

# Crime detection and criminal identification in India using data mining techniques

Devendra Kumar Tayal · Arti Jain ·  
Surbhi Arora · Surbhi Agarwal · Tushar Gupta ·  
Nikhil Tyagi

Received: 15 July 2013 / Accepted: 10 March 2014 / Published online: 1 April 2014  
© Springer-Verlag London 2014

**Abstract** In the current paper, we propose an approach for the design and implementation of crime detection and criminal identification for Indian cities using data mining techniques. Our approach is divided into six modules, namely—data extraction (DE), data preprocessing (DP), clustering, Google map representation, classification and WEKA<sup>®</sup> implementation. First module, DE extracts the unstructured crime dataset from various crime Web sources, during the period of 2000–2012. Second module, DP cleans, integrates and reduces the extracted crime data into structured 5,038 crime instances. We represent these instances using 35 predefined crime attributes. Safeguard measures are taken for the crime database accessibility. Rest four modules are useful for crime detection, criminal identification and prediction, and crime verification, respectively. Crime detection is analyzed using *k*-means clustering, which iteratively generates two crime clusters that are based on similar crime attributes. Google map improves visualization to *k*-means. Criminal identification and prediction is analyzed using KNN classification. Crime verification of our results is done using WEKA<sup>®</sup>. WEKA<sup>®</sup> verifies an accuracy of 93.62 and 93.99 % in the formation of two crime clusters using selected crime attributes. Our

approach contributes in the betterment of the society by helping the investigating agencies in crime detection and criminals' identification, and thus reducing the crime rates.

**Keywords** Clustering · Classification · Crime · Data mining · Google map · *k*-Means · K-NN · WEKA<sup>®</sup>

## 1 Introduction

Crime is an offense against the society that is often prosecuted and punishable by the law (Brantingham and Brantingham 1984; Siegel et al. 2003; Becker 2008). It has been observed that criminals commit crimes at any place and in any form. One of the well-known crimes in the world is the terror attack on World Trade Centre on September 11, 2001 (Okonkwo and Enem 2011). Some well-known crimes in India are given here in chronological order of their occurrence—(1) Jessica Lal who worked as a celebrity barmaid in Delhi was shot dead on April 30, 1999.<sup>1</sup> (2) Nithari serial murders in Uttar Pradesh during 2005–2006, where the dismembered bodies of several children were found in the sewers.<sup>2</sup> (3) Terrorist attacks in Mumbai on November 26, 2008, by terrorist organization that killed 166 people and wounded at least 308.<sup>3</sup> (4) Rape case in Delhi on December 16, 2012, in which the victim who was a paramedical student died from her injuries while

---

D. K. Tayal  
Department of CSE, Indira Gandhi Delhi Technical University  
for Women, New Delhi, India  
e-mail: dev\_tayal2001@yahoo.com

A. Jain (✉)  
Department of CSE/IT, Jaypee Institute of Information  
Technology, Noida, Uttar Pradesh, India  
e-mail: arti.jain@jiit.ac.in

S. Arora · S. Agarwal · T. Gupta · N. Tyagi  
Department of IT, Jaypee Institute of Information Technology,  
Noida, Uttar Pradesh, India

---

<sup>1</sup> Jessica Lal murder case, <http://www.ndtv.com/topic/jessica-lal>  
Accessed on April 28, 2013.

<sup>2</sup> Nithari case, <http://wcd.nic.in/nitharireport.pdf> Accessed on April  
30, 2013.

<sup>3</sup> Mumbai blast, [http://www.ipcs.org/pdf\\_file/issue/SR71-Final.pdf](http://www.ipcs.org/pdf_file/issue/SR71-Final.pdf)  
Accessed on April 28, 2013.

undergoing emergency treatment.<sup>4</sup> All of these crimes remained as headlines of the news for the very long time; sentiments of crores of people of India were attached to these crimes in sympathy, and many campaigns were raised in the protest. This reveals that crimes terrifically affect not only the victims but also the people of the country as a whole. So, the check on crimes and target to the criminals are inevitable that need to be performed by the law enforcement agencies to secure the country. These agencies along with additional computer data analysts are responsible for unambiguous and competent crime investigation from the voluminous crime data. We therefore propose an approach for crime detection and criminal identification (CDCI) using data mining techniques (DMT) (Chen et al. 2004; Adderley et al. 2007; Thongtae and Srisuk 2008; Yu et al. 2008; Malathi and Baboo 2011; Malathi et al. 2011; Hussain et al. 2012) for Indian cities.

In this paper, the proposed CDCI approach is divided into six modules, namely—(1) data extraction (DE) (Witten et al. 2011; Han et al. 2012), which extracts the unstructured crime data from various crime Web sources viz. National Crime Records Bureau (NCRB),<sup>5</sup> Committee to Protect Journalists (CPJ)<sup>6</sup> and other Web sources during the period of 2000–2012.<sup>7, 8</sup> (2) data preprocessing (DP) (Malathi and Baboo 2011; Malathi et al. 2011; Han et al. 2012) cleans, integrates and reduces the extracted crime data into structured 5,038 crime instances. We represent these instances by our 35 crime attributes. (3) Clustering (Green et al. 1976; Sayal and Kumar 2011) using *k*-means (Kulis and Jordan 2011; Hornik et al. 2012; Kaur et al. 2012), which groups crime instances iteratively into two clusters with similar attributes for crime detection. (4) Google Map Application Programming Interface (GMAPI), which embeds Google maps through JAVA<sup>®9</sup> Netbeans for user-friendly and improved visual aids to *k*-means.<sup>10</sup> (5) Classification (Witten et al. 2011; Han et al. 2012) using K-NN (KNN) (Okonkwo and Enem 2011), which discovers similarities among different crimes and

organizes them into predefined classes for criminal identification and prediction. (6) WEKA<sup>®</sup> (Malathi et al. 2011), which uses JAVA<sup>®</sup>-based graphical user interface for crime verification of our *k*-means results.<sup>11</sup>

CDCI here uses an integrated technology that makes it more secure and differentiated from others. Password-protected user interface is designed to access the tool, to view and analyze the generated results. In case the user wants to access the criminals' database, he needs to provide his identity proof. Once verified and permitted by the tool's administrator, the user can access the criminals' database.

The remainder of this paper is organized as follows: Sect. 2 discusses about the related literature review and its shortcomings. Section 3 discusses about the proposed CDCI methodology. Section 4 discusses about the CDCI experimentation and associated results. Finally, Sect. 5 concludes the paper.

## 2 Background and related work

Crimes in India are stoked up at an alarming rate, and criminals are opting for queer activities to commit them. Newspapers, Web blogs, etc. are day to day filled with various crime incidents. Some of the mystified crimes that occurred in India in last couple of years are mentioned here.<sup>12</sup>

A professor was beaten to death by his own students in Ujjain, Madhya Pradesh. A gang of nine taxi drivers from Gurgaon, Haryana, robbed and killed at least 35 people after offering them lift. Unruly mob stripped and molested a girl in full public view at the Gateway of India, Mumbai, on the New Year eve. Days after horrible Nithari killing, 4 decomposed bodies of children were recovered from abandoned godown in Punjab. Sexually assaulted teenage girls in the Kashmir valley are still struggling to cope up with trauma. These incidents reveal how crimes are becoming a growing blight in India and have become a dominant fact of an Indian life as well.

Some responsible factors that prevail in India for sheer increase in crimes are poverty, migration, unemployment, frustration, starvation, illiteracy, corruption, nepotism, inflation, etc. Impact of such crimes is that today people living in India now focus their eyes toward crime investigation agencies and security agencies to check and control crimes. Currently, physical investigation by agencies has the probability to ignore and neglect the supportive crime features. Most of these agencies are searching manually the

<sup>4</sup> Delhi rape case, <http://www.dailymail.co.uk/news/article-2269725/Indian-teenager-accused-Delhi-gang-rape-faces-maximum-year-jail-term.html> Accessed on April 28, 2013.

<sup>5</sup> National Crime Records Bureau, <http://ncrb.gov.in> Accessed on March 10, 2013.

<sup>6</sup> Committee to Protect Journalists, <http://www.cpj.org> Accessed on March 20, 2013.

<sup>7</sup> Crime alert, <http://www.crimelert.org> Accessed on March 25, 2013.

<sup>8</sup> NSW bureau of crime statistics and research, <http://www.bocsar.nsw.gov.au> Accessed on March 22, 2013.

<sup>9</sup> Jin F, Wang W, Xiao Y, Pan Z Proposal of Crime Data Mining Project. [https://filebox.vt.edu/users/xykid/dataAnalysisProject/Check-point-II\\_Jin\\_Xiao\\_Pan\\_Wang.pdf](https://filebox.vt.edu/users/xykid/dataAnalysisProject/Check-point-II_Jin_Xiao_Pan_Wang.pdf). Accessed on May 30, 2013.

<sup>10</sup> Netbeans, <http://netbeans.org> Accessed on April 2, 2013.

<sup>11</sup> WEKA manual, <http://www.inf.ufpr.br/lesoliveira/aprendizado/wekamanual.pdf> Accessed on March 15, 2013.

<sup>12</sup> Crime wave in India, <http://www.merineews.com/article/the-crime-wave-in-india/132433.shtml> Accessed on April 18, 2013.

database of criminals, which is a tedious process and takes much more time. Few of them work with the help of computer data analysts and are responsible for crime detection, criminal identification and prediction, and crime verification to ensure safety to the citizens of India. To contribute in this aspect, we propose our CDCI approach using DMT for Indian cities by consideration of selected crime features. Our methodology can help these agencies to filter crime database to find out the most probable criminals. This will save a lot of time for the agencies.

Literature survey details that earlier work related to the crime investigation carries some intrinsic limitations. Some of the authors have discussed primary clustering (Chen et al. 2004; Kulis and Jordan 2011; Malathi and Baboo 2011; Malathi et al. 2011; Sayal and Kumar 2011) and classification (Okonkwo and Enem 2011) techniques for crime detection, criminal identification theoretically; however, none of them provides a sound implementation for the same. Although some papers (Nath 2006; Malathi and Baboo 2011; Malathi et al. 2011) discuss application of *k*-means for crime detection, but these and other works (Ehlers 1998; Visher and Weisburd 1998; Gorr and Harries 2003; Gorr et al. 2003; Chen et al. 2004; Hussain et al. 2012) are deficient in integration among crime detection, criminal identification and prediction, and crime verification. Malathi et al. (2011) work with crime attributes—*number of crimes of a particular crime type*, e.g., murder and burglary, versus *years*. Our CDCI in addition to Malathi et al. (2011) works with attributes “number\_of\_crimes\_committed\_in\_year” versus “crime\_year” (please refer Sect. 4). These crime attributes are considered as follows: (1) independent of attributes “crime\_location” and “crime\_type”; (2) dependent on attribute “crime\_location,” but independent of attribute “crime\_type,” etc. Nath (2006) tries to detect crime suspects based on their *rac*es, *age* and *sex*. On the other hand, our CDCI speculates suspects based on the fine-grained attributes—“suspect\_name,” “suspect\_age,” “suspect\_sex,” “suspect\_facial\_feature,” “suspect\_other\_physical\_feature” and “suspect\_nationality.” Mande et al. (2012a, b) states that criminal identification is based on autocorrelation/Gaussian mixture models. They solely depend on the eye-witness information. They confine to only one state of India, i.e., Andhra Pradesh for criminal records. Our CDCI works on crime data of several Indian cities that are selected based on their crime rates. Jin et al. (see footnote 9) in their proposal define the position of crime events with *longitude* and *latitude* using *k*-means. They visualize intra-cluster distance through Google map using different colors. Our CDCI defines the formation of clusters firstly using *k*-means and then using GMAPI. GMAPI uses the crime attributes—“crime\_location,” “crime\_location\_longitude,” “crime\_location\_latitude” and “number\_of\_crimes\_committed\_in\_location” (please refer Sect. 4). Okonkwo and Enem (2011) confer about terrorism—9/11

attack as a type of crime. They recommend government to set up data mining agencies within the law enforcement agencies where various criminal data should be consolidated and mined. They focus on KNN’s theoretical details, but there is no implementation provided. Li and Juhola (2014) say that crime research is an area that can benefit from better visualization and DMT. Our CDCI proposal provides a consolidated and visualized approach for crime detection, criminal identification and prediction, and crime verification to shield India from heinous crimes.

### 3 Methodology

This section is divided into two subsections: Sects. 3.1 and 3.2. Section 3.1 describes the CDCI dataset for applying DMT and Sect. 3.1.1 for CDCI data protection. Section 3.2 describes the proposed CDCI approach.

#### 3.1 CDCI dataset

First of all, we generate the CDCI crime dataset through two sequential steps (1) DE extracts the unstructured crime data from various crime Web sources, namely—NCRB (see footnote 5), CPJ (see footnote 6) and other (see footnotes 7, 8) Web sources during the period of 2000–2012. (2) DP cleans, integrates and reduces the extracted crime data into structured 5,038 crime instances (.csv format). The structured CDCI crime dataset is represented using 35 crime attributes. The formulated dataset is implemented using two JAVA<sup>®</sup> tools—(1) Netbeans for crime detection, criminal identification and prediction. (2) WEKA<sup>®</sup> for crime verification. Some of the crime attributes are mentioned in Sect. 4. Figure 1 shows the CDCI sample dataset.

##### 3.1.1 CDCI data protection

Crime experts use their knowledge skills, intuition and past experiences when they deal with criminals and associated crime cases. They are not deviated by the incorrect crime data (e.g., person who is not a criminal but wrongly identified as criminal by the system). On the other hand, our machine learning model is highly dependent on the input of the crime data. To reduce this risk, we have extracted the crime data from reliable sources such as NCRB and CPJ.

Day by day, crime records are expanding in size which gives rise to some key problems such as data storage (Korukonda 2007) and data reliability. (1) Data storage aids in efficient storing of the crime database. But data storage need not be compatible and feasible at all times due to computers’ processing limitations. (2) Data need not be reliable when an agency has the incorrect crime data. In these scenarios, our data mining-based model, i.e., CDCI results may deviate

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
	crime_type	crime_date	crime_year	crime_reason	suspect_name	victim_name	victim_age	crime_location	suspect_age	suspect_sex	suspect_origin	suspect_facial_feature	suspect_weapon_type	victim_nationality
3963	Murder	22June.	2010	Criminal Grou	Mudassar	Ankur	Medium	Delhi	Young	Female	Rajasthan	Limp	Metal Rod	India
3965	Sexual Abu	17June.	2009	Government (James	Surya	Surya	Young	Delhi	Medium	Male	Haryana	Tattoo (Bicep)	Knife	India
3966	Murder	06June.	2012	Political Grou	Irfan	sunny	Young	Delhi	Young	Male	Maharashtra	Cut (Lip)	Gun	India
3967	Sexual Abu	31May.	2011	Military Offici	Bunty	Rishabh	Medium	Delhi	Medium	Male	Rajasthan	Tattoo (Bicep)	Unarmed	Sri Lanka
3968	Assault	28May.	2009	Government (Kamla	Sahil	Sahil	Young	Delhi	Young	Female	Haryana	Cut (Lip)	Chain	India
3969	Sexual Abu	27May.	2012	Political Grou	Raju	Akash	Elder	Delhi	Medium	Male	Bihar	Scar (Head)	Others	India
3970	Assault	27May.	2011	Political Grou	Ali Shaaban	Akshat	Elder	Delhi	Young	Male	Maharashtra	Limp	Others	India
3971	Rioting	21May.	2010	Unknown Fire	Mahad Salad A	Omkar	Elder	Mumbai	Medium	Male	Ghaziabad	Cut (Lip)	Gun	Iraq
3972	Assault	09May.	2009	Political Grou	Justin	Raunak	Young	Delhi	Elder	Male	Uttar Pradesh	Scar (Head)	Gun	Russia
3973	Sexual Abu	07May.	2008	Political Grou	Ali Ahmed Abc	Tauseen	Elder	Kolkata	Elder	Male	Rajasthan	Tattoo (Bicep)	Others	India
3974	Assault	07May.	2007	Political Grou	Rajesh Mishra	Manoj	Medium	Delhi	Medium	Male	Rajasthan	Scar (Head)	Unarmed	India
3975	Robbery	24April.	2006	Criminal Grou	Abukar Hassan	Parth	Elder	Kolkata	Medium	Male	Haryana	Scar (Head)	Others	India
3976	Rioting	20April.	2005	Government ( Lee	Manish	Manish	Young	Kolkata	Young	Male	Rajasthan	Cut (Lip)	Metal Rod	Brazil
3977	Mugging	20April.	2004	Political Grou	Rémi Ochlik	Shyam	Medium	Delhi	Elder	Male	Uttar Pradesh	None	Unarmed	India
3978	Robbery	19April.	2011	Military Offici	Marie Colvin	Nikhil	Medium	Mumbai	Medium	Male	Bihar	Cut (Lip)	Chain	India
3979	Assault	26March.	2002	Military Offici	Rami al-Sayed	Ashwin	Elder	Pune	Young	Male	Ghaziabad	Limp	Others	India
3980	Sexual Abu	22March.	2001	Military Offici	Mario Randolfi	Palash	Young	Mumbai	Elder	Male	Bihar	Tattoo (Bicep)	Unarmed	India
3981	TSNS	19March.	2000	Military Offici	Mazhar Tayyar	Prathamesh	TSNS	Bangalore	TSNS	Male	Uttar Pradesh	TSNS	TSNS	UAE
3982	Robbery	18March.	2009	Military Offici	Hassan Osman	Nithin	Elder	Delhi	Medium	Male	Rajasthan	Tattoo (Bicep)	Gun	India
3983	Mugging	18March.	2012	Political Grou	Nicholas	Nishant	Young	Mumbai	Medium	Male	Haryana	Limp	Unarmed	Iraq
3984	Rioting	07March.	2010	Political Grou	Ahmed Ismail	Ajith	Elder	Mumbai	Young	Male	Uttar Pradesh	Cut (Lip)	Others	India
3985	Sexual Abu	21Feb.	2011	Criminal Grou	Anas al-Tarsha	Siddharth	Medium	Kolkata	Young	Male	Uttar Pradesh	None	Chain	India
3986	Mugging	14Feb.	2005	Government (Nicole	Rajeev	Rajeev	Elder	Delhi	Medium	Male	Haryana	Cut (Lip)	Knife	India
3987	Rioting	11Feb.	2010	Criminal Grou	Rita	Karan	Elder	Pune	Elder	Female	Haryana	Scar (Head)	Knife	India
3988	Assault	18Jan.	2009	Local Reside	Stacy	Judy	Young	Kolkata	Medium	Male	Maharashtra	Tattoo (Bicep)	Knife	Nicaragua

Fig. 1 CDCI sample database

from their actual analysis. To reduce these risks, we have tried to compress the dataset using DP techniques and DE is done from reliable sources.

Some measures that are taken care to safeguard the personal information comprise of the data quality process. This phenomenon states that personal data should not be disclosed, but should be relevant to the purposes for which it is intended to be used and to the extent necessary for those purposes—be accurate, complete and kept up-to-date. Keeping this in mind, we provide two types of login in our CDCI tool—Admin login and Guest login. Admin login is provided to the administrator of the CDCI. Admin has complete authority to access the entire criminal database and personal information also. In case the user wants to access the criminals' database, he needs to provide his identity proof. Once verified and permitted by the tool's administrator, the user gains the right to access the criminals' database. Guest login is for limited access and is provided according to the type of the user to view/analysis the CDCI tool or other related information. CDCI is thus more secure and differentiated from existing crime investigating processes.

### 3.2 Proposed CDCI approach

This section depicts the work flow of our proposed CDCI (Fig. 2) using DMT for Indian cities. The work flow starts with DE step followed by DP step, which generated CDCI database. This database is then supplied to other CDCI modules—clustering and classification. CDCI clustering uses *k*-means which replaces missing value of an instance attribute with mean/mode that is computed from other given instances over the same attribute. *k*-Means groups crime instances iteratively into clusters with similar

attributes for crime detection. CDCI clustering is then followed by GMAPI, which embeds Google maps through Netbeans for user-friendly and improved visual aid to *k*-means. The CDCI classification uses KNN which discovers similarities among different crimes and organizes them into predefined classes for criminal identification and prediction. CDCI then employs WEKA<sup>®</sup> for crime verification of our *k*-means results. To do the same, WEKA<sup>®</sup> converts the CDCI database (.csv) format to WEKA<sup>®</sup> workable (.arff) format (see footnote 11). WEKA<sup>®</sup> verifies high accuracy in the formation of two crime clusters using selected crime attributes (please refer Sect. 4).

CDCI being an integration of various data mining modules such as DE, data preprocessing, clustering, visualization and classification. It thus gains insight into the crimes and facilitates into the detection of prime crime suspects by filtering out the huge crime data. CDCI can help the police and justice departments to narrow down the identification of criminals. This in turn will reduce the cost and time of crime investigation.

## 4 Experimentation and results

Whenever a crime takes place, detection organizations look into their criminal database to identify the criminals. In this process, each and every criminal get investigated by them, whereas our CDCI uses DMT, e.g., *k*-means and KNN, which improve the filtration of huge crime database for the identification of prime crime suspects. Thus, the investigation is set to be impact upon and influenced by the reduction in time and effort by our CDCI development. Hence, this section discusses about the experimentation

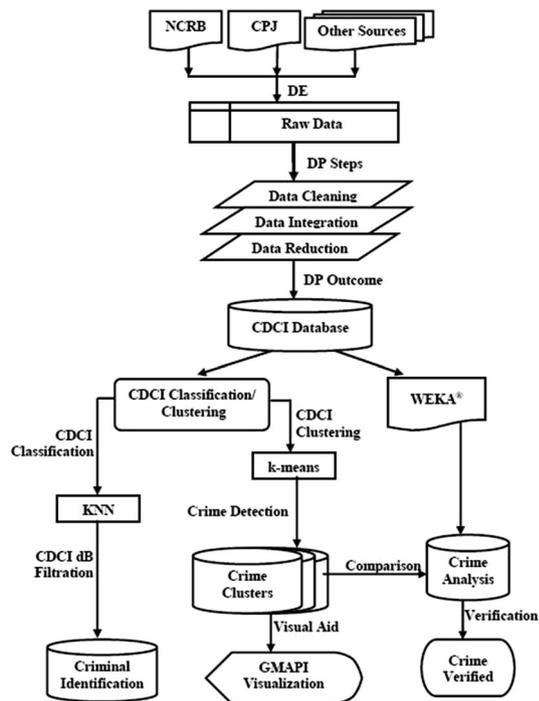


Fig. 2 CDCI work flow

and results that our CDCI performs using two tools—Netbeans 7.2.1 (see footnote 10) and WEKA® 3.6.2 (see footnote 11). This section is divided into five subsections: Sects. 4.1–4.5 as mentioned below.

#### 4.1 CDCI for selection of Indian cities

CDCI approach chooses seven Indian cities (Delhi, Kolkata, Mumbai, Pune, Jaipur, Hyderabad and Bengaluru) based on the percentage of their crime rates in diminishing order (please refer Fig. 3). Total numbers of crimes in these cities during 2000–2012 are analyzed with line graphs (please refer Fig. 4). Figure 4 is generated from CDCI crime dataset that uses the attribute “crime\_year” (represented by *Years* on *X*-axis) versus attribute “number\_of\_crimes\_committed\_in\_year” (represented by *Number of Crimes* on *Y*-axis).

#### 4.2 CDCI *k*-means implementation

CDCI searches for immanent patterns and relations in the given crime data by using the *k*-means and GMAPI techniques. These techniques provide an overview of large amount of the crime data and facilitate in handling, searching and retrieving of the desired crime information. CDCI can also be useful for crime prevention. Clusters are formulated for seven Indian cities that are selected based on their crime rates. Also, particular crime based on

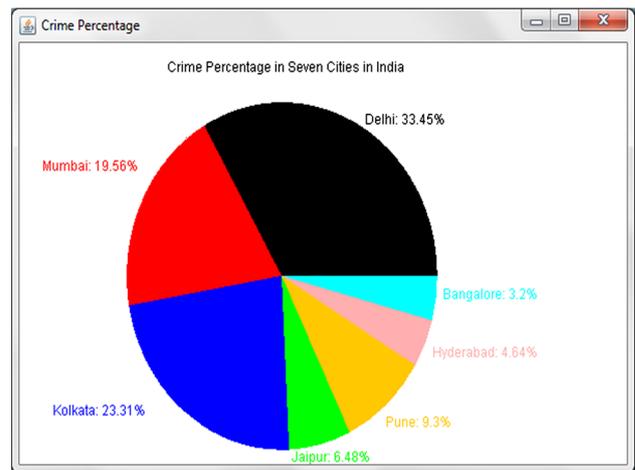


Fig. 3 Crime rates in % of seven Indian cities

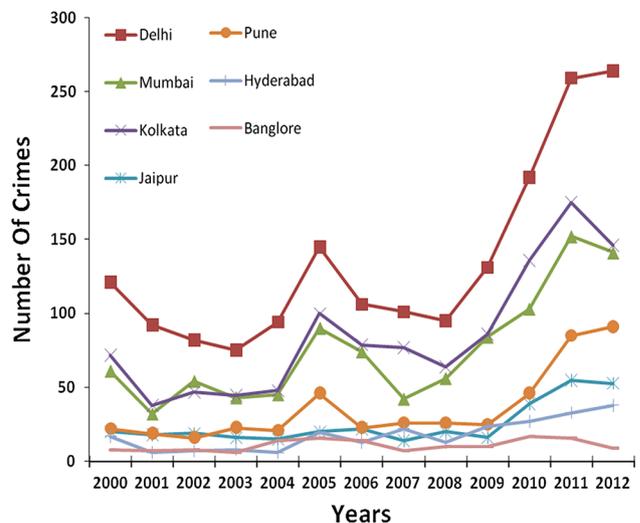


Fig. 4 Number of crimes versus years for seven Indian cities

location type can help the investigating agencies to take proper security measures against that crime. For example, Delhi has the highest number of rape reports among the Indian cities. So security officials should arrange special security for women in Delhi, which may prevent the crime occurrence to sensitive areas to some extent. In this section, CDCI implements *k*-means clustering. Clustering is done through Netbeans in Sect. 4.2.1 as Case 1 to Case 4 (as mentioned below) for crime detection in India. To define these cases, *k*-means uses crime attributes “crime\_year,” “crime\_type” and “crime\_location.” And Netbeans derive an additional important crime attribute “number\_of\_crimes\_committed\_in\_year” from the attribute “crime\_year.”

4.2.1 CDCI *k*-means for Case 1 to Case 4

*Case 1 crime detection in India during 2000–2012* *k*-Means aims to group objects (crimes in India during 2000–2012) as—*A* number of crimes in 2000, *B* number of crimes in 2001, *C* number of crimes in 2002 ... *L* number of crimes in 2011, *M* number of crimes in 2012 into precise clusters. Clusters are based on the two crime attributes “crime\_year” and “number\_of\_crimes\_committed\_in\_year” irrespective of attributes “crime\_location” and “crime\_type.”

We choose the number of clusters = 2 and name them as  $G_1$  and  $G_2$ . Initial seed points (or centroids) of these clusters are  $c_1$  and  $c_2$  that are based on (“crime\_year,” “number\_of\_crimes\_committed\_in\_year”). Here, we select  $c_1$  as (2002, 233) and  $c_2$  as (2009, 376). We assign each object (*A*, *B*, *C* ... *N*, *M*) to that cluster ( $G_1$  or  $G_2$ ), which has the minimum distance with respect to object. Figure 5 shows that objects *A*, *B*, *C*, *D*, *E*, *G*, *H*, *I* belong to  $G_1$  and *F*, *J*, *K*, *L*, *M* belong to cluster  $G_2$ . Since clusters are generated iteratively from CDCI crime dataset. So, at the end of first iteration, we get  $c_1$  as (2004.17, 246.16) and  $c_2$  as (2007.57, 508.71). Figure 6 shows that objects *A*, *B*, *C*, *D*, *E*, *G*, *H*, *I*, *J* belong to  $G_1$  and *F*, *K*, *L*, *M* belong to cluster  $G_2$ . Similarly,

at the end of second iteration, we get  $c_1$  as (2004.0, 278.0) and  $c_2$  as (2009.5, 633.08). Figure 7 shows that objects *A*, *B*, *C*, *D*, *E*, *F*, *G*, *H*, *I*, *J* belong to  $G_1$  and *K*, *L*, *M* belong to cluster  $G_2$ . At the end of third iteration, we get  $c_1$  as (2004.5, 294.2) and  $c_2$  as (2011.0, 698.0). Figure 8 shows that objects *A*, *B*, *C*, *D*, *E*, *F*, *G*, *H*, *I*, *J* belong to  $G_1$  and *K*, *L*, *M* belong to cluster  $G_2$ . We now compare the grouping of the last iteration and the current iteration which reveals that the objects does not move group anymore (please refer Figs. 7 and 8). Thus, the

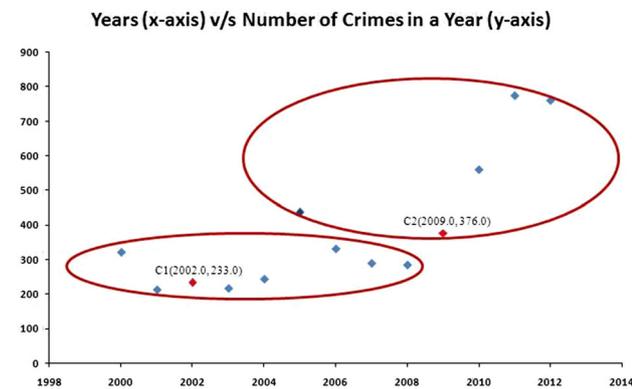


Fig. 5 CDCI *k*-means clusters with initial centroids

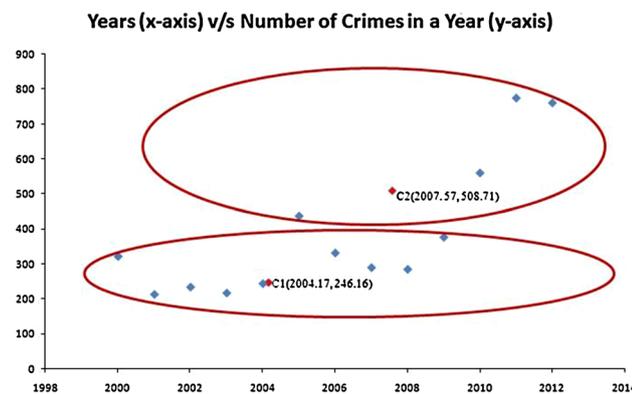


Fig. 6 CDCI *k*-means clusters at the end of first iteration

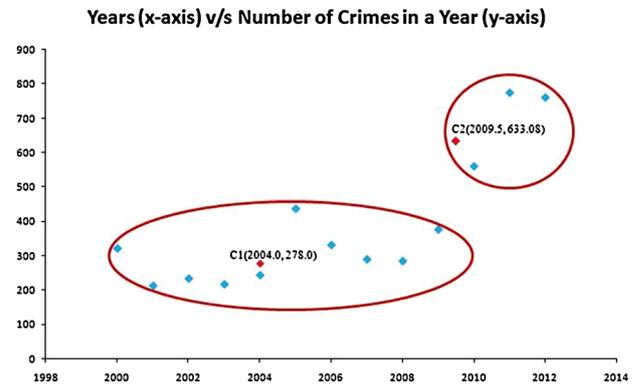


Fig. 7 CDCI *k*-means clusters at the end of second iteration

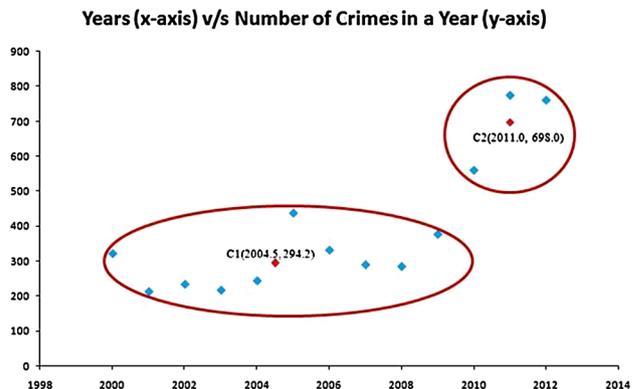


Fig. 8 CDCI *k*-means clusters at the end of third iteration

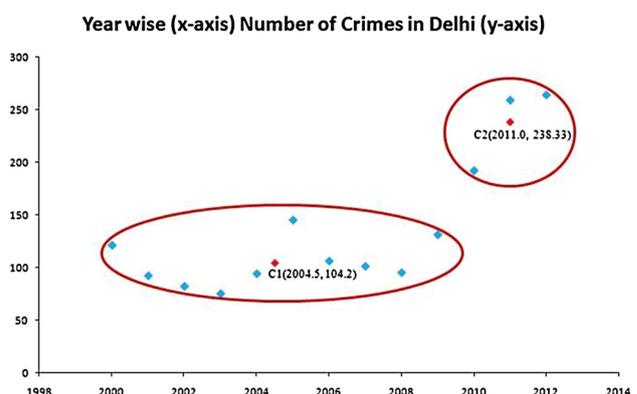
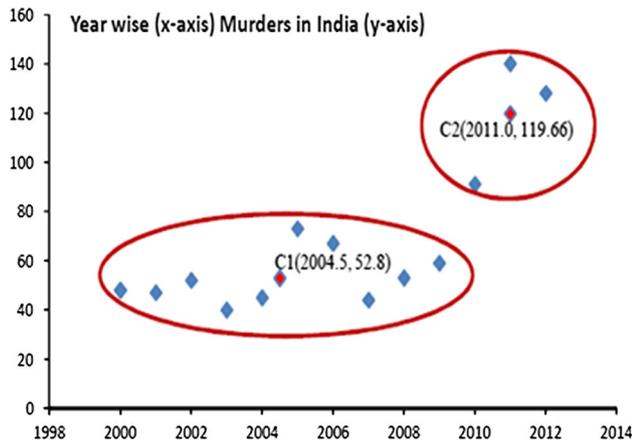


Fig. 9 CDCI for crimes in Delhi (2000–2012)

computation of *k*-means clustering reaches to its stability, and no more iteration is needed. It is pertinent to note that these clusters are independent of attributes “crime\_location” and “crime\_type.”

*Case 2 crime detection in Delhi during 2000–2012* Clusters are generated to detect the number of crimes in a



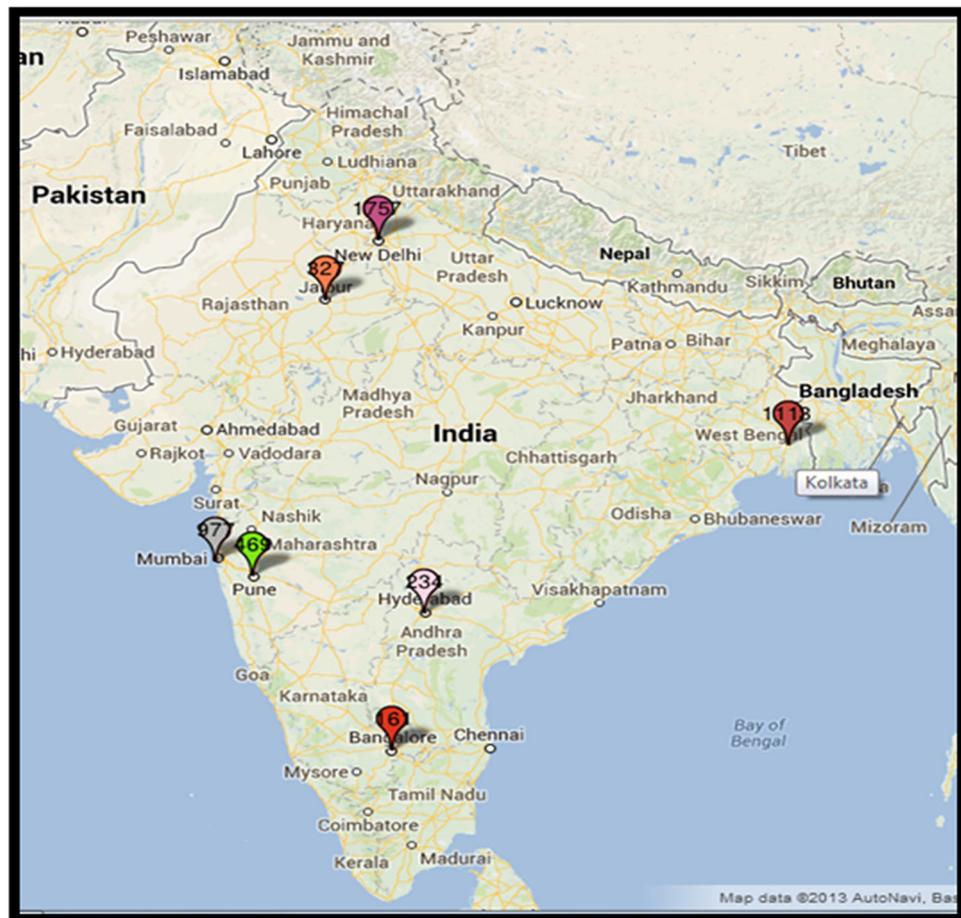
**Fig. 10** CDCI clusters for murders in India (2000–2012)

specified location (say Delhi) during 2000–2012 (please refer Fig. 9). Here again, attributes “crime\_year” and “number\_of\_crimes\_committed\_in\_year” are used to generate clusters, but they depend on the attribute “crime\_location” = Delhi and independent of the attribute “crime\_type.” In the same way, for other crime locations in India, similar clusters can be generated.

*Case 3 crime detection of type murders in India during 2000–2012* Clusters are generated to detect the number of crimes of specific type (say murder) during 2000–2012 in India (please refer Fig. 10). Here also, attributes “crime\_year” and “number\_of\_crimes\_committed\_in\_year” are used to generate clusters, but they depend on attribute “crime\_type” = Murder and independent of attribute “crime\_location.” In the same way, for other crime types in India, similar clusters can be generated.

*Case 4 crime detection of type murders in Delhi during 2000–2012* As we generate clusters for Case 1 to Case 3, similar clusters can be generated for this case where attributes “crime\_type” and “crime\_location” are specified explicitly. This case helps to detect which crime type is at

**Fig. 11** CDCI performs GMAPI for selected Indian cities



peak in a given location. For instance, Delhi has high crime rate for murders during 2000–2012.

### 4.3 CDCI GMAPI implementation

In order to enhance *k*-means results, CDCI performs GMAPI. GMAPI embeds Google maps through Netbeans (see footnote 10) for user-friendly and improved visual aid to *k*-means. Figure 11 shows total number of crimes (during 2000–2012 in seven selected Indian cities) as cluster values, along with cities within the map of India. For this purpose, we choose the crime attributes for GMAPI as “crime\_location,” “number\_of\_crimes\_committed\_in\_location,” “crime\_location\_longitude” and “crime\_location\_latitude.” The two attributes “crime\_location\_longitude” and “crime\_location\_latitude” are used to plot the crime location markers on Google map, for example, “crime\_location\_longitude” = 28.63°N and “crime\_location\_latitude” = 77.22°E for crime marker to be placed at “crime\_location” = Delhi. Rest two attributes “crime\_

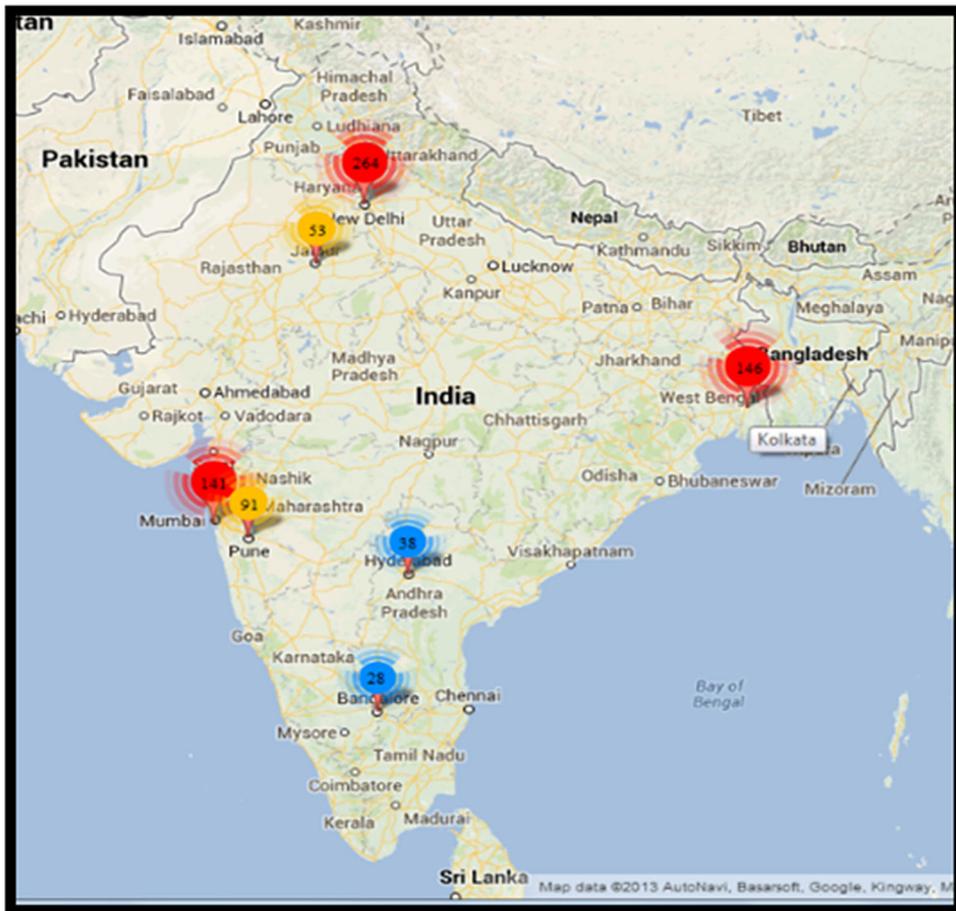
location” and “number\_of\_crimes\_committed\_in\_location” are used for crime detection during 2000–2012.

Now, if we want to know about the crimes in Indian cities for a particular year (say 2012), we can directly locate with GMAPI (please refer Fig. 12). Clusters with different number of crimes are represented by different color markers as:

-  : blue color cluster for number of crimes (0–50)
-  : orange color cluster for number of crimes (50–100)
-  : red color cluster for number of crimes (>100)

In GMAPI, we neither require to select the number of clusters ahead of time nor there need to select the initial centroids. Also, however large is the crime dataset, GMAPI easily runs on that data. GMAPI creates crime clusters that identify the hot spots of crime locations. Thus, GMAPI speeds the crime investigation and points to enforce security measures in those effected locations.

**Fig. 12** CDCI crime clusters in GMAPI for year 2012



4.4 CDCI KNN implementation

KNN approach stores all available objects and classifies new objects based on the similarity measure (Han et al.



Fig. 13 KNN user interface in CDCI

Fig. 14 Criminal identification in CDCI using KNN

Suspects																				
Shafiqah Khan	Male	Young	Haryana	None	Assault	2	Gun	Mumbai	02July	Mosab al-Obaidah	Male	Young	88.85970715-76.98830981	Syria	SEVENTH	Tishreen	Syria	Print/Reporter/Print Reporter	Culture/Politics/Sports/War	
Chishti Mujahid	Male	Young	Ghazabad	Cut (Lip)	Assault	2	Gun	Delhi	26May	Nur Mose Hussein	Male	Young	88.91332247-77.00832491	Somalia	FIFTH	Radio IQK	Somalia	Radio	Broadcast	Politics/War
Juan Emilio Andujar Matos	Male	Medium	Haryana	Cut (Lip)	Assault	5	Gun	Pune	07Nov	Ganjam Das	Male	Young	88.89751775-76.94089449	Bangladesh	SIXTH	Samakal	Bangladesh	Print/Reporter	Corruption/Crime/Politics	
Michael Kelly	Male	Elder	Uttar Pradesh	Scar (Head)	Assault	1	Gun	Delhi	09May	Afian Khasanov	Male	Young	88.92704908-77.03246938	Russia	THIRD	Reuters	Russia	Television/Camera Operator	Politics/War	
Joito Evarado	Male	Elder	Bihar	Scar (Head)	Assault	2	Gun	Pune	01June	Milan Pantic	Male	Young	88.90593793-77.06286386	Yugoslavia	SECOND	Vecernje Novosti	Yugoslavia	Print/Reporter	Business/Corruption/Crime	
Reynaldo Monay	Male	Medium	Maharashtra	Tattoo (Bicep)	Assault	1	Gun	Delhi	29March	Kerem Lawton	Male	Young	88.92919037-77.02762565	Yugoslavia	THIRD	Associated Press Television News	United Kingdom	Television/Producer	War	
Jamilah Hashimzade	Male	Young	Ghazabad	Limp	Assault	6	Gun	Mumbai	03May	Azzefine Saifj	Male	Young	88.91153258-76.93800067	Algeria	SIXTH	El-Ouma	Algeria	Print/Editor	Culture/Politics/War	
Mazk Saha	Male	Elder	Bihar	Scar (Head)	Assault	3	Gun	Delhi	29July	Cetin Abovay	Male	Young	88.88993849-76.9230789	Turkey	SIXTH	Ogur Halk	Turkey	Print/Editor	TSNS	
MAGGIE	Male	Young	Ghazabad	Cut (Lip)	Assault	10	Gun	Delhi	06July	ZULMA	Male	Young	88.9522529-77.0377303	Sri Lanka	THIRD	Voice of Tigers	United Kingdom	Radio/Alcoholist	War	
EMOIA	Female	Medium	Bihar	Limp	Assault	5	Gun	Delhi	04July	MUSA	Male	Young	88.91334914-77.03078712	Somalia	THIRD	Radio Jowhar	United States	Radio/Architect	Politics	
ROBERTO	Male	Young	Rajasthan	Cut (Lip)	Assault	1	Gun	Mumbai	09July	SHEREEN	Male	Young	88.92051076-77.03335993	Iraq	THIRD	Al-Iraqiya	Colombia	Television/Actor	War	
ANDRES	Male	Elder	Maharashtra	Cut (Lip)	Assault	1	Gun	Jaypur	09July	PUIJA	Male	Young	88.94850497-77.01847208	India	FOURTH	Asomiya Khabar	Mozambique	Print/Income tax officer	Corruption	
				Tattoo															Corruption	

2012). In our paper, we have used KNN for criminals’ identification by looking at the past crimes and finding similar ones that match the current crime based on  $k$  ( $k$ : number of nearest neighbors matched). In Fig. 14, output of 12 prime suspects is shown with KNN when only three input attributes (“crime\_type” = Assault && “victim\_sex” = Male && “victim\_age” = Young) are given and rest of the attributes have null values. We have designed the user interface to input the crime attribute values for KNN (please refer Fig. 13). Here, KNN is shown to apply on only 6 attributes; however, it is valid for all 35 attributes but for the sake of brevity it is not shown here.

We can say that CDCI uses KNN for criminals’ identification by filtering the number of suspected criminals. Criminals may then be executed and prosecuted by the law and justice of India. This may currently take long time, may be years, depending on the severity of the crimes.

Again for criminal prediction, KNN assigns the new case to the same crime class to which most of its neighbors belong. When a new criminal is observed with no past records, then CDCI identifies its nearest neighbors, i.e., criminals with the same crime pattern. KNN assumes that

in order to predict criminal, CDCI looks for records with similar predictor values in the crime database that is nearest. But if no pattern is matched, then the new criminal is added in our database.

#### 4.5 CDCI using WEKA®

WEKA® is a useful tool in the analysis of the real-world datasets. Since WEKA® (please refer Fig. 15) undergoes testing of several data mining algorithms, so it acts as a base system in the verification process. In this section, the results of *k*-means clusters (as obtained in Sect. 4.2.1, Case 1) are verified with WEKA® (please refer Table 1). WEKA® verifies an accuracy of 93.62 and 93.99 % in the formation of two crime clusters using the selected crime attributes.

Error Calculation:

$$G_1 \text{ Cluster} = \text{MOD}(2,445 - 2,289) / 2445 = 0.0638$$

$$G_2 \text{ Cluster} = \text{MOD}(2,593 - 2,749) / 2593 = 0.0601$$

So, Accuracy Measure:

$$G_1 \text{ Cluster} = 93.62 \%$$

$$G_2 \text{ Cluster} = 93.99 \%$$

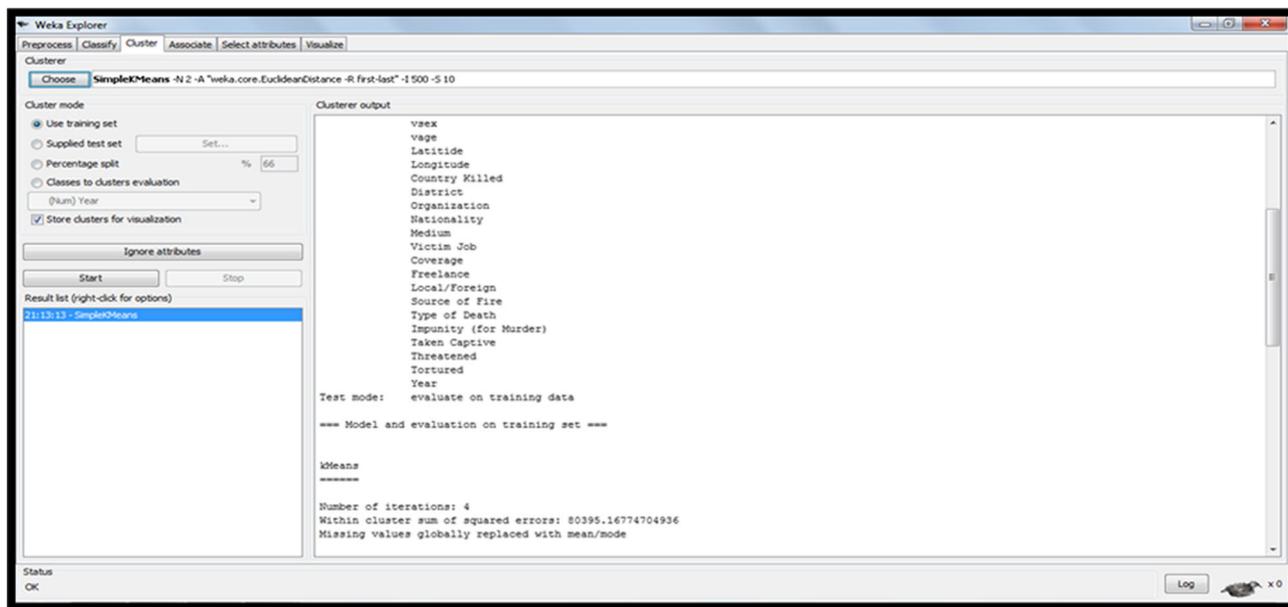
### 5 Conclusion

Crimes in India are rising at an alarming rate because of the factors such as increase in poverty, migration, unemployment, frustration, illiteracy and corruption. Crime investigating agencies search the database of criminals manually

**Table 1** *k*-Means clusters ( $G_1$  and  $G_2$ )

Cluster ID	<i>k</i> -means (WEKA®)	<i>k</i> -means (code)
$G_1$	2,445	2,289
$G_2$	2,593	2,749

or with some computer data analyst which is a tedious process and takes much more time. So to contribute toward combating crimes and to identify criminals, we propose an integrated technology of CDCI using DMT for Indian cities. Selection of seven Indian cities is based on their crime rates. Our CDCI extracts unstructured crime data from various crime Web sources and then preprocessed the crime data into structured 5,038 instances that are represented using 35 predefined crimes attributes. Password-protected user interface is designed to access the CDCI tool. CDCI then applies *k*-means clustering for crime detection during 2000–2012 through four cases. Case 1 detects crimes in India irrespective of crime location and crime type. Case 2 detects crimes in specific location, e.g., Delhi, irrespective of crime type. Case 3 detects crimes of specific type, e.g., murders, irrespective of crime location. And Case 4 detects crimes of specific type and in specific location. To enhance *k*-means results, the CDCI performs GMAPI which embeds Google maps through Netbeans. CDCI also applies KNN classification for criminals' identification and prediction. KNN looks at the past crimes and finds similar ones that match the current crime based on the number of nearest neighbors' matched. CDCI then uses WEKA® to verify *k*-means, Case 1 results. We



**Fig. 15** WEKA® tool in CDCI

measure an accuracy of 93.62 and 93.99 %, respectively, in the formation of two crime clusters using selected crime attributes.

Investigating agencies can utilize our proposed data mining tool to ease their crime investigation process. CDCI can speed up the crime solving process by processing and filtering the voluminous crime data within a short span of time. Thus, CDCI can aid the law enforcement agencies to enforce the security of citizens of India.

## 6 Future work

In future, we can enhance data privacy, reliability, accuracy and other security measures of our crime-based data mining system. We shall also collaborate with security agencies in India.

## References

- Adderley R, Townsley M, Bond J (2007) Use of data mining techniques to model crime scene investigator performance. *Knowl Based Syst* 20(2):170–176
- Becker RF (2008) *Criminal investigation*, 3rd edn. Jones and Bartlett Learning Publishers, Burlington, MA
- Brantingham PJ, Brantingham PL (1984) *Patterns in crime*. Macmillan/McGraw-Hill School Division, New York
- Chen H, Chung W, Xu JJ, Wang G, Qin Y, Chau M (2004) Crime data mining: a general framework and some examples. *Comput IEEE* 37(4):50–56
- Ehlers D (1998) Predicting crime: a statistical glimpse of the future? *Nedbank ISS Crime Index*. Halfway House: Inst for Security Stud 2(2)
- Gorr W, Harries R (2003) Introduction to crime forecasting. *Int J Forecast* 19(4):551–555
- Gorr W, Olligschlaeger A, Thompson Y (2003) Short-term forecasting of crime. *Int J Forecast* 19(4):579–594
- Green EJ, Booth CE, Biderman MD (1976) Cluster analysis of burglary M/Os. *J Police Sci Adm* 4(4):382–388
- Han J, Kamber M, Pei J (2012) *Data mining: concepts and techniques*, 3rd edn. Morgan Kaufmann Publishers Inc, San Francisco, CA
- Hornik K, Feinerer I, Kober M, Buchta C (2012) Spherical k-means clustering. *J Stat Softw Am Stat Assoc* 50(10):1–22
- Hussain KZ, Durairaj M, Farzana GRJ (2012) Criminal behavior analysis by using data mining techniques. In: International conference on advances in engineering, science and management, IEEE, pp 656–658
- Kaur N, Sahiwal JK, Kaur N (2012) Efficient k-means clustering algorithm using ranking method in data mining. *Int J Adv Res Comput Eng Technol (IJARCET)* 1(3):85–91
- Korukonda AP (2007) Technique without theory or theory from technique? An examination of practical, philosophical and foundational issues in data mining. *AI Soc* 21(3):347–355
- Kulis B, Jordan MI (2011) Revisiting k-means: new algorithms via bayesian nonparametrics. In: 29th International conference machine learning. Omnipress, Edinburgh, pp 513–520
- Li X, Juhola M (2014) Country crime analysis using self organizing map, with special regard to demographic factors. *AI Soc* 29(1): 53–68
- Malathi A, Baboo SS (2011) Evolving data mining algorithms on the prevailing crime trend—an intelligent crime prediction model. *Int J Sci Eng Res* 2(6)
- Malathi A, Baboo SS, Anbarasi A (2011) An intelligent analysis of a city crime data using data mining. In: International conference information electronic engineering, vol 6. IACSIT Press, Singapore, pp 130–134
- Mande U, Srinivas Y, Murthy JVR (2012a) Feature specific criminal mapping using data mining techniques and generalized gaussian mixture model. *Int J Comput Sci Commun Netw* 2(3):375–379
- Mande U, Srinivas Y, Murthy JVR (2012b) An intelligent analysis of crime data using data mining & auto correlation models. *Int J Eng Res Appl (IJERA)* 2(4):149–153
- Nath SV (2006) Crime pattern detection using data mining. International conference on web intelligence and intelligent agent technology, IEEE/WIC/ACM, pp 41–44
- Okonkwo RO, Enem FO (2011) Combating crime and terrorism using data mining techniques. In: 10th International conference IT people centred development, Nigeria Computer Society, Nigeria
- Sayal R, Kumar VV (2011) A novel similarity measure for clustering categorical data sets. *Int J Comput Apps* 17(1):25–30
- Siegel D, Bunt H, Zaitch D (2003) *Global organized crime: trends and developments*. Kluwer, London
- Thongtae P, Srisuk S (2008) An analysis of data mining applications in crime domain. In: 8th International conference on computer and information technology workshops (ICCIT), IEEE, pp 122–126
- Visher CA, Weisburd D (1998) Identifying what works: recent trends in crime prediction strategies. *Crime Law Soc Change* 28:223–242
- Witten IH, Hall MA, Frank E (2011) *Data mining: practical machine learning tools and techniques*, 3rd edn. Morgan Kaufmann Publishers, Los Altos, CA
- Yu G, Shao S, Luo B (2008) Mining crime data by using new similarity measure. In: 2nd International conference genetic and evolutionary computing (WGEC), IEEE, Washington, USA, pp 389–392