complete package with many different flexible supporting options [11]. Figure 2.1 given below shows the proposed system architecture.

2.10 Some Other Frameworks

Another tool has been proposed by Mohler to identify the patterns of offender behavior which is changing significantly using the extracted features including frequency, seriousness, duration and nature of the crime incidences committed by them. Using those factors, the similarities between pairs of criminals are compared with reference to a new distance measure and the crime data set is clustered accordingly [7].

Different researches have used different techniques such as K-means clustering to detect crime pattern to speed up the process of solving crimes, Self Organizing Map (SOM) to link the offenders of serious sexual attacks and two phase clustering algorithm called AK-modes to automatically find similar case subsets from large datasets (Information Gain Ratio was used for attribute weighing) [3].

Table 1.1 The summary of the literature survey

Technology	Algorithm
Clustering	k-means Clustering
	AK-mode Algorithm
	Expectation-Maximization Algorithms
Classification	Decision Tree Algorithm
Frequent pattern mining	Fuzzy Association Rule Mining
	MV Algorithm

	Apriori Algorithm
	Apriori Growth Algorithm
	Predictive Apriori Algorithm
	FP-growth Algorithm
Outlier Detection	

2.11 Summary

This chapter discussed the findings of our literature study and summary of the literature survey is indicated in Table 1. It shows how different data mining techniques and data mining algorithms are used in crime record analysis in order to achieve different goals.

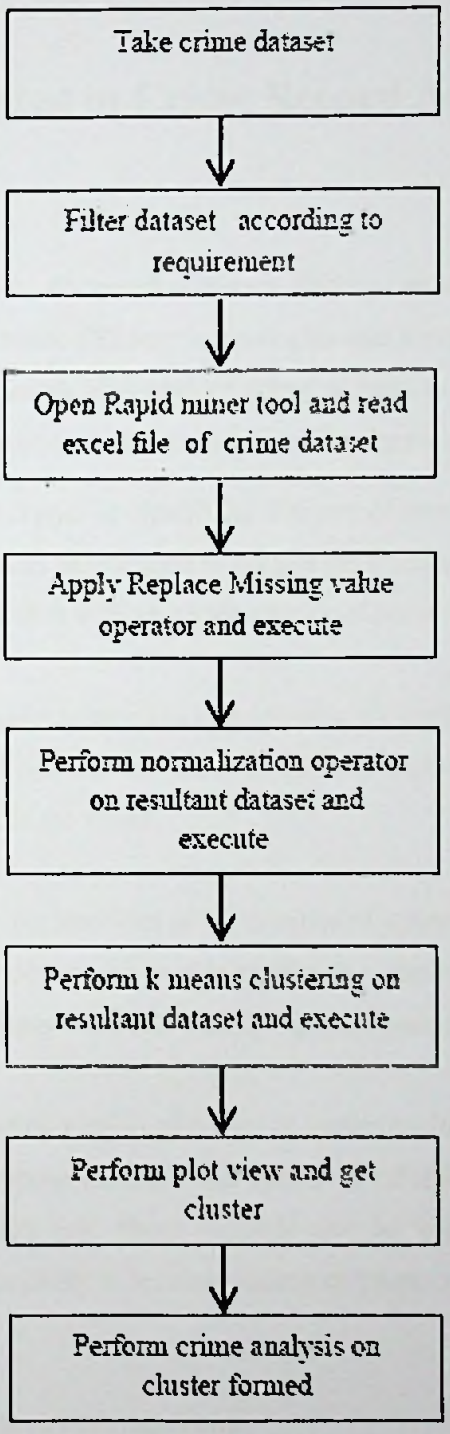


Figure2.1 Flow chart for Crime Analysis [11]

Chapter 3

Technology adapted in Crime Record Analysis

3.1 Introduction

In the previous chapter we discussed different findings in the area of crime record analysis. This chapter presents different technologies and methodologies used in crime record analysis. In our research of analyzing crime records to identify the patterns, we found that the methods can be divided into four broad categories:

1. Methods to predict crimes or identifying hotspots of crimes:

   These methods can be used to predict the place of next crime incident and the places and time slots with an increased risk of crime.

2. Methods for predicting suspects or offender/suspect profiling:

   These methods can be used to identify individuals at higher risk of committing crimes in the future.

3. Methods to predict the identities of the criminal of a new crime incident:

   These methods can be used to predict the criminal profiles that accurately match with more likely offenders with the specific past crime patterns.

4. Methods for predicting victims of crimes or victim profiling:

   Similar to those methods that focus on offenders, crime locations, and times of heightened risk, these methods can be used to identify groups or individuals who are likely to become victims of future crime incidents.

3.2 Methods related to predicting crimes

Major problem is identifying areas at increased risk and identifying geographic features that increase the risk of crime. This can be achieved using historical crime data and a range of additional data such as 119 calls and economical data. Different conventional

crime analysis techniques such as crime mapping for hot spot identification, basic regression models from a spread sheet application, graphing or mapping the frequency of crimes in a given area by time/day and finding locations with the greatest frequency of crime incidents can be applied while predictive analytics such as advanced hot spot identification models, risk terrain analysis, regression, classification, clustering models, near-repeat modeling, spatiotemporal analysis methods can be used to enhance the results of the analysis.

With the inspiration to spread out each crime's projected involvement to future crime risk within a certain area, Single and Dual Kernel Density Estimation use data related to individual crime incidents in order to find hot spots. This is done based on how much they are close to past actual crime incidents. This uses a mathematical function called a kernel. To estimate hot spots, Single Kernel Density Estimation uses only a single variable: crime incidents where Dual Kernel Density Estimation uses two variables, crime incidents and population density. This method generates a contour map, a heat map, or a surface view map with the more heavily weighted areas of high crime visually represented in either case. The hot spots are defined as areas above a certain threshold on each of the above types of maps [2].

Moreover, other approaches apply statistical regression to the problems related to crime for many years. This method is still has been used in crime record analysis in many approaches, in spite of the more complicated techniques which can provide additional information or avoid some pitfalls. Also, regression techniques can be used to get answers for the questions where the answer is a number with some confidence value associated with it. E.g.: "How many robberies will a neighborhood have next week?" [2].

Different approaches, such as, hot spot analysis, statistical regression, data mining, and near-repeat methods are widely used to find information on different questions related to the specific geographical location a crime will occur over a specified time horizon (when, varying from a day to a year, depending on the method and application) and, hence, type of person is likely to be the victim of the predicted crime [3].

In addition to that, temporal and spatiotemporal methods can be used to identify the exact time the next crime incident is most likely to be committed. These methods can also be

used to identify the victims (who) of the above predicted crime incident because they account for the ambient population, as well as local residents.

Furthermore, risk terrain analysis is another appropriate approach to discern the geographical factors that build the corresponding risk for crime incident to happen and look for some geographical locations that has higher risk for a specific type of crime (where).

3.3 Methods to identify individuals at high risk of offending in the future

Major problem is finding a high risk of an aggressive occurrence between criminal groups and identifying individuals who is likely to become offenders in future crimes, especially the novice at great risk of committing crimes repeatedly, domestic crime cases with a higher risk of having injuries or death and patients who is lack of mental health at greatest risk of future criminal behavior or violence. Different predictable crime analysis techniques such as manual review of income gang/criminal intelligence reports and health care data that summarizes known risk factors can be applied while predictive analytics such as near-repeat modeling on recent intergroup violence and regression and classification models using the risk factors can be used to enhance the results of the analysis [2].

3.4 Methods used to identify likely suspects of a new crime

Major problem is identifying suspects using a suspects' criminal history and the pattern of committing crimes or other partial data, determining which crimes are part of series (i.e. most likely committed by the same criminal), finding suspect's most likely anchor point, and finding suspects using sensor information such as GPS tracking, license plate readers and CCTV camera around the crime scene. Different conventional crime analysis techniques such as manually reviewing crime intelligence reports and drawing inferences, crime linking using a table to compare the attributes of crimes known to be in a series with other crimes, locating areas both near and between crimes in a series and manual requests and review of sensor data can be applied while predictive analytics such as computer assisted queries, statistical modeling to perform crime linking and geographic

proofing tools to statistically infer most likely points can be used to enhance the results of the analysis [4].

3.5 Methods used to identify crime victims

Major problem is identifying groups likely to be victims of various types of crime or vulnerable populations which share the same characteristics, identifying individuals who would be directly affected by at-risk locations or crime hotspots, identifying individuals at higher risk for being a victim (e.g., people engaged in high-risk criminal behavior) and identifying individuals at risk of being a victim of a domestic violence. Different predictable crime analysis techniques such as crime mapping to identify crime type hot spots, manually graphing or mapping most frequent crime sites and identifying people most likely to be at these locations, review of criminal records of individuals known to be engaged in repeated criminal activity and manual review of domestic disturbance incidents can be applied. Moreover, predictive analytics such as advanced models to identify crime types by their hot spot or the geographical locations with higher risk of happening crimes; risk terrain analysis, advanced crime-mapping tools to generate crime locations and identify workers, residents, and others who frequently visit these locations and advanced data mining techniques can be used on local and other accessible crime databases to identify the individuals at risk of being a victim and computer-assisted database queries of multiple databases to identify domestic and other disturbances involving local residents when in other jurisdictions can be used to enhance the results of the analysis [10].

3.6 Summary

This chapter summarized different technologies used in crime record analysis. To predict crimes, corresponding predictive analytics methods start, at the most basic level, with regression analyses and extend all the way to cutting-edge mathematical models that are the subjects of active research. Next chapter discusses our approach to solve the current problems related to crime record analysis in Sri Lanka.

Chapter 4

A Novel Approach for Crime Record Analysis in Sri Lanka

4.1 Introduction

Previous chapter summarized different methods used in different aspects, especially focusing four areas of crime record analysis: hotspot analysis, offender profiling, victim profiling and predicting suspects. This chapter describes the selected approach for crime analysis.

4.2 Proposed Model

Basically, the development of a crime analysis tool has six steps, namely, data cleaning, integration, transformation, reduction, mining and visualizing as shown in Figure 4.1. Data preprocessing is a very important stage in the process of data mining since the results are considerably affected by outliers, noisy data and missing values. Thus, the outliers has to be detected and eliminated using suitable algorithms because those data may reduce the quality of data clustering and classification and, consequently, reduce the accuracy of prediction and increase the percentage of incorrectly classification instances. Moreover, data from different sources has to be integrated together in to a unified schema. Different inconsistencies have to be handled during this phase. Data has to be transformed in to other representations in order to improve the use of different algorithms in the analysis process. E.g.: decision tree algorithm only work with categorical algorithms. Apriory algorithms only work with boolean attributes. Due to these reasons, nominal data has to be transformed into categorical data or boolean data.

Since the original data set has a large set of attributes, but only few of them are relevant next important phase is reducing the data set by attribute subset selection.

Splitting of a set of data or objects in to a number of clusters based on the values of the different attributes of different entities is called clustering. Thereby, a cluster is a

collection of a set of similar data which behave same as a group. It can be said that the clustering is somewhat equal to the classification technique, with the only difference that the classes are not defined and determined in advance, and grouping of the data is done without supervision.

Simple k-means is the simplest and most frequently used clustering algorithm among other clustering algorithms in data mining. The reason behind this acceptance of the simple k-means is mainly the simplicity of the process. This algorithm is also suitable for smaller datasets and, as well as, for larger datasets since it has much less computational complexity, though this complexity grows linearly with the increasing volume of records. Beside simplicity of this technique, it however suffers from some disadvantages such as determination of the number of clusters by user, affectability from outlier data, high-dimensional data, and sensitivity toward centers for initial clusters and thus possibility of being trapped into local minimum may reduce efficiency of the K-means algorithm [11].

Classification is another important feature in the data mining process as a technique which can be used for modeling for forecasting. On the other hand, classification process divides the records in the dataset in to some groups that can act either dependently or independently. Also, this process can be used to make some examples of hidden and future decisions on the basis of the model generated using historical data. Decision tree learning, neural network, nearest neighborhood, Naive Bayes method and support vector machine are different algorithms which are used for the purpose of classification. [12]

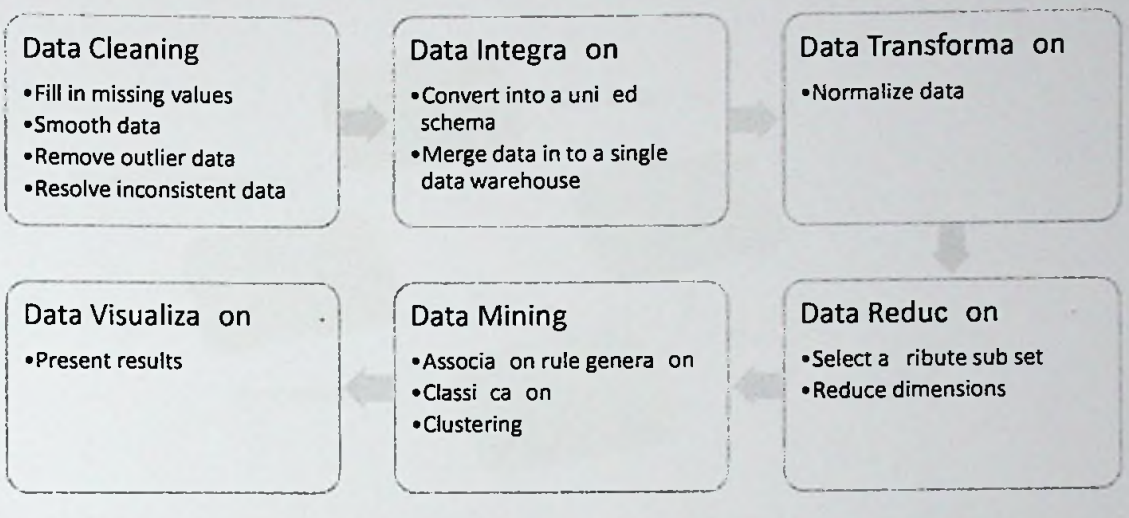


Figure 4.1 Proposed Model

4.3 System Overview

Moreover, this can be summarized into the following model given in Figure 4.2. The first two steps collect and analyze crime data, and offender data related to individual crime incidents in the history in order to predict future crimes. Data from different sources in the community may be required additionally. Efforts to combine these data are often far from easy, however. The third step is to inform police authorities about the knowledge and information gained from the analysis, so that, police operations can be adjusted to intervene against the predicted crime (or help solve past crimes). The type and the amount of this intervention will be different with the situation and the area charged with intervening. The types of interventions can be identified as generic intervention, crime-specific intervention, and problem-specific intervention. Thus, using the information provided, the need for situational awareness among officers and staff has to be established and this is a crucial aspect of any intervention plan. These interventions lead to a change in the patterns of criminal and ideally reduce or solve crime (the fourth step).

Figure 4.2 Summarized Model

4.4 Data Collection

Both the quantity and the quality of the data collected will determine the quality of the results of any approach. The collected data sets need to be updated periodically to ensure that current data is taken for the analysis and this reflects the effects of interventions. The lack of strong analytics may result in law quality and less accurate outcomes even with the perfect data. Moreover, missing values and noise values in the dataset decrease the quality of the result. And, some important information may be missing due to the misunderstanding and law level of sensitivity of the person who enters the data in to the system.

4.5 Data Integration

Mining techniques rely heavily on the data set not only on crimes but also on the environment in which the crimes took place. Most of the crime records will likely be collected by individual police stations, but information describing the environment may come from many other sources, such as, statistical department for data related to population, etc. in addition to that, free and commercial data sets may be available for use

with crime data. Analysts must be able to merge dissimilar information sources as part of the data collection and integration process. There are a number of methods for combining information, ranging from very simple techniques that offer an approximate picture.

4.6 Data Analysis

There are different types of analytic methods which can be used for predicting crime hotspots, identifying the patterns of offenders and victims and predicting offenders of novel crime incidents.

Using regression and data mining techniques, available data sets collected from different sources can be explored to provide some insight into the patterns of crimes that may be unique to a given region. The trends identified in this analysis can be used to inform the law enforcement authorities and to design an approach to identify hot spots. For example, these techniques can tell how far back to look for crime patterns or whether there are seasonal or weekly trends that should be included in the analysis. Global Positioning System (GPS) data can also be used in the process of data mining. Geographical profiles derived from clustering techniques on crime data related to geographical locations can reveal patterns indicating a serial criminal.

4.7 Making Predictions about Potential Crimes

Predictions about crimes and its victims and suspects to find when and where the next crime is most likely to happen, what is likely to be the motive for it, who is most likely to commit the crime (criminal) and who is most likely to be a victim. As with most forecasting methods, predicting future criminal events, whether from a tactical (next incident) or strategic (long-term) perspective, involves studying data on past crimes and victims, often using a variety of methods but generally always looking for patterns. "The past is prologue" is considered as the underlying assumption (days for tactical approaches, months to a few years for strategic approaches). Although this is true for addressing the questions of when, where, what, and who, the methods used will differ with the context and the goal. Hot spot analysis, statistical regression, data mining, and

near-repeat methods are generally used to identify where a crime will occur over a specified time horizon (when, varying from a day to a year, depending on the method and application) and therefore who is likely to be a victim and the suspect for the crime. Temporal and spatiotemporal methods discussed in the previous chapter can be used to identify when a crime is most likely to occur. Risk terrain analysis is appropriate for discerning the geospatial factors that create crime risk and looking for physical locations that might be ripe for a specific type of crime (where).

Police tend to operate in areas with high crime, but it would be a mistake to say that police cause high crime.

4.8 Predictive Analysis with Data Mining

Officially, data mining is the practice of analyzing any large amount of computerized data to investigate useful patterns and trends. Classification methods can be used to create models to predict a category or a class for an outcome (e.g., "There is an 85-percent chance of a robbery here next month"), rather than a continuous number, as in regression (e.g., "We predict an average of 1.24 robberies here next month").

Clustering methods cluster the records into different groups in which the records are "similar" mathematically. These models can be used to make predictions by stating that a future situation will likely be similar to a previous cluster of situations (e.g., "This neighborhood is showing attributes similar to those of other neighborhoods labeled as high-crime").

There are methods that allow for far more complicated (or at least very different) relationships between input data and output predictions than is normal in regression models. There are simple methods that select a relative handful of the possible variables and build a mathematical model with fairly simple relationships between the input data and the forecast. These models run comparatively quickly, and the results are usually directly interpretable by a person. Most regression analyses fit into this category, as do decision-tree methods.

There are three families of data mining techniques that will be of primary interest for predictive policing: regression, clustering, and classification.

4.9 Spatiotemporal Analysis

Features generally used in spatiotemporal analysis include time of day, day of week, and time and day cycles, temporal proximity to other events (e.g., payday, sporting events, concerts), season, weather, interval between offenses in a crime series (including correlations of those intervals to other factors, such as the value of stolen property), repeat locations, geographic progression of incidents in a crime series, spatial arrangement of incidents, type of location (e.g., parks, convenience stores, public housing), geographic correlates (e.g., near bus stops, near establishments licensed to sell liquor), environmental and target factors (e.g., lighting, neighborhood condition, traffic level), demographic and economic data from the crime area.

All of these features, alone or in combination, have predictive value in analyzing both short-term series and long-term problems or hot spots.

4.10 Users

The results of this analysis can be used by law enforcers to find general and specific crime trends, patterns, and series in an ongoing, timely manner in order to take advantage of the abundance of information existing in law enforcement agencies, the criminal justice system, and public domain, to maximize the use of limited law enforcement resources, to have an objective means to access crime problems locally, regionally, nationally within and between law enforcement agencies, to be proactive in detecting and preventing crime, to meet the law enforcement needs of a changing society and to understand the criminal behaviors.

4.11 Summary

We have discussed how we adopt the technology to solve the identified problem in this chapter. A major challenge in crime analysis is accurately and efficiently analyzing the

growing volumes of crime data. Crime data is not so precisely collected. Sometimes the analysis process needs to have assumptions about data. Cleaning of data to fill the missing values and inconsistencies is necessary to have a quality mining result. Preprocessing data consists of data cleaning, data integration and data transformation using a computer program. The intention is to reduce some noises, incomplete and inconsistent data. The results from preprocessing step can be later used by data mining algorithm. The methods that support clustering are best suited for the crime analysis of the poorly planned settings and is used to group data according to the different type of crime. From the clustered results it is easy to identify crime trends over years and can be used to design precaution methods for future. The classification of data is mainly used predict future crime trend. Outlier detection is mainly used to identify future crimes that are emerging newly. We discuss the analysis and design of our solution in the next chapter.

Chapter 5

Analysis and Design of the Proposed Solution

5.1 Introduction

In the previous chapter, we briefly discussed the technology used in our approach to solve the identified problem. This chapter describes the system design which includes four subsystems of crime record analysis. Moreover, this describes individual modules and the technologies used.

5.2 System Design

We have identified four major modules in the research which can be listed as follows.

1. Predicting crimes/Hot spot analysis
2. Predicting offenders/Offender profiling
3. Predicting perpetrators' identities
4. Predicting victims of crimes/Victim profiling

System Design can be represented as in the Figure 5.1 below.

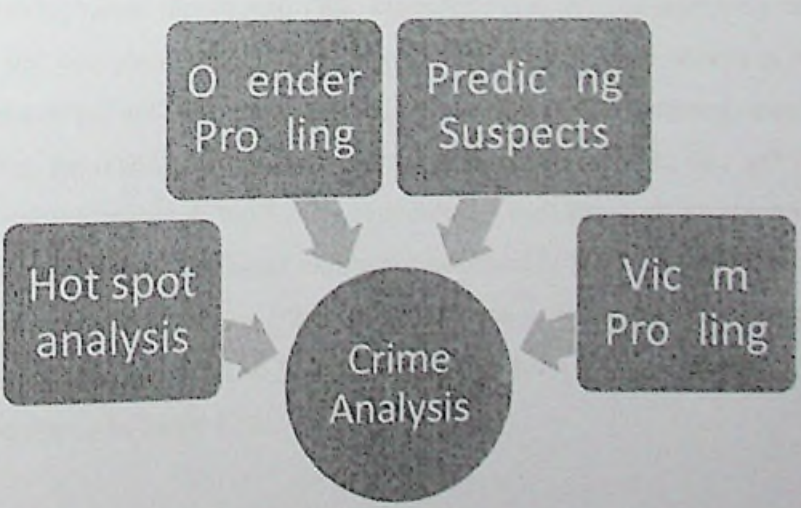


Figure 5.1 System Design of the Proposed Solution

5.2.1 Hot spot analysis

Hot spot methods predict areas of increased crime risk based on historical crime data. Hot spot methods seek to take advantage of the fact that crime is not uniformly distributed, identifying areas with the highest crime volumes or rates. The underlying assumption and prediction is that crime will likely occur where crime has already occurred: The past is prologue. Clustering with the attributes longitude and latitude data of each crime can be used to find the clusters or hotspots of crime incidents as described in the Implementation chapter. We selected k-means algorithm in order to find clusters since it has been used widely to identify crime hotspots. In addition to that, Microsoft Excel Power Maps can be used to generate thematic map. Only the longitude and latitude of crime incidents can be used to identify patterns or crime hotspots.

5.2.2 Offender profiling

The geographical approach looks at patterns in the location and timing of offences to make judgments about links between crimes and suggestions about where offenders live and work. Clustering techniques are used to identify clusters of suspects which shares same characteristics. Since it is widely used, we have selected k-means algorithm. It has been identified that important variables for this analysis are location of entry (walls, roof, window, etc), method of entry (Climbing, drilling, destroying, breaking, tunnel, etc.), type of dwelling house (apartment, villa, bungalow, etc), type of searching tidy, untidy, all rooms, just one place, etc, location of exit (Walls, roof, window, etc.), methods of offender interaction with the environment (Lock the door after entering, manipulate the alarm, killing the watchdog, etc.) All suspect related information, i.e., age of suspect, gender of suspect, approach for the crime incident which has been numbered from 1 to 10, where the meanings of different numbers mentioned in Table 5.1, marital status of the suspect, where the value can be either M for married or S for unmarried/single and the education level of suspects which has been numbered from 1 to 11(meanings of different numbers are shown in Table 5.2).

Table 5.1 The Meaning of Different Approach for the Crime

Approach level	Meaning
1	Kidnapped by outsider
2	Kidnapped by boyfriend of the victim
3	Deceived by boyfriend
4	On the way back home while walking
5	At home y a servant of the house
6	At home from the main door
7	At home by breaking a wall/window/door
8	At home by father/brother
9	At home by a relative
10	At a relative's house

Table 5.2 The Meanings of Different Education Levels

Education Level	Meaning
1	No schooling
2	Primary school
3	Secondary school up to grade 7
4	Secondary school up to grade 10
5	Sat for G. C. E. Advanced Level but not passed
6	G. C. E. ordinary Level passed but no A/L
7	Sat for G. C. E. Advanced Level but not passed
8	G. C. E. Advanced Level passed
9	Other higher education diploma
10	Completed bachelors degree
11	Completed post graduate degree

In addition to the suspect's details, we can use victim's details such as age, gender and education level in order to identify similar characteristics. Moreover, we need to have

longitude and latitude of the crime location in order to identify geological patterns in criminals.

5.2.3 Victim profiling

Most of the time the victim is identifiable and in most cases is the person reporting the crime. We can use k-means clustering technique here, as it is one of the most widely used data mining clustering technique. This module can be used to identify groups likely to be victims of various types of crime (vulnerable populations) and to identify people directly affected by at-risk locations. k-means clustering algorithm can be used to identify different clusters of victims who shares same similar characteristics. For this purpose, we have to use the details related to victims, such as, age, gender and level of education. The level of victim's education is categorized in to eleven levels as discussed in Table 5.2.

5.2.4 Predicting suspects

This module can be used to predict primary suspects to speed up the investigation process. Clearly, examining individual offenders based on their known modus operandi (MO) to crimes that have been committed creates reasonable leads for investigators to follow when following up on past incidents. Using this same information on offenders, analysts can project future behavior. Unfortunately, our data does not have the details on individual suspects or identities of suspects of individual crimes. But, having clustered suspects, we can predict the cluster of the suspect for a recent crime where the suspect is unidentified, so that, the investigation can begin with the suspects in the predicted suspect cluster. Considering the details of victims and suspects with longitude and latitude data of the place where the crime has occurred, different clusters of suspects can identified. Using that cluster number as the class label and the victim's details and geographical data as other attributes we can generate a classifier model with J48 algorithm. This model can be used to predict the category of the suspects, and later crime investigation can begin with the suspects in the identified cluster.

5.3 Summary

This chapter explained the design of the crime record analysis using datamining techniques. In this research some of the most significant capabilities of data mining techniques were leveraged through a multi-purpose framework for intelligent crime investigation. Next chapter gives details on the implementation of the design.

Chapter 6

Implementation of the Solution

6.1 Introduction

This chapter provides implementation details of each of the four modules mentioned in the previous chapter. Moreover, this presents software and algorithms used in each module with sample outputs.

6.2 Weka

Weka tool show in Figure 4.3 below is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java code. We have selected Weka since it contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes [13].

The weka.filters package is concerned with classes that transform datasets by removing or adding attributes, resampling the dataset, removing examples and so on. This package offers useful support for data preprocessing, which is an important step in machine learning.

6.3 Data Collection and Preprocessing

Data for the analysis is collected from Department of Police, Sri Lanka. A sample data set is attached in Appendix A. collected data had lots of missing values and noisy values. So, we followed the process shown in Figure 6.1 to preprocess data. Fill in missing values is done with weka using the filter ReplaceMissingValues which replaces all missing values for nominal and numeric attributes in a dataset with the modes and means from the training data. The class is ignored while applying it. Then we applied AttributeSelection filter to select only necessary attributes with the evaluator InfoGainAttributeEval which evaluates the worth of an attribute by measuring the

information gain with respect to the class and BestFirst search method which searches the space of attribute subsets by greedy hillclimbing augmented with a backtracking facility. Then all numeric values are converted in to binary values using NominalToBinary filter. Some screen shots which shows how we applied the filters using weka is included in Appendix A. A sample of the preprocessed data set is shown in Figure 6.2.

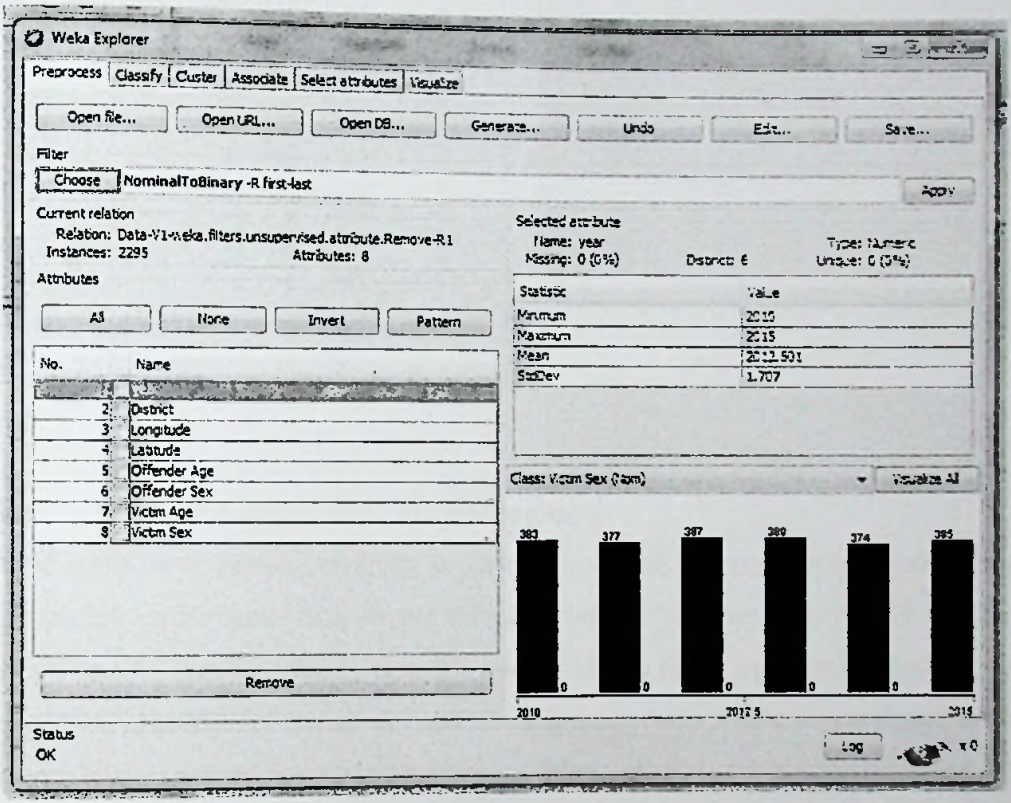


Figure 6.1 WEKA GUI

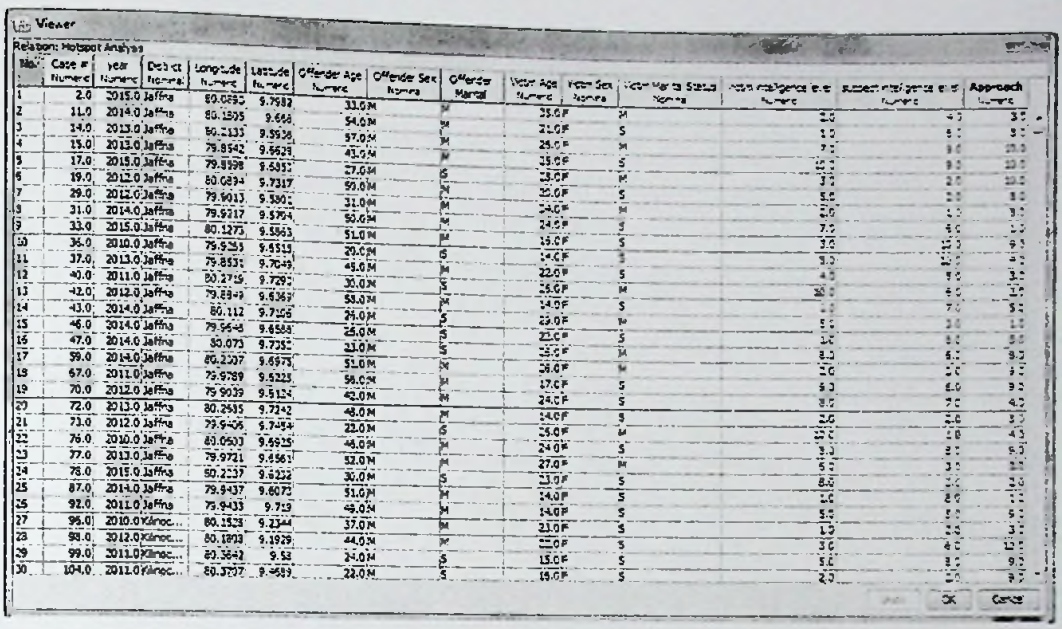


Figure 6.2 Sample of Preprocessed Dataset

6.4 Hot Spot Analysis and Crime Mapping

Grid mapping or thematic mapping, are commonly used, but these methods can be highly dependent on the initial data set and the partitioning of the map. Microsoft Excel Power Maps can be used to generate thematic map as shown in the figure 6.3 below. Hotspots could be identified as shown in Figure 6.4. Similar maps might be generated to show crime levels for particular districts, beats, or other jurisdictional boundaries, crime levels by month, crime levels during particular holidays, crime levels during special events (e.g., sporting events, major expos) and crime levels by weather conditions (assuming the data include a field for weather conditions in that jurisdiction at that time).

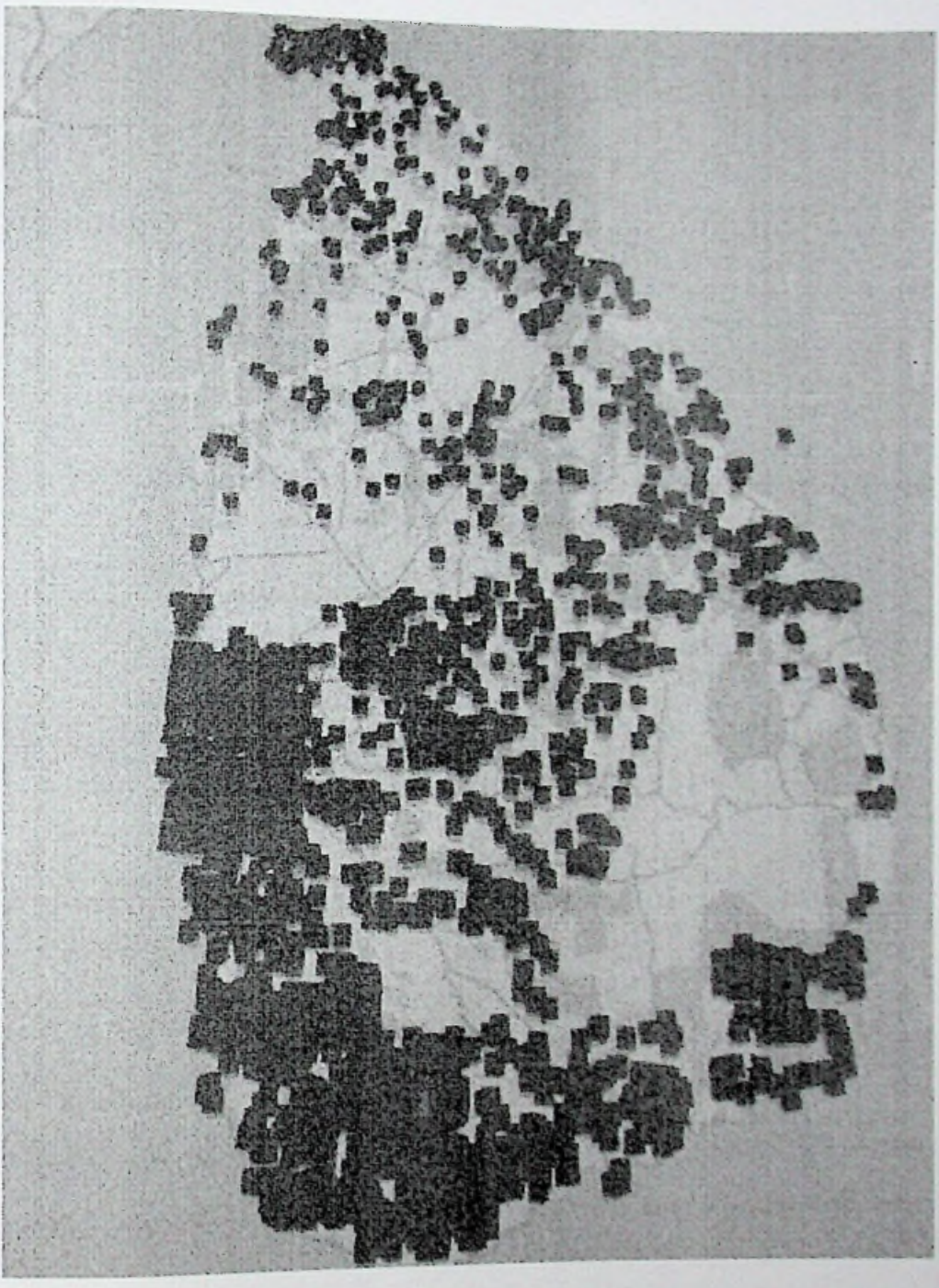


Figure 6.3 Hot spot Analysis

Figure 6.4 Identified Hotspots

K-nearest neighbor algorithm can be used to identify the clusters as shown in the Figure 6.5. This algorithm can use either the Euclidean distance (default) or the Manhattan distance. If the Manhattan distance is used, then centroids are computed as the component-wise median rather than mean. We have chosen the Euclidean distance function to use for instances comparison.

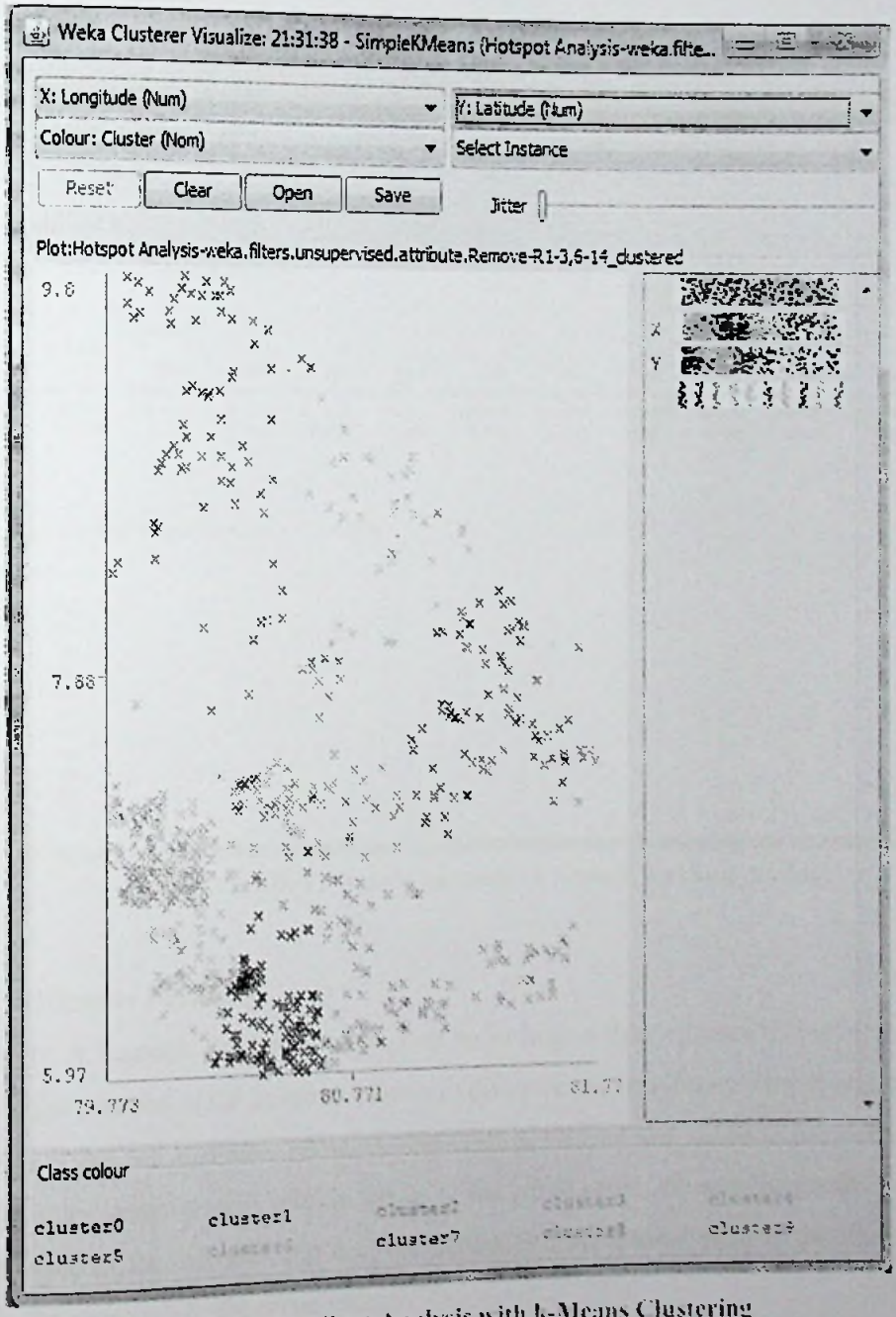


Figure 6. 5 Hot Spot Analysis with k-Means Clustering

The dataset is divided in to two segments where two third is taken for training the model and one third is used for testing the model generated. 10 major clusters could be identified. While the algorithm is processing, missing values are globally replaced with mean/mode. Figure 6.6 shows 10 major clusters and their centroids in Weka graphical user interface.

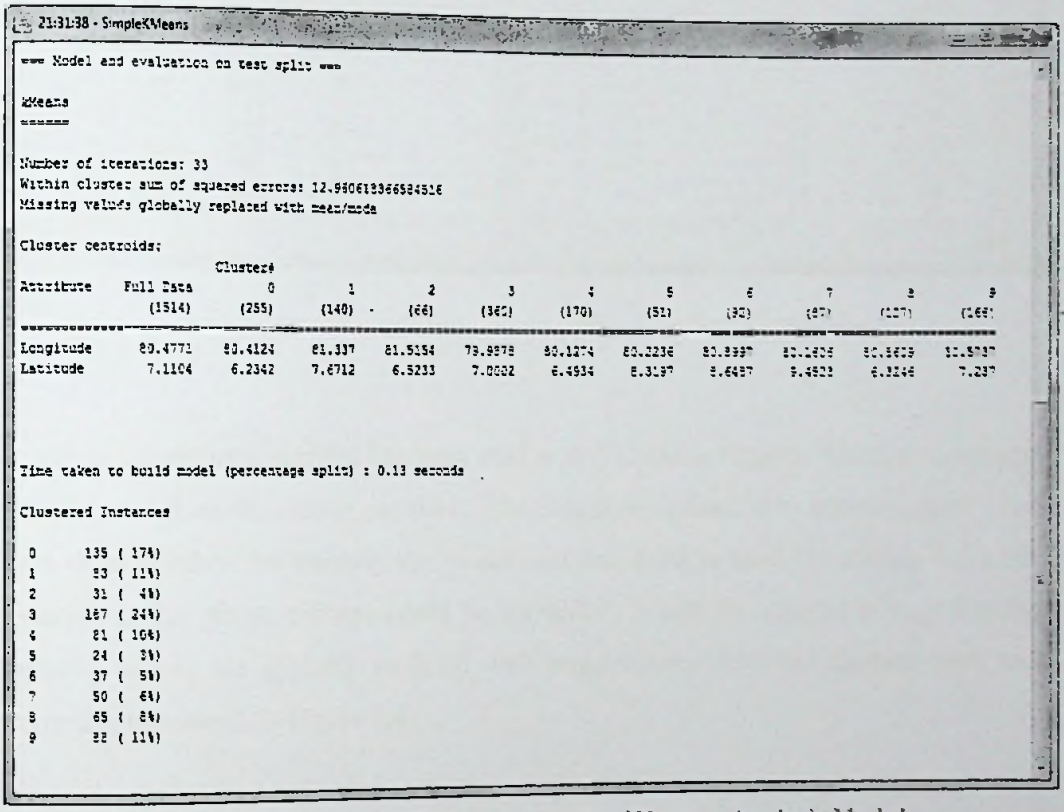


Figure 6. 6 Identified Cluster Centroids of Hotspot Analysis Module

6.5 Offender Profiling

Offenders or suspects has been categorized according to their characteristics in order to identify the patterns of the suspects belong to different clusters. Sample input is shown in Figure 6.7. All the attributes related to the crimes, victims and suspects has been used. This includes longitude and latitude values of the crime scene, the age, the gender and the marital state of the offender, the age, the gender and the marital state of the victim and finally, the approach to the crime scene.

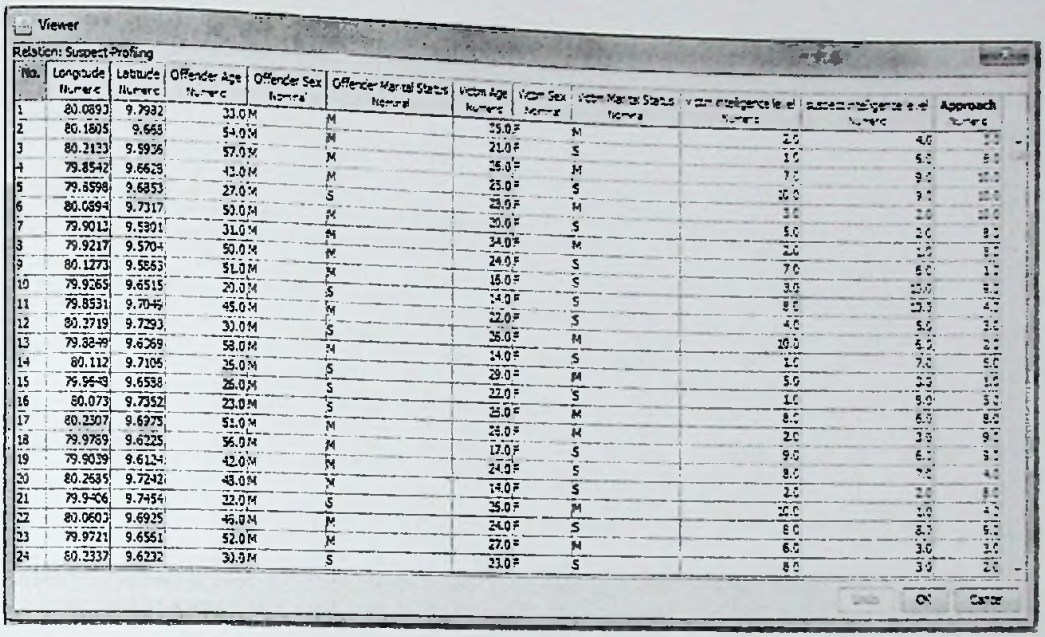
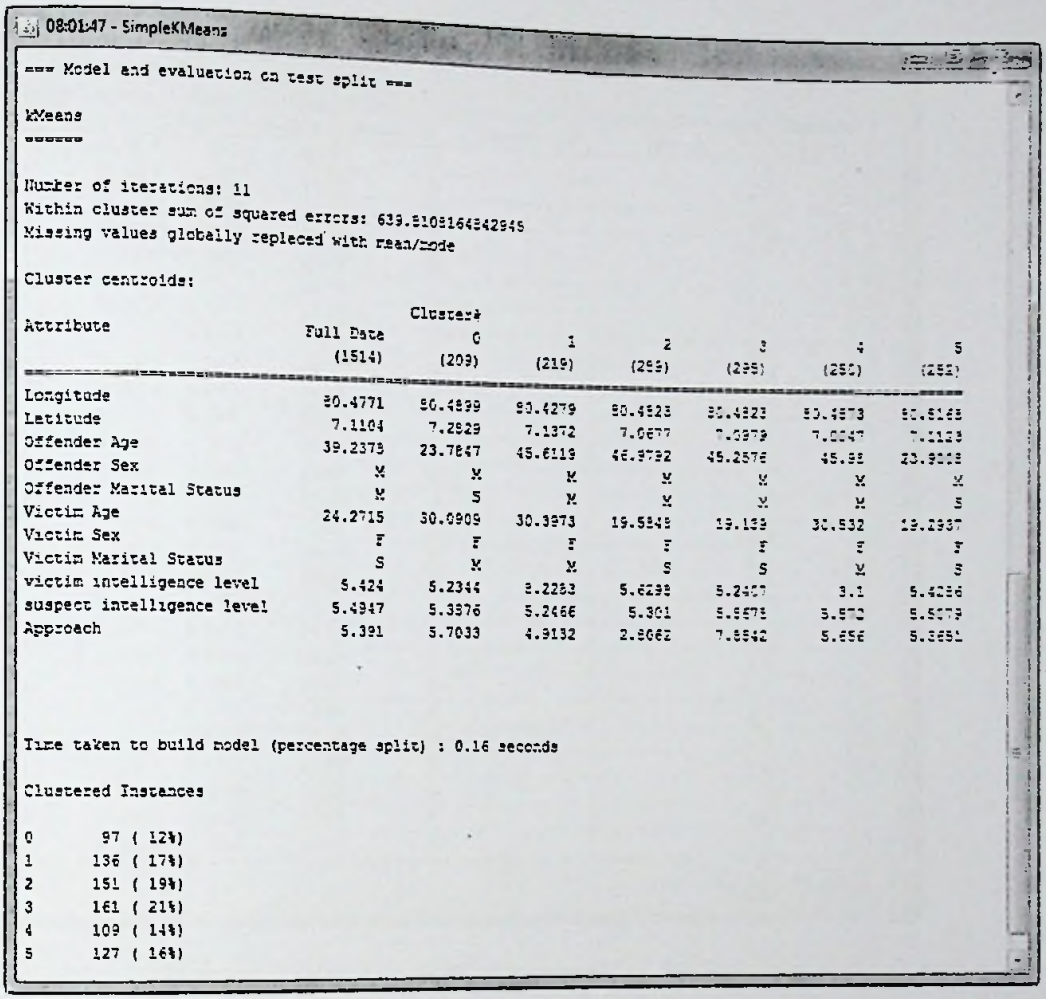


Figure 6. 7 Sample Input to the Offender Profiling Module

k-means clustering algorithm has been used with Euclidean distance function to measure the distance from the cluster centroids. The dataset is divided in to two segments where two third is taken for training the model and one third is used for testing the model generated. Six major clusters could be identified. While the algorithm is processing, missing values are globally replaced with mean/mode. Identified clusters with their centroids are shown in Figure 6.8.



08:01:47 - SimpleKMeans

```
=== Model and evaluation on test split ===

kMeans
======

Number of iterations: 11
Within cluster sum of squared errors: 639.8108164842845
Missing values globally replaced with mean/mode

Cluster centroids:
```

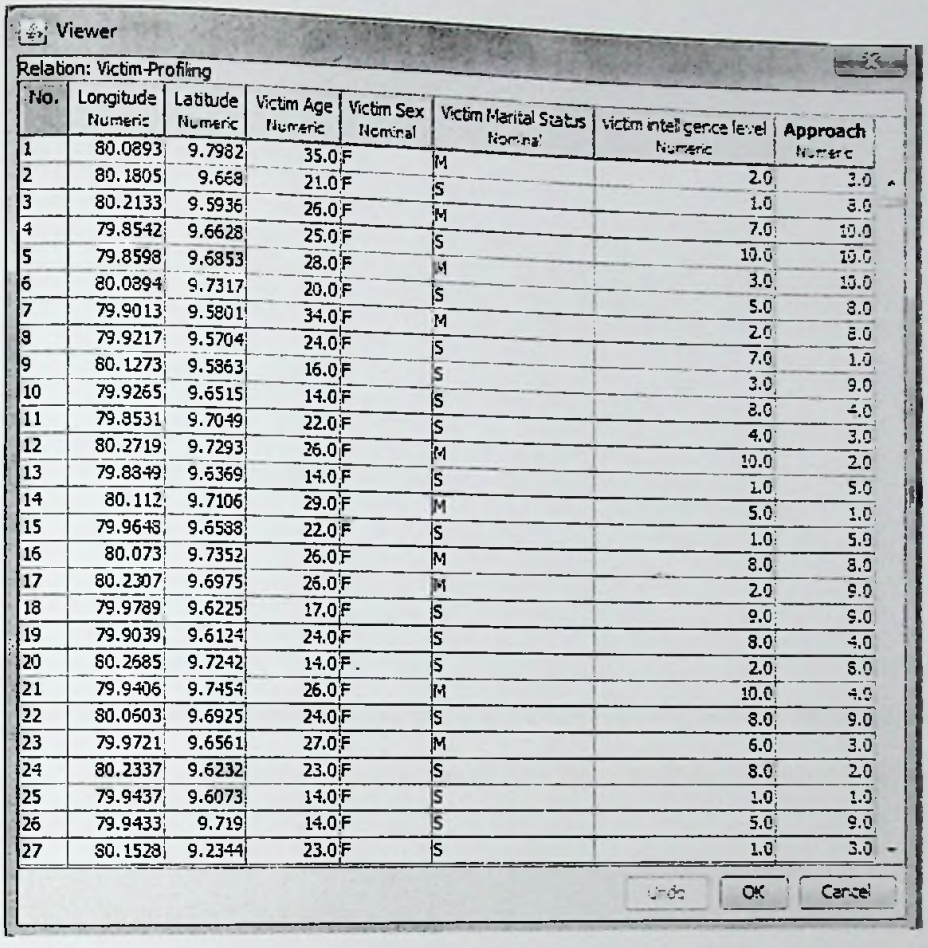
Attribute	Full Data (1514)	Cluster# 0 (209)	1 (219)	2 (259)	3 (295)	4 (250)	5 (252)
Longitude	80.4771	80.4899	80.4279	80.4823	80.4823	80.4573	80.5165
Latitude	7.1104	7.2529	7.1372	7.0677	7.0979	7.0047	7.1129
Offender Age	39.2379	23.7847	45.6119	46.9792	45.2574	45.88	23.9008
Offender Sex	M	M	M	M	M	M	M
Offender Marital Status	M	S	M	M	M	M	S
Victim Age	24.2715	30.0909	30.3973	19.5849	19.139	30.532	19.2937
Victim Sex	F	F	F	F	F	F	F
Victim Marital Status	S	M	M	S	S	M	S
victim intelligence level	5.424	5.2344	5.2283	5.6293	5.2407	3.1	5.4286
suspect intelligence level	5.4947	5.3876	5.2466	5.301	5.5575	5.572	5.5079
Approach	5.391	5.7033	4.9132	2.8062	7.8542	5.656	5.3651

```
Time taken to build model (percentage split) : 0.16 seconds

Clustered Instances

0       97 ( 12%)
1      136 ( 17%)
2      151 ( 19%)
3      161 ( 21%)
4      109 ( 14%)
5      127 ( 16%)
```

Figure 6. 8 Identified Cluster Centroids in Offender Profiling Module

6.6 Victim Profiling

Using the same methodology used in offender profiling, victims has been categorized according to their characteristics in order to identify the most vulnerable victims into clusters. Sample input is shown in Figure 6.9. All the attributes related to the victims has been used in the input. This includes longitude and latitude values of the crime scene, the age, the gender and the marital state of the victim and finally, the approach to the crime scene.



Viewer

Relation: Victim-Profiling

No.	Longitude Numeric	Latitude Numeric	Victim Age Numeric	Victim Sex Nominal	Victim Marital Status Nominal	victim intelligence level Numeric	Approach Numeric
1	80.0893	9.7982	35.0	F	M	2.0	3.0
2	80.1805	9.668	21.0	F	S	1.0	8.0
3	80.2133	9.5936	26.0	F	M	7.0	10.0
4	79.8542	9.6628	25.0	F	S	10.0	10.0
5	79.8598	9.6853	28.0	F	M	3.0	10.0
6	80.0894	9.7317	20.0	F	S	5.0	8.0
7	79.9013	9.5801	34.0	F	M	2.0	8.0
8	79.9217	9.5704	24.0	F	S	7.0	1.0
9	80.1273	9.5863	16.0	F	S	3.0	9.0
10	79.9265	9.6515	14.0	F	S	8.0	4.0
11	79.8531	9.7049	22.0	F	S	4.0	3.0
12	80.2719	9.7293	26.0	F	M	10.0	2.0
13	79.8849	9.6369	14.0	F	S	1.0	5.0
14	80.112	9.7106	29.0	F	M	5.0	1.0
15	79.9648	9.6588	22.0	F	S	1.0	5.0
16	80.073	9.7352	26.0	F	M	8.0	8.0
17	80.2307	9.6975	26.0	F	M	2.0	9.0
18	79.9789	9.6225	17.0	F	S	9.0	9.0
19	79.9039	9.6124	24.0	F	S	8.0	4.0
20	80.2685	9.7242	14.0	F	S	2.0	8.0
21	79.9406	9.7454	26.0	F	M	10.0	4.0
22	80.0603	9.6925	24.0	F	S	8.0	9.0
23	79.9721	9.6561	27.0	F	M	6.0	3.0
24	80.2337	9.6232	23.0	F	S	8.0	2.0
25	79.9437	9.6073	14.0	F	S	1.0	1.0
26	79.9433	9.719	14.0	F	S	5.0	9.0
27	80.1528	9.2344	23.0	F	S	1.0	3.0

Undo OK Cancel

Figure 6. 9 Sample Dataset for the Victim Profiling Module

k-means clustering algorithm has been used with Euclidean distance function to measure the distance from the cluster centroids. The dataset is divided in to two segments where two third is taken for training the model and one third is used for testing the model generated. Ten major clusters could be identified. While the algorithm is processing, missing values are globally replaced with mean/mode. Identified clusters with their centroids are shown in Figure 6.10.

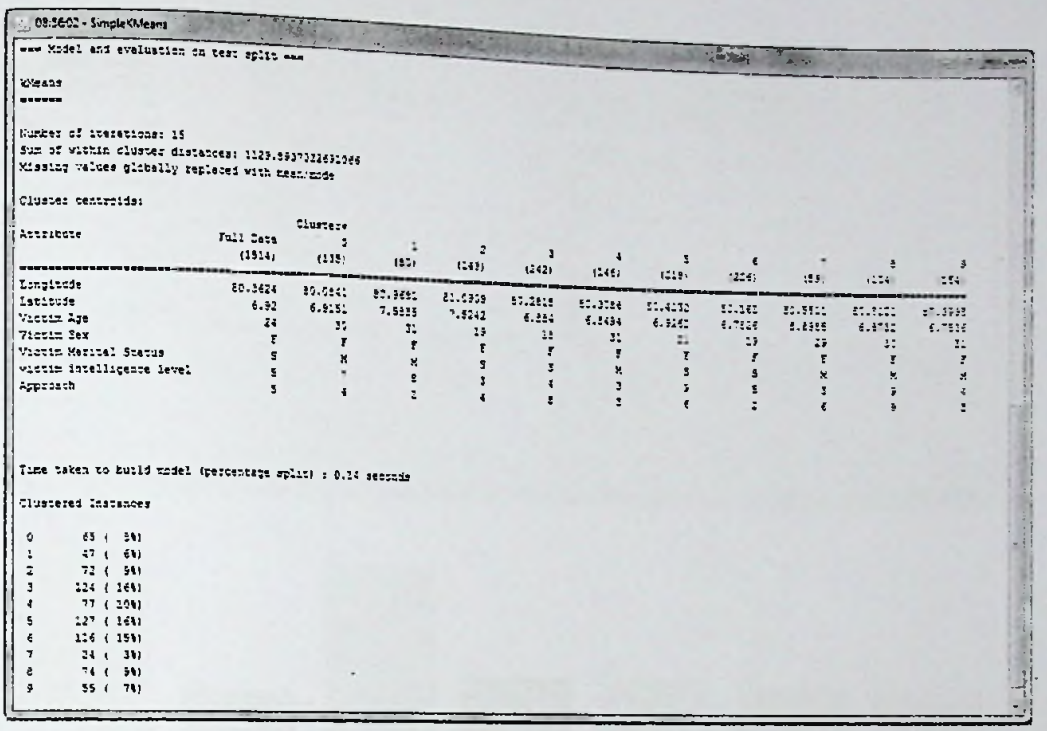


Figure 6.10 Identified Clusters and Their Centroids in the Victim Profiling Module

6.7 Predicting Suspects

All the suspects are clustered according to their characteristics including details of crime scene, details of victims and details of suspects. Seven major clusters has been identified from the algorithm AddCluster which is a preprocessing filter that adds a new nominal attribute representing the cluster assigned to each instance by the simple k-means clustering algorithm. Moreover, Euclidean function has been used with the clustering algorithm. The summary of the resulting attribute *cluster* is shown in Figure 6.11.

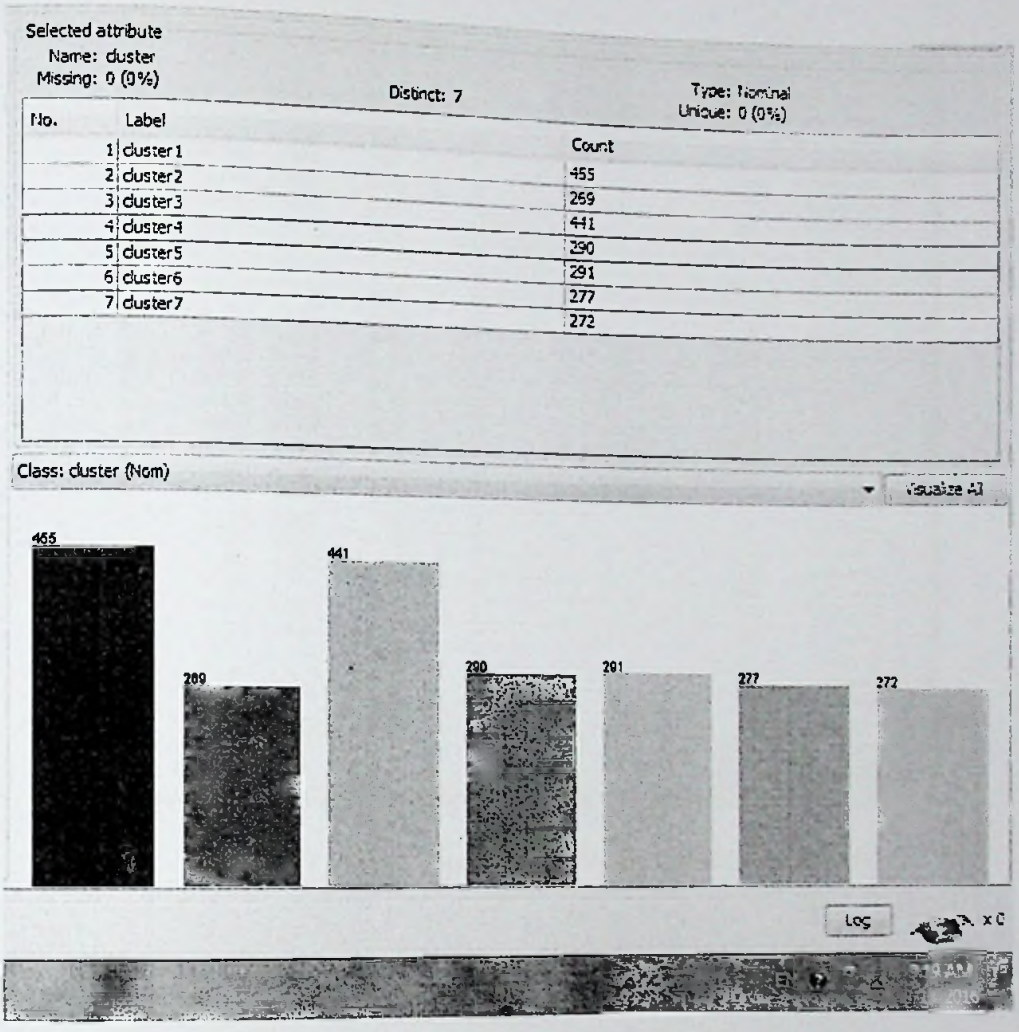
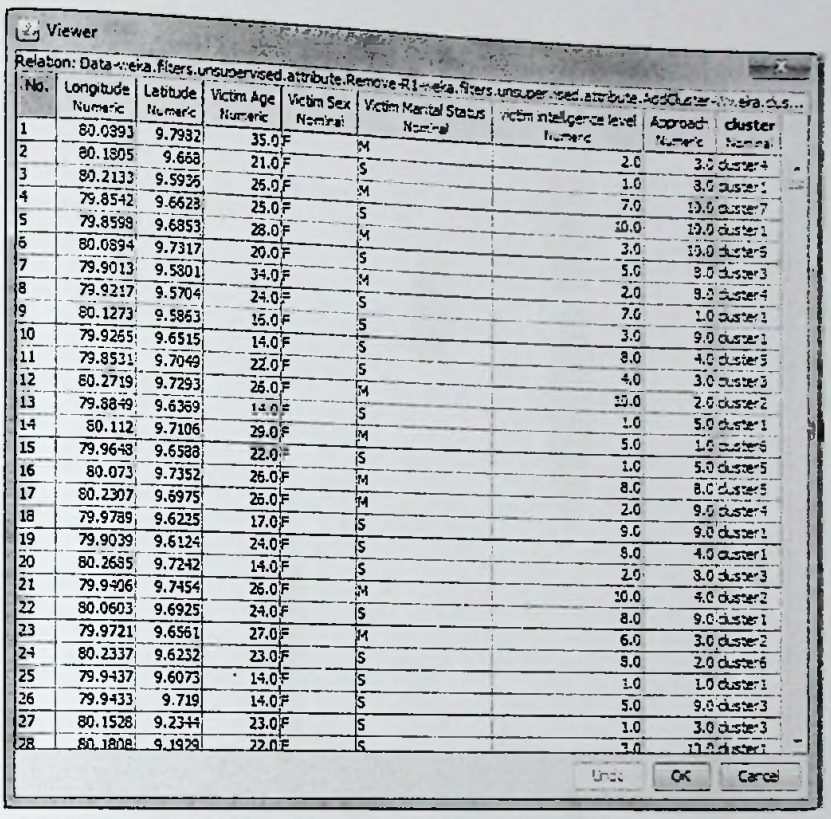


Figure 6.11 Summary of the new attribute 'cluster' in the dataset

Then, the details of the suspects including the age, the gender and the marital status of the offender has been removed from the dataset to prepare it for the classification. Sample dataset used for the suspect prediction module is shown in Figure 6.12.

J48 algorithm which is used to generate decision tree with the confidence factor 0.3 is used to generate the classifier model which can be used to predict the cluster of the suspect once a crime occurred. Resulting classifier is shown in Figure 6.13 as a decision tree. The dataset is divided in to two segments where two third is taken for training the model and one third is used for testing the model generated. The result of evaluation of the model on test split method is shown in Figure 6.14.



Viewer

Relation: Data-weka.filters.unsupervised.attribute.Remove-R1-weka.filters.unsupervised.attribute.AddCluster-Weka.clus...

No.	Longitude Numeric	Latitude Numeric	Victim Age Numeric	Victim Sex Nominal	Victim Marital Status Nominal	victim intelligence level Numeric	Approach Numeric	cluster Nominal
1	80.0393	9.7932	35.0	F	M	2.0	3.0	cluster4
2	80.1805	9.668	21.0	F	S	1.0	8.0	cluster1
3	80.2133	9.5926	26.0	F	M	7.0	10.0	cluster7
4	79.8542	9.6628	25.0	F	S	10.0	10.0	cluster1
5	79.8598	9.6853	28.0	F	M	3.0	10.0	cluster5
6	80.0894	9.7317	20.0	F	S	5.0	8.0	cluster3
7	79.9013	9.5801	34.0	F	M	2.0	9.0	cluster4
8	79.9217	9.5704	24.0	F	S	7.0	1.0	cluster1
9	80.1273	9.5863	15.0	F	S	3.0	9.0	cluster1
10	79.9265	9.6515	14.0	F	S	8.0	4.0	cluster5
11	79.8531	9.7049	22.0	F	S	4.0	3.0	cluster3
12	80.2719	9.7293	26.0	F	M	10.0	2.0	cluster2
13	79.8849	9.6369	14.0	F	S	1.0	5.0	cluster1
14	80.112	9.7106	29.0	F	M	5.0	1.0	cluster6
15	79.9648	9.6588	22.0	F	S	1.0	5.0	cluster5
16	80.073	9.7352	26.0	F	M	8.0	8.0	cluster5
17	80.2307	9.6975	26.0	F	M	2.0	9.0	cluster4
18	79.9789	9.6225	17.0	F	S	9.0	9.0	cluster1
19	79.9039	9.6124	24.0	F	S	8.0	4.0	cluster1
20	80.2685	9.7242	14.0	F	S	2.0	8.0	cluster3
21	79.9406	9.7454	26.0	F	M	10.0	4.0	cluster2
22	80.0603	9.6925	24.0	F	S	8.0	9.0	cluster1
23	79.9721	9.6561	27.0	F	M	6.0	3.0	cluster2
24	80.2337	9.6232	23.0	F	S	8.0	2.0	cluster6
25	79.9437	9.6073	14.0	F	S	1.0	1.0	cluster1
26	79.9433	9.719	14.0	F	S	5.0	9.0	cluster3
27	80.1528	9.2344	23.0	F	S	1.0	3.0	cluster3
28	80.1808	9.1929	22.0	F	S	3.0	11.0	cluster1

Undo OK Cancel

Figure 6.12 Sample Dataset for the Suspect Prediction Module

Figure 6.13 Classification Model for the Suspect Prediction Module

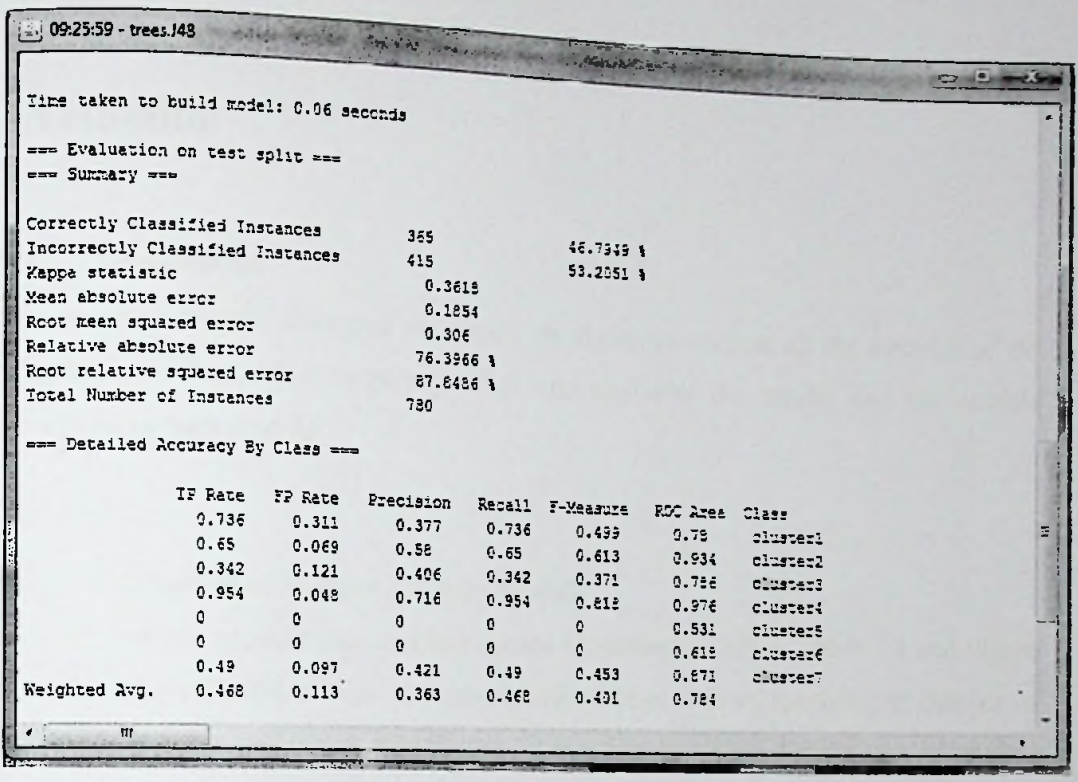


Figure 6.14 Evaluation of the Model for the Suspect Prediction Module

6.8 Summary

This section provided implementation details of each module of the proposed solution. Moreover, it mentioned software, algorithms, different parameters and resulting models of each module in the design. Next chapter evaluates all the modules implemented in the solution.

Chapter 7

Evaluation

7.1 Introduction

The previous chapter discussed the details on implementation of all the modules of the proposed solution. This chapter justifies and evaluates the results and the models generated in each module.

7.2 Evaluation of Hotspot Analysis Module

Hotspot analysis module uses simple k-means clustering algorithm. Table 7.1 and Figure 7.1 show the variation of sum of squared errors within clusters for different number of clusters. From the graph shown in Figure 7.1 it is clearly visible that the sum of squared errors within clusters gets in to a stable state after ten clusters. Hence, we have selected ten as the number of clusters of crime scenes to identify hotspots in this module.

Table 7. 1 Variation of Squared Errors within Clusters for Different number of clusters in Hotspot Analysis Module

Number of Clusters	Sum of Squared Errors Within Cluster
2	97.225
3	82.902
4	58.372
5	49.029
6	35.804
7	31.059
8	29.694
9	26.16
10	12.96
11	12.171
12	10.572

13	9.704
14	9.083
15	7.949

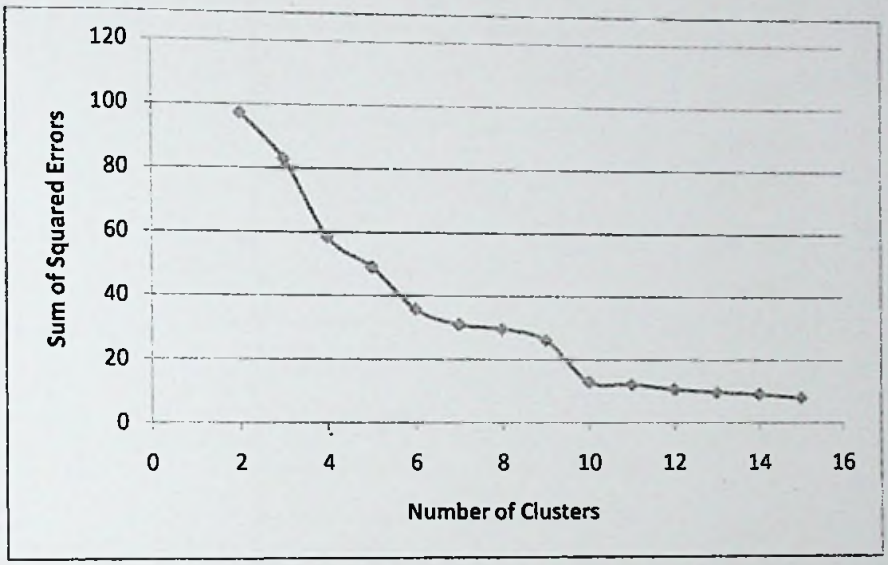


Figure 7.1 Variation of Squared Errors within Clusters for Different number of clusters in Hotspot Analysis Module

Moreover, Sum of squared clusters within cluster with Manhatton distance function is 160.009 where it is 12.961 with Euclidean distance function. Thus, we selected Euclidean distance function as the distance function to measure the distances of the objects with the cluster centroids.

7.3 Evaluation of Suspect/Offender Profiling Module

Suspect profiling module uses simple k-means algorithm to identify clusters of suspects according to their characteristics and patterns. Table 7.2 and Figure 7.2 show the variation of sum of squared errors within clusters for different number of clusters. From the graph shown in Figure 7.2 it is clearly visible that the sum of squared errors within clusters gets in to a stable state after six clusters. Hence, we have selected six as the number of clusters to identify clusters of suspects who share similar characteristics.

Table 7.2 Variation of Squared Errors within Clusters for Different number of clusters in Suspect Profiling Module

Number of Clusters	Sum of Squared Errors within Cluster
2	1507.637
3	963.983
4	952.709
5	870.616
6	639.811
7	619.589
8	603.981
9	580.79
10	564.222
11	535.865
12	522.671
13	502.434
14	496.207
15	482.46

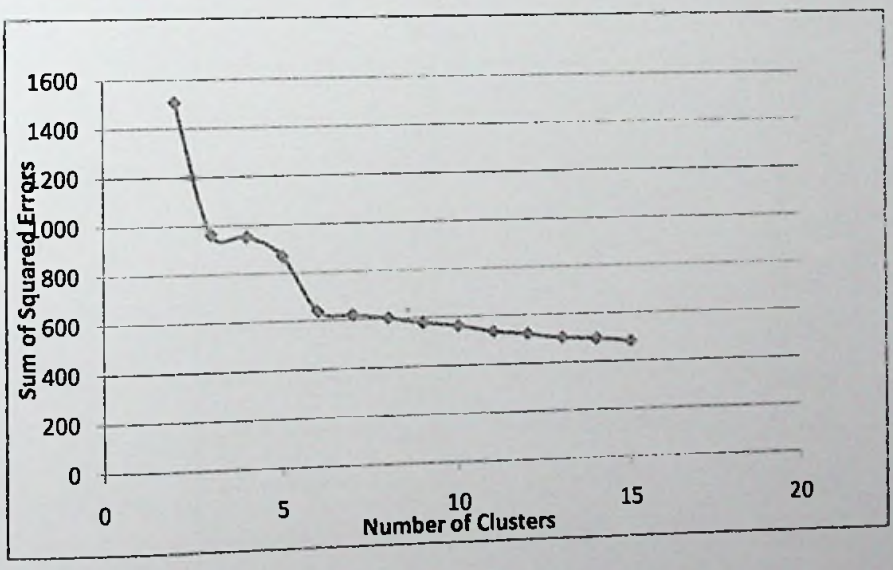


Figure 7.2 Variation of Squared Errors within Clusters for Different number of clusters in Suspect Profiling Module

Moreover, Sum of squared errors within cluster with Manhatton distance function is 3447.073 where it is 977.988 with Euclidean distance function. Thus, we selected Euclidean distance function as the distance function to measure the distances of the objects with the cluster centroids.

7.4 Evaluation of Victim Profiling Module

Victim profiling module uses simple k-means algorithm to identify clusters of victims to identify clusters of more vulnerable victims. Table 7.3 and Figure 7.3 show the variation of sum of squared errors within clusters for different number of clusters. From the graph shown in Figure 7.3 it is clearly visible that the sum of squared errors within clusters gets in to a stable state after ten clusters. Hence, we have selected ten as the number of clusters to identify clusters of victims who share similar characteristics in this module.

Table 7.3 Variation of Squared Errors within Clusters for Different number of clusters in Victim Profiling Module

Number of Clusters	within cluster sum of squared errors
2	529.197
3	477.092
4	410.611
5	375.377
6	333.918
7	314.103
8	298.573
9	289.122
10	273.884
11	256.283
12	245.035
13	224.319
14	219.789
15	210.74

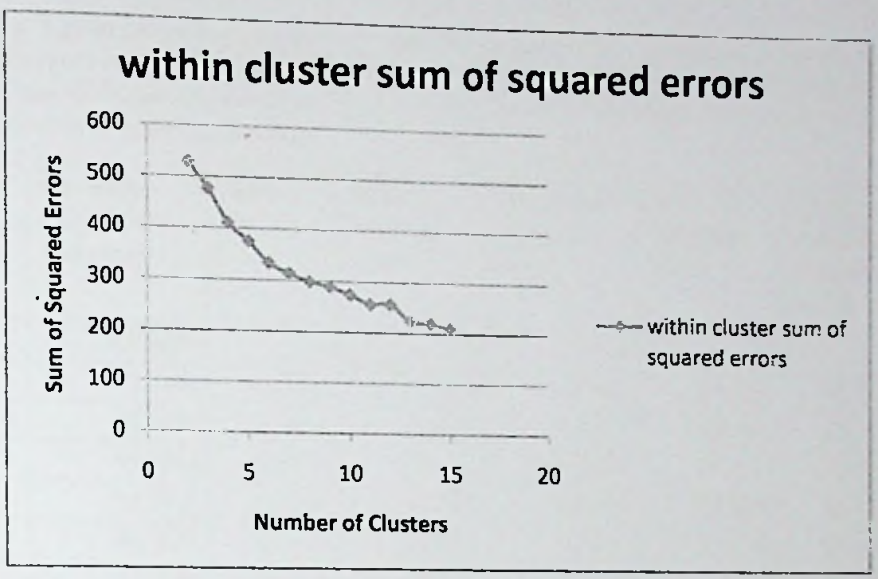


Figure 7. 3 Variation of Squared Errors within Clusters for Different number of clusters in Victim Profiling Module

Moreover, Sum of squared errors within cluster with Manhatton distance function is 1703.698 where it is 399.536 with Euclidean distance function. Thus, we selected Euclidean distance function as the distance function to measure the distances of the objects with the cluster centroids.

7.5 Evaluation of Suspect Prediction Module

Suspect prediction module uses simple k-means algorithm to identify clusters of suspects first as a preprocessing step. Table 7.4 and Figure 7.4 show the variation of percentage of correctly classified instances, Kappa statistic, mean absolute error, root mean squared error, relative absolute error and root relative squared error of the classifier model when different number of clusters for the suspect category is used. From the graphs shown in Figures 7.4 to 7.9, it is clearly visible that the correctly classified instances, Kappa statistic, mean absolute error, root mean squared error, relative absolute error and root relative squared error of the classifier model gets in to a stable state after seven clusters. Hence, we have selected seven as the number of clusters to categorize suspects in order to generate the model to predict the cluster of the suspect of a crime.

Table 7.4 Variation of percentage of correctly classified instances, Kappa statistic, mean absolute error, root mean squared error, relative absolute error and root relative squared error of the classifier model when different number of clusters for the suspect category is used in Suspect Prediction Module

Number of Clusters	Correctly Classified Instances	Kappa statistic	Mean absolute error	Root mean squared error	Relative absolute error	Root relative squared error
2	100.00%	1	0	0	0	0
3	71.28%	0.5607	0.1872	0.3059	43.28%	65.71%
4	71.28%	0.5607	0.1872	0.3059	43.28%	65.71%
5	61.03%	0.5095	0.1786	0.2944	55.77%	73.51%
6	51.79%	0.4087	0.1851	0.3032	66.88%	81.53%
7	44.23%	0.3322	0.1889	0.3084	77.82%	88.54%
8	48.97%	0.418	0.1627	0.2858	74.41%	86.41%
9	44.23%	0.3747	0.1468	0.2719	74.41%	86.52%
10	36.28%	0.2905	0.1364	0.2627	75.95%	87.64%
11	44.87%	0.3877	0.1197	0.245	72.77%	85.43%

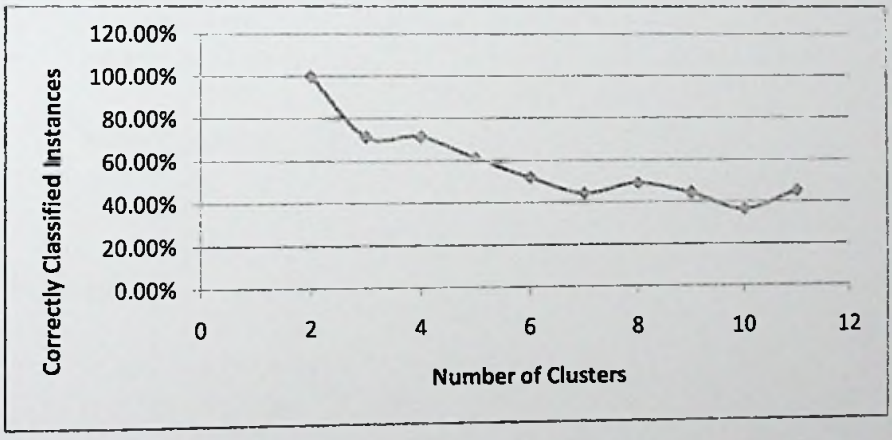


Figure 7.4 Variation of the Percentage of Correctly Classified Instances for Different number of suspect clusters in Suspect Prediction Module

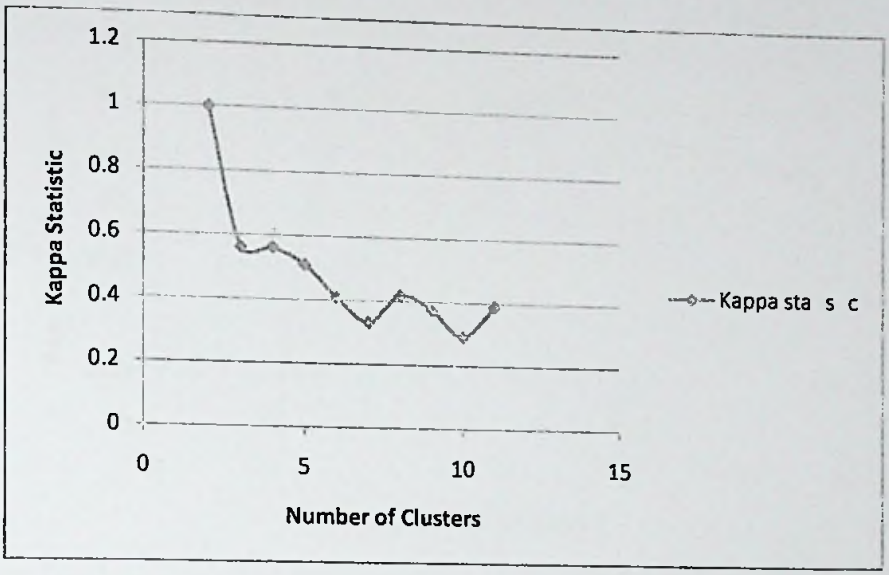


Figure 7.5 Variation of the Kappa Statistic for Different number of suspect clusters in Suspect Prediction Module

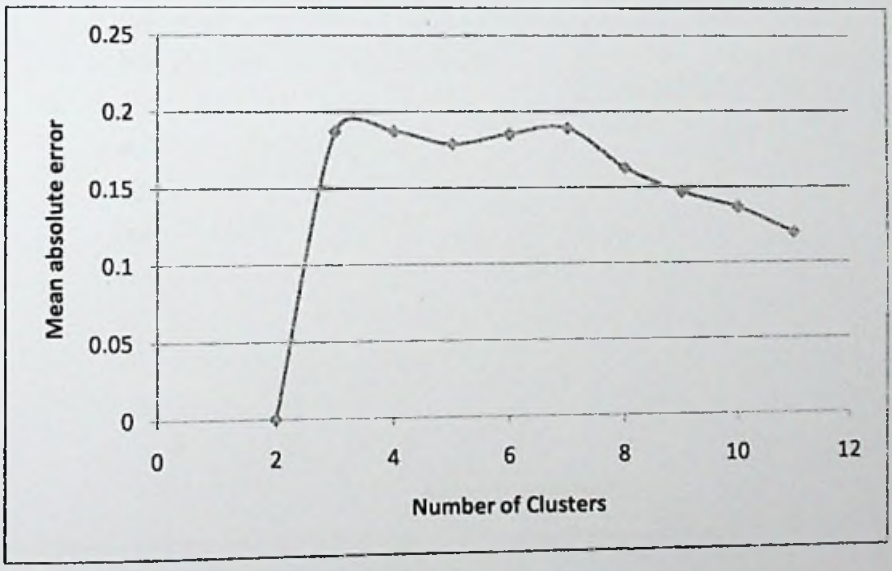


Figure 7.6 Variation of the Mean Absolute Error for Different number of suspect clusters in Suspect Prediction Module

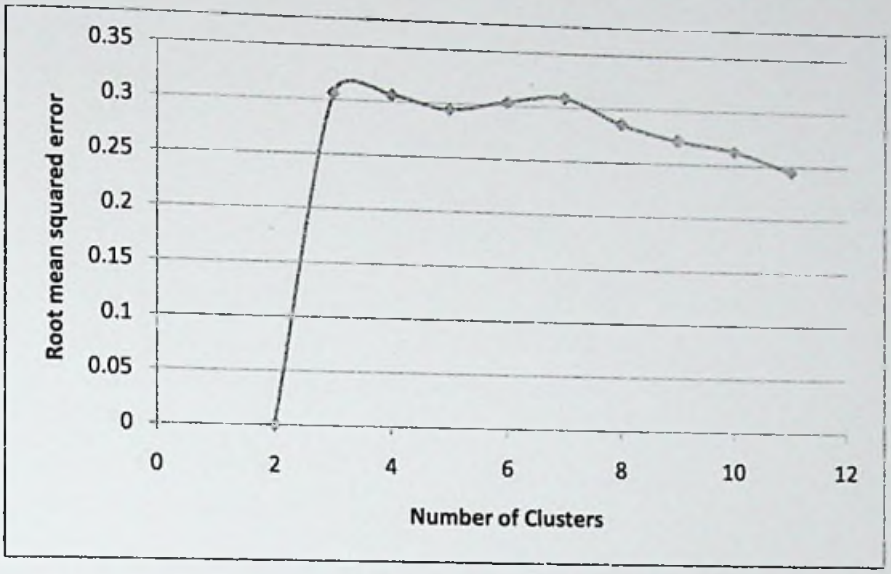


Figure 7.7 Variation of the Root Mean Squared Error for Different Number of Suspect Clusters in Suspect Prediction Module

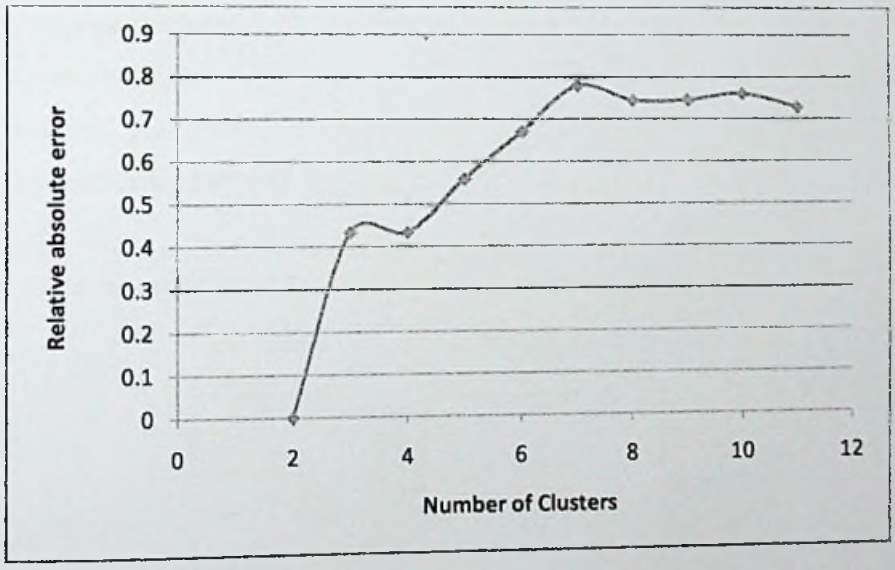


Figure 7.8 Variation of the Relative Absolute Error for Different Number of Suspect Clusters in Suspect Prediction Module

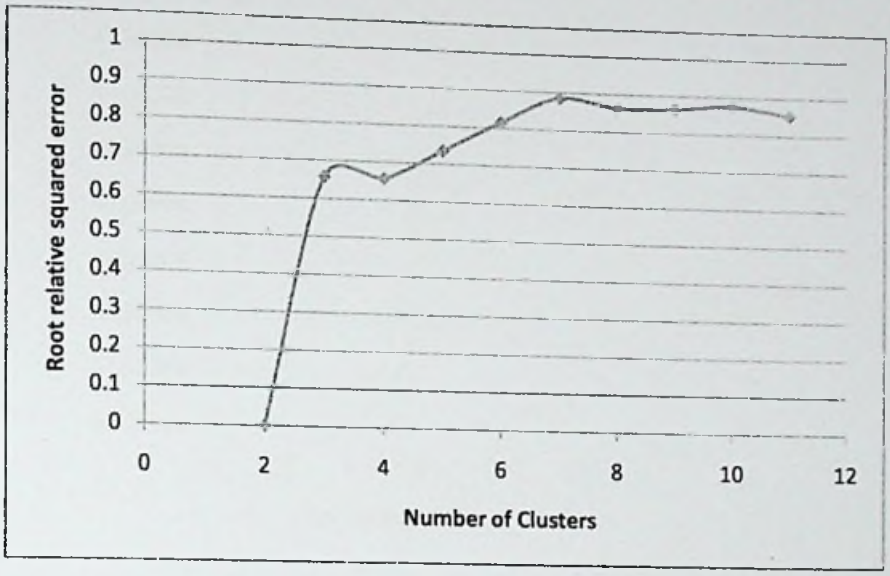


Figure 7.9 Variation of the Root Relative Squared Errors for Different Number of Suspect Clusters in Suspect Prediction Module

Moreover, Table 7.5 and Figure 7.10 show variation of the percentage of correctly classified instances with different confidence factors for J48 algorithm. From the graph shown in Figure 7.10, it is clearly visible that the percentage of correctly classified instances gets stabilized after the confidence factor value 3.0. Hence, we selected 0.3 as the confidence factor of our model.

Table 7.5 Variation of the Percentage of Correctly Classified Instances with Different Confidence Factors for J48 Algorithm in Suspect Predicting Module

Con dence Factor	Correctly Classi ed Instances
0.25	44.2308%
0.3	46.7949%
0.35	46.7949%
0.4	46.7949%
0.45	47.0513%

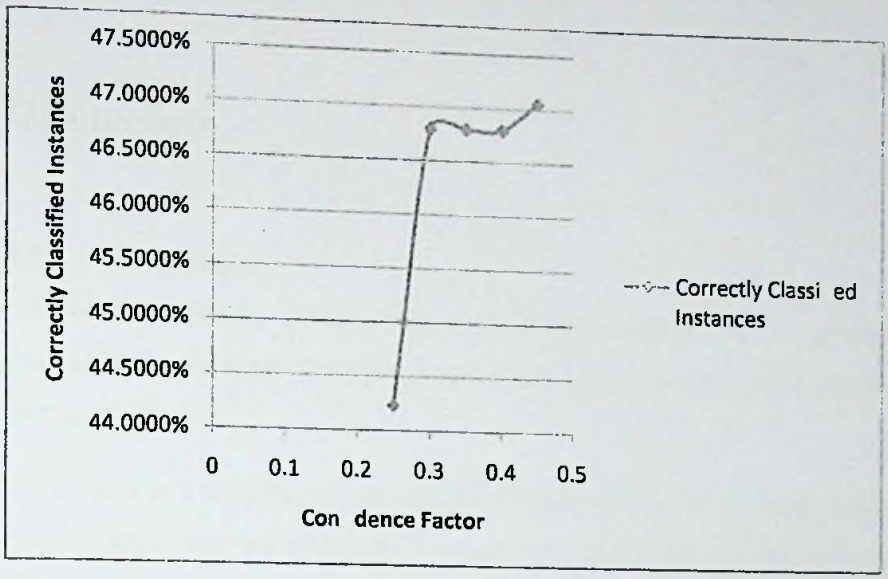


Figure 7.10 Variation of the Percentage of Correctly Classified Instances with Different Confidence Factors for J48 Algorithm in Suspect Predicting Module

7.6 Summary of Evaluation

This chapter evaluated the methodologies and the results discussed in the Implementation chapter. Next chapter discusses some limitations and future improvements for the proposed solution.

Chapter 8

Discussion

8.1 Introduction

All previous chapters discussed the problem identified and the proposed solution. This chapter discusses some limitations and future improvements which can be proposed for further work.

Crime data is a sensitive domain where efficient clustering techniques play vital role for crime analysts and law-enforcers to precede the case in the investigation and help solving unsolved crimes faster. Crime record analysis with data mining techniques is conducted by different researchers in two directions: Classification of Crime and Clustering Technique of Crime. Some researchers propose new methods to detect outliers by discovering frequent pattern from the data set. The outliers are defined as the data transactions that contain less frequent pattern in their item sets. It can be applied to law enforcement to discover unusual patterns from multiple actions of a criminal entity, especially fraud committed in financial transactions, trading activity or insurance claims.

Recent developments in crime control applications aim at adopting data mining techniques to aid the process of crime investigation. Raw data are being preprocessed before mining because data are in different format, collected from various sources and stored in the data bases and data warehouses. Similarity measures are an important factor which helps to find unsolved crimes in crime pattern. Partition clustering algorithm is one of the best method for finding similarity measures. K-means algorithm mainly used to partition the clusters based on their means. Ak- mode clustering algorithm is a two step process such as attribute weighing phase and clustering phase. Expectation-Maximization algorithm is an extension of K-means algorithm which can be used to find the parameter estimates for each cluster. Knox's method requires critical distance in time as well as space defining closeness has to be set but the determination of these critical distances requires subjective decision. Mantel approach does not however require use of critical distances but uses both time and space matrices. The Regional Crime Analysis

Program uses data mining and data fusion techniques in order to catch professional criminals. Another framework for crime trends uses a new distance measure for comparing all individuals based on their profiles and then clustering them accordingly. This method also provides a visual clustering of criminal careers and identification of classes of criminals. Exploratory Data Analysis techniques are interactive and visual, and there are many effective graphical display methods for relatively small data sets. Spatial point patterns (SPP) are based on coordinates of locations of crime incidences and is typically interpreted as analysis of clustering.

8.2 Limitations

Some of the limitations of crime record analysis using data mining includes crime pattern analysis can only help the detectives, not replace them. Moreover, data mining is sensitive to quality of input data that may be inaccurate, have missing information, be data entry error prone etc. And, mapping real data to data mining attributes is not always an easy task and often requires skilled data miner and crime data analyst with good domain knowledge. They need to work closely with a detective in the initial phases.

8.3 Further Developments

As a future extension of this study we will create models for predicting the crime hot-spots that will help in the deployment of police at most likely places of crime for any given window of time, to allow most effective utilization of police resources. We also plan to look into developing social link networks to link criminals, suspects, gangs and study their interrelationships. Additionally the ability to search suspect description in regional crime databases, traffic violation databases from different provinces, etc to aid the crime pattern detection or more specifically counter terrorism measures will also add value to this crime detection paradigm.

8.4 Summary

By making use of the available crime data and appropriate spatial and temporal methods in crime mapping, this study aims to observe crime and offender dynamics, and the factors affecting crime negatively and positively. The methods used for the study have practical implications due their high relevance to policy and wide applicability for law enforcement in other areas.

Chapter 9

Reference

[1] AdderleyR.William and Musgrove Peter, (2001), "*Police crime recording and investigation systems: A user's view*", An International Journal of Police Strategies and Management, Vol. 24

[2] Akpinar E. and Usul N. (2004). "*Geographic Information Systems Technologies in Crime Analysis and Crime Mapping*"

[3] Bharathi A., Shilpa R., (2014), "*A Survey on Crime Data Analysis of Data Mining Using Clustering Techniques*", International Journal of Advance Research in Computer Science and Management Studies, vol 2

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

[5] Chen H., Chung W., Chan Y.M., Xu J., ang G., Zheng R. and Atabakch H.,"*Crime Data Mining: An Overview and Case Studies*", proceedings of the annual national conference on digital government research, Boston, pp.1-5, (2003)

[6] de Bruin, J.S. ,Cocx, T.K. , Kosters, W.A. , Laros, J. and Kok, J.N. (2006) "*Data mining approaches to criminal career analysis*", in Proceedings of the Sixth International Conference on Data Mining (ICDM"06)

[7] G. O. Mohler, M. B. Short, P. J. Brantingham, F. P. Schoenberg, and G. E. Tita, "*Self-Exciting Point Process Modeling of Crime*", Journal of the American Statistical Association, Vol. 106, No. 493, 2011.

[8] Gwinn, S.L., Bruce, C., Cooper, J.P., Hick, S.: "*Exploring crime analysis. Readings on essential skills*", Second edition. Published by BookSurge, LLC (2008)

[9] Hand D., Mannila H., Smyth P.., (2001), "*Principles of Data Mining*" Prentice Hall.

[10] Hanmant N. Renushe, Prasanna R. Rasal, Abhijit S. Desai, (2012). " *Data Mining Practices for Effective Investigation of Crime*", International Journal of Computer Technology & Applications,Vol.3, Issue.3, pp.865-870.

[11] J. Agarwal, R. Nagpal, and R. Sehgal, "*Crime analysis using k-means clustering* ", International Journal of Computer Applications, Vol. 83 No4, December 2013

[12] Malathi. A and Dr. S. SanthoshBaboo, (2011), "*An Enhanced Algorithm to Predict a Future Crime using Data Mining*", International Journal of Computer Applications (0975 – 8887), Vol. 21

[13] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten (2009); "*The WEKA Data Mining Software: An Update* "; SIGKDD Explorations, Volume 11, Issue 1.

[14] Reza Fadaei-Tehrani, Thomas M. Green, (2002) "*Crime and society*" International Journal of Social Economics

[15] VijayaKumar M., Karthick S. and N.Prakash. (2013). "*The Day-To-Day Crime Forecasting Analysis of Using Spatial temporal Clustering Simulation*", International Journal of Scientific & Engineering Research, Vol.4, Issue.1, pp. 1-6.

[16] www.police.lk

Appendix A

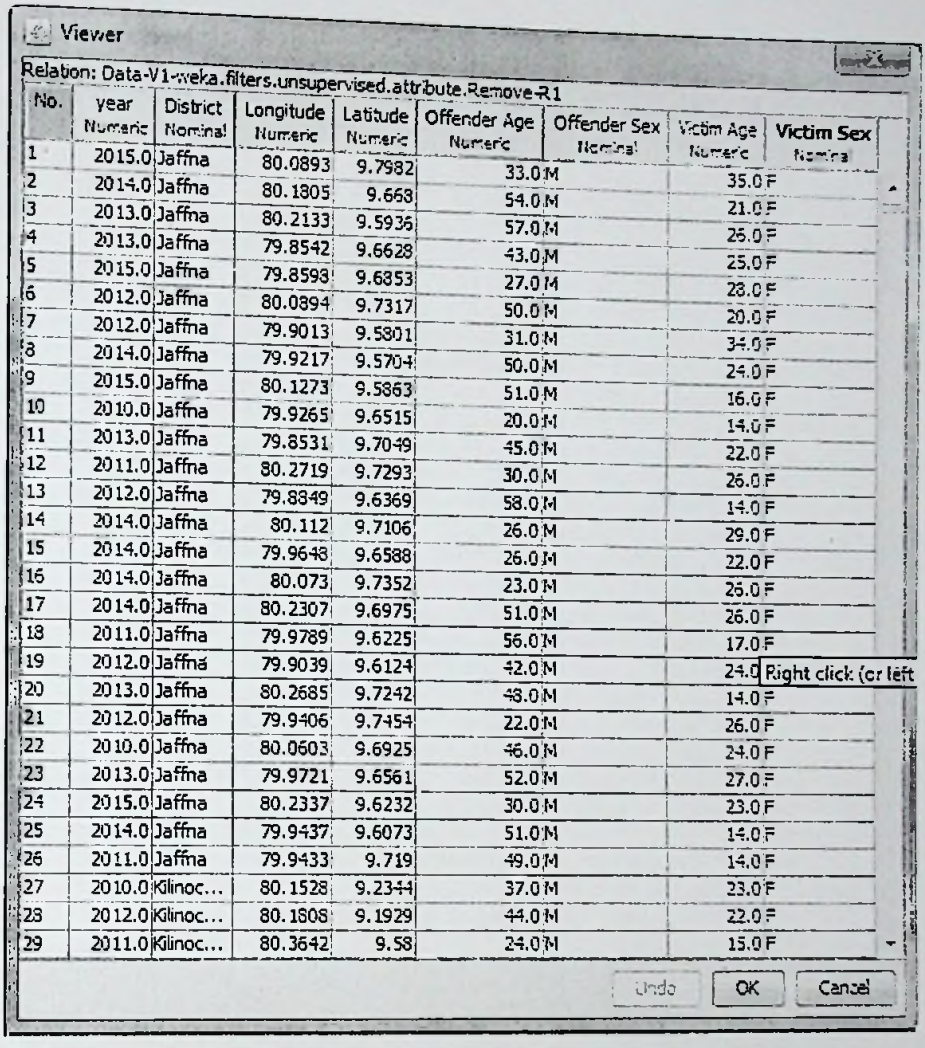


Figure A.1: Sample Data Set

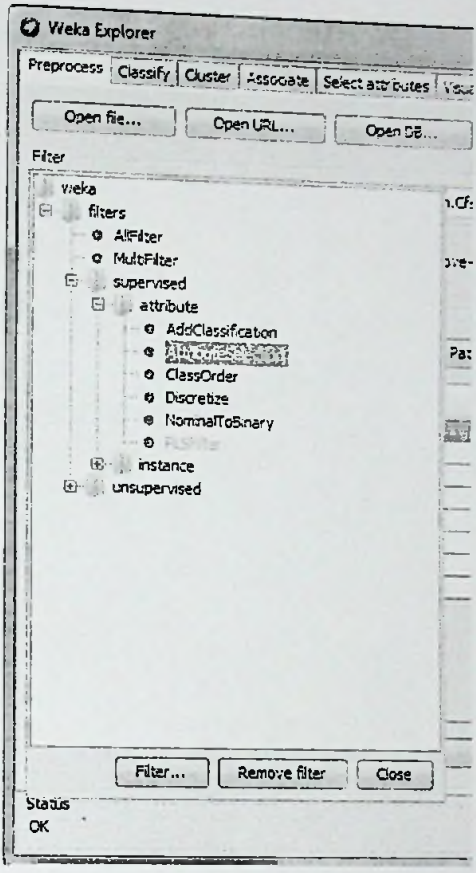


Figure A.2: Applying A ribute Subset Selec on

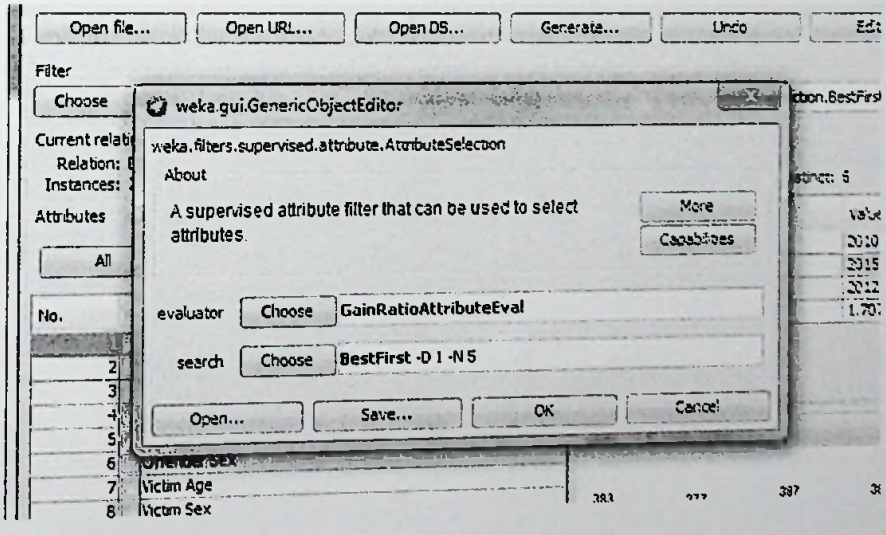


Figure A.3: Applying evaluator and search method for a ribute subset selec on

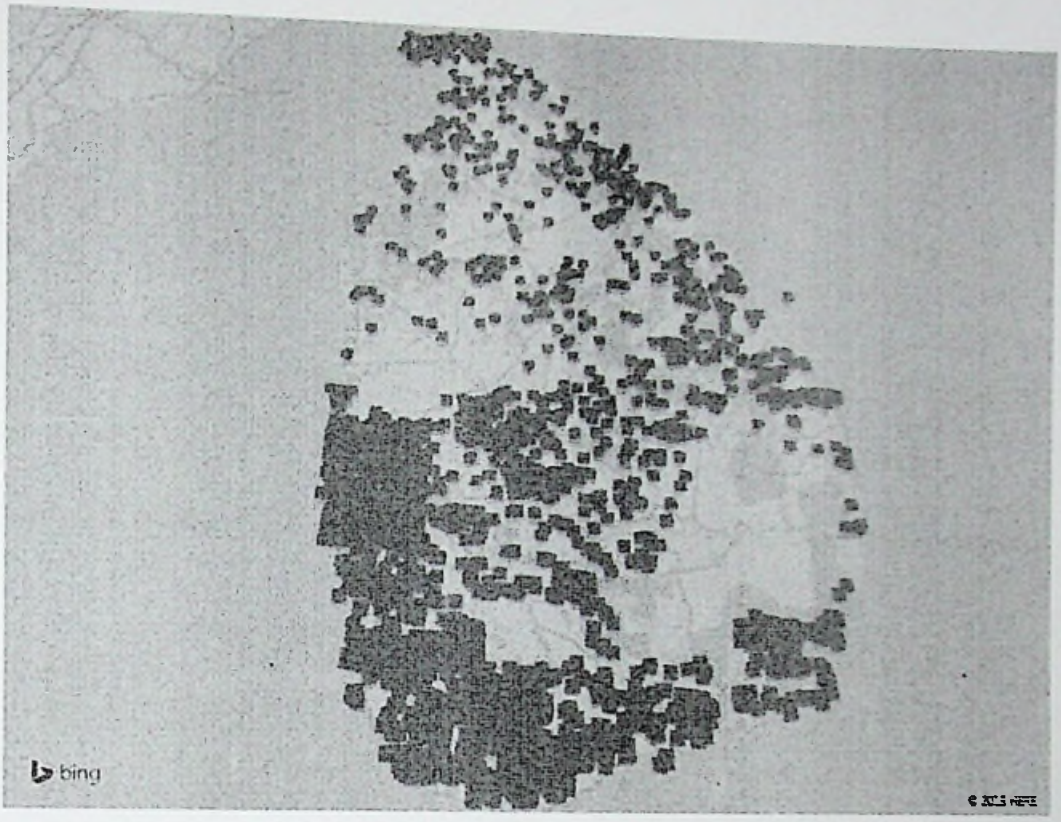


Figure A.4: Using Power Map to Iden fy Crime Hot Spots