Hot Spot Identification Using Kernel Density Estimation for Serial Crime Detection

S. Sivaranjani[✉], M. Aasha, and S. Sivakumari

Department of Computer Science and Engineering, Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore 641 108, India
sivaranjanicse@gmail.com

Abstract. A hot-spot mapping is an advanced crime detection technique which helps police personnel to identify high-crime areas and the best way to respond. However identification of crime location with less man power would be a more difficult process. In this work, Social crime data aware kernel density estimation based serial crime detection approach (SAKDESD) is implemented to group the serial and social crime data set in terms of more similarity. Social crime data set consists of various user comments about the crime happening in different locations which can provide the in-depth information about the serial crimes. The unstructured social crime data set is pre-processed to obtain meaningful structured format. This work also adapts the latent semantic approach for finding the similar topics present in the social crime data set which can lead to accurate prediction and efficient grouping of serial crimes. The experimental tests were conducted in matlab simulation environment which proves that the proposed approach SAKDESD provides a better result than the existing approach such as Modified graph cut clustering algorithm (MGCC).

Index Terms: Serial crime · Social crime data · Kernel density estimation
Latent semantic

1 Introduction

Serial crimes are constant threats happening in different locations made by same person in similar manner. These crimes need to be identified and addressed for ensuring public safety. Investigation departments are responsible for assuring the social protection by finding the criminal who are responsible for the series threats happening in the world. Detection of serial crime hotspots and the persons who involved in that would be a tedious process which needs to be analysed and processed in efficient manner. Inadequate man power and the voluminous data about the crimes happening in multiple locations reduce the efficiency of the investigation process. To address these issues, several attempts were made by various researchers using data mining approaches to make ease of investigation processes [1].

In this work, we attempt to find and group the similar kind of crimes happening at various locations in terms of crime features by using the kernel density estimation approach which overcomes the issues like grouping and overlapping problem faced by

I. Zelinka et al. (Eds.): ICSCS 2018, CCIS 837, pp. 253–265, 2018.
https://doi.org/10.1007/978-981-13-1936-5_28

modified graph cut clustering (MGCC) algorithm which was introduced in our previous work [2]. The social crime data is considered in this work along with the serial crime data set for the accurate prediction of the similar kind of crime features is done in terms of their characteristics. Social crime data would be in the unstructured format which must be pre-processed to handle it in an efficient manner. The latent semantic approach is employed to identify the most similar topics present in the crime data set. Hence this work can provide flexibility in the investigation processes involved in the serial crime detection.

2 Related Works

Susmita and Sharmistha [3] analysed the crimes that were happening against women at Tripura district using direct methods with multiple object. The fuzzy membership value was applied for the conversion of qualitative data into quantitative data. This work predicts the nature of crimes by gathering details from women in the Tripura district. Data mining approaches were utilized in this work for analysing the nature of crimes in terms of different risk factors. This research work concludes that the proposed methodology helps to deal with all type of data.

Clare et al. [4] analysed the risk factors of the serial killers in terms of their physiological behaviour. To do so, this approach implements the Neuro development methodology in which serial crimes would be identified in terms of interrelationship between different types of crimes which are located in different places. These interrelationships are identified based on the factors called the risk factors such as types of crime, level of crime effect, and so on. The neurological development approach leads to an improved finding of the serial crimes in terms of various risk factors associated with them.

Duygu and Murat [5] introduced the novel way of crime analysis process in the computer science student's point of view. This approach would lead to an improved finding of the various research methodologies. This analysis is conducted over the undergraduate students of the Trakya University where the finding of the work is demonstrated in the computer implemented programming language which was proved that the crime analysis was better than the manual analysis. This approach was employed to analyse the crime factors that are occurred in the environment in terms of the ethics and law of the crime factors.

Nabeela et al. [6] studied the socio-economic factors which are reason for the crime that are occurring in different locations of the Pakistan. The factors are which includes unemployment, educational qualification, poverty, and the economic growth. This study concludes that the government of Pakistan needs to concentrate on creating more job opportunities, alleviate poverty and promote education in order to reduce the rate of crime.

Omowunmi et al. [7] interrogate the crime situation and their behaviour by constructing the pattern model in terms of their location, time and the type of crime behaviour. The Author proposed an automatic threshold selection method based on quartile floor-ceiling functions. This approach is based on the pruning process which was done based on the Apriori techniques using which the unnecessary data samples would be eliminated from the data set, so that the crimes can be identified accurately. The author concluded with the result indicating that Revised Frequent Pattern Growth was more promising than Traditional FP-Growth model.

This section provides a detailed analysis about the different related works which has been conducted previously in terms of crimes that are happening in the different crime location in terms of their better provisioning of the resources. From these analyses, it is concluded that the serial crime is the most frequently occurred threat in the real world environment which need to be detected in order to provide the secured environment to the people.

3 Kernel Density Estimation Based Serial Crime Detection Approach

Crimes are the behaviour disorder which cost our society dearly in several ways. Those crimes needs to be spotted and the person committed must be identified to avoid further crime actions performed by same persons. The crime might happen in different form based on their structure and the characteristics. Two most important forms of crimes happening in different locations are "personal trait crime" and the "serial crime". Personal trait crimes are the one which would be done by individual for their personal reasons. Serial crimes are the one which are repetitive in nature done by the individual or group of people continuously in different places. Serial crimes are the most dangerous threat than the personal trait crime that needs to be identified for ensuring the security.

The serial crimes can be identified by finding the similar features present among the crimes characteristics that are happening in the different locations. The crime mapping will be effective if the crime locations are classified into different forms and that are shown in the separate regions in the visualization [8–10]. In our previous proposed research modified cut clustering approach was used for finding the serial crimes which would group the similar features that are present in the crime type of T. This approach works on structured serial crime data set but doesn't support well for unstructured data like social media data. This serial crime data set is gathered from the police department which will contain the police investigation parameters like crime details and behaviours which is not only enough for predicting the serial crimes in accurate manner.

In this proposed research work, Social crime data aware kernel density estimation based serial crime detection approach is introduced which would find the serial crimes by finding the similar features that reside in the crimes of type T happened in different spatial points p. Social media data's about crime behaviour are gathered from the multiple social media's which is used to predict the crime behaviour along with the serial crime data set for predicting the serial crime in an accurate manner. Social media data would be in the unstructured format which is pre-processed to eliminate the unnecessary words and represent the dataset in the structured format. After pre-processing, Latent Dirichlet Allocation (LDA) is applied to find the most important features present in the document based on their relevance. Then the kernel density would be estimated for the extracted features using LDA [11]. This kernel density is calculated by using the characteristics of the serial crime which can be identified from the serial crime data set.

Finally, the spatial point ‘p’ with more kernel density is taken as the hot spot location of the serial crime.

The Overall flow of the proposed research work is given in Fig. 1 as follows:

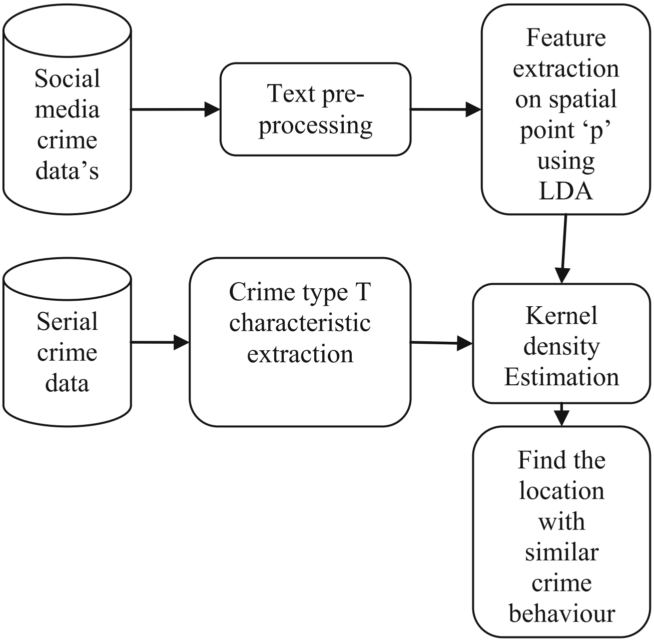
Social media crime data’s
Text pre-processing
Feature extraction on spatial point ‘p’ using LDA
Serial crime data
Crime type T characteristic extraction
Kernel density Estimation
Find the location with similar crime behaviour

Fig. 1. Flow of kernel density estimation based serial crime detection of different spatial point ‘p’.

The above flow diagram provides the steps in serial crime detection based on public comments on social media data in different locations. This process flow is explained in the following sub sections.

3.1 Data Collection

In this proposed research work, two types of data sets namely, Crime Data Set and Social Media Crime Data Set are considered for finding the serial crimes that are happening in different crime locations in terms of their similar characteristics.

Crime Data Set: The crime data considered in this work consists of attributes like different types of crimes for e.g., Murder, dacoity, robbery, theft and housing breaking etc., latitude and longitude values of crime which denotes the exact location where the crimes had happened. By using these information, the different types of crimes are analysed and finally the similar types of crimes are grouped together to predict the serial crimes [12–14].

Social Media Crime Data Set: The social media crime data sets are gathered from different public social media web sites such as Facebook, Twitter and so on which reveal the semantic meaning of the crime in terms of comments about different crimes at different locations. It would be in the unstructured format with various unnecessary tags and words. Pre-processing would be done to represent it in the structured format, so that semantic meaning of those data's can be obtained.

3.2 Text Pre-processing

Text pre-processing is the most essential process which is used to convert the unstructured data into structured format. It will also avoid the unnecessary words from the data set that are irrelevant to the concept. The social crime data's gathered from the online social media web sites would consists of the noises, html tags, advertisement messages and so on. This might cause the high dimensionality problem at the time of processing the social crime data set where every tag would be considered as one feature. The kernel density would be calculated for those features which are unnecessary, so that computation overhead would be increased. The text pre-processing would consist of the following steps:

Online Text Cleaning: Online text cleaning is the process of eliminating the online text characters such as html tags, scripting tags, and advertisement contents and so on. This step will remove all the contents that are irrelevant to the concepts.

White Space Removal: After performing online text cleaning white space removal would be done, so that the complete representation of the documents can be obtained. White space removal is used to provide the sentences in complete format, from which semantic meaning can be extracted in the flexible manner.

Expanding Abbreviations: In this step, abbreviation expansion would be done by parsing the sentences from start to end. By abbreviating the sentences, complete meaning of the document can be obtained.

Stemming: Stemming is the process of avoiding the derived words through which repetition is avoided. In this research work, porter stemming algorithm is used to perform stemming process which will reduce the inflected words from their base words.

Start and Stop Word Removal: Start and Stop word removal is the process of eliminating the Starting and ending words from the sentences from which meaning of the document cannot be extracted. Some of the start words are ‘a, an, the’ and so on. Some of the stop words that are considered in this work are ‘the, is, at, which’, which are removed before processing the social media crime data’s.

3.3 Feature Detection Using Latent Dirichlet Allocation

Latent dirichlet allocation is the natural language processing based generative model which is used to observe the more similar parts of the documents which are given as input. LDA provides better performance in finding the similarity in two ways. Those are word distribution and the topic distribution. This would be identified by taking crime type T detail from the crime data set which is compared with the documents. Here documents are public comments about the crime data's gathered from the social websites.

Word distribution and Topic distribution are defined as the probability of corresponding crime type T belongs to the words and topics of document D respectively. The Algorithm is discussed as follows:

Algorithm 1:

```
    Input: Social crime data set gathered in particular time period, Crime types
    Output: Features
    1. Repeat
    2. Collect the input documents
    3. Initialize weight values of all documents as null
    4. Find the log likelihood to find the words that are most related
                    $-2\ln\lambda = 2\sum_i O_i \ln\left(\frac{o_i}{E_i}\right)$          (1)
    5. Assign the temporary crime types for every word that has more log likelihood present in the user review
    5. a.  If any word is repeated multiple times assign with different crime types
    6. Find the similarity of the word with the corresponding crime types
    7. Find the similarity of the crime type with the user review.
    8. Update the weight values of reviews based similarity
    9. until final solution obtained
where
$-2\ln\lambda$ → Log likelihood ratio
$O_i$ → Observed Value
$E_i$ → Expected Value
```

The above algorithm provides a pseudo code of the latent Dirichlet allocation procedure which is used to find the most similar and relevant words that are present in the social crime data set in terms of different crime types represented in the serial crime data set. This procedure will assign the labelling of crime type for every word present in the document. This process would be done for every crime types that are considered in this research methodology. Finally all the words that are labelled using corresponding crime types T would be considered as the features f(p) of the crime types at particular spatial point p. These features would represent the serial crimes that are happening in the different location of the world in similar manner. Based on these features, the location in which these crimes are happening more would be identified by calculating the kernel density of those particular features.

3.4 Kernel Density Estimation of Features at Spatial Point 'P'

Kernel density estimation is defined in terms of statistics as the non-parametric way which is used to identify the probability density distribution of the corresponding feature in a particular location. In this research work KDE is used to find the density level of the features f(p) in the corresponding spatial location 'p' in terms of crime type T. KDE is one of the most improved data smoothening approach which can find the density estimation of the corresponding feature on the particular spatial location by omitting the noises present in the location. This smoothening is achieved in this work by setting the bandwidth parameter value of ‘h’ as optimal. This ‘h’ value would indicate the surface of data which need to be processed.

KDE estimation would be done for all feature values in terms of every crime types T within a particular time period. The equation that is used to estimate the kernel density of features at the spatial point ‘p’ for the crime types are calculated by using the formulae as like follows:

$$f(p) = k(p, h) = \frac{1}{Ph}\sum_{j=1}^{P} K\left(\frac{\left\|p - p_j\right\|}{j}\right) \tag{2}$$

where,

p → spatial point in which density to be calculated
h → bandwidth of KDE used to smoothen the data processing problem
P → total number of crime types T that are considered
j → single crime location during the time period
K → density function
$\|.\|$ → Euclidean distance

Density function is calculated in this research work by using the probability density function procedure which will find the probability of likelihood of occurrence of particular feature in the spatial point ‘p’ in terms of different crime types T.

The overall working flow of this kernel density based serial crime detection for the social crime data is given in the following algorithm.

Algorithm 2:

```
Input: Social crime data set, Crime types from serial crime data set, time window
Output: Location in which serial crimes are happened
1. Gather the serial crime data from different social web sites
2. Pre-process the data’s gathered from the social media web sites
  a. Remove the html tags, advertisement
  b. Remove the white spaces present in the data set
  c. Abbreviate the acronyms
  d. Remove the start and stop words from the data sets
  e. Apply porter stemmer algorithm to remove the stemming words
3. Find the most relevant features that are related to the crime types present in the data set by using LDA
4. For every features $f_i \in F$
5. Find the kernel density function k of features f in spatial point p for every crime type T
6.        $f(p) = k(p,h) = \frac{1}{Ph}\sum_{j=1}^{P} K$        (3)
7. End for
8. Return spatial point p with more kernel density
```

The above algorithm provides kernel density estimation calculation for the features at different spatial point p in the particular time window. This approach provides a better findings of the locations in which serial crimes are happened most in an accurate manner. This proposed research work is implemented in the matlab simulation environment in terms of performance measure values which is evaluated and compared with the existing approach which is discussed in detail as follows.

4 Experimental Results

SAKDESD is adapted for analysing and predicting the different number of features from the social crime data set which is most similar with the crime type T. This methodology is implemented in the matlab simulation environment and compared with the existing approach MGCC. This performance evaluation is done based on the metrics called as the mantel index and the jaccard index. This analysis is graphically represented in the proceeding sections. The results reveal that the proposed methodology produces better results than the existing one.

4.1 Data Set

Social crime data set obtained from multiple social media web sites in terms of different crimes activities happening in the different crime locations and the data gathered from the police stations of Coimbatore city, India are considered in this work for predicting the locations in which serial crimes are happening mostly in the real world environment. Both data set would consists of the details like crime type, number of crimes happening in given time period, people opinion about those particular crime happened in different location. By using these information, the different types of crimes are analyzed and finally the similar types of crimes are grouped together to predict the serial crimes.

4.2 Mantel Index

The mantel index is defined as metric used to calculate the correlation between the different features that are located as similar crime type in the geographical area. Mantel index of the proposed research approach should be high than the existing research approach which represents the high data correlation. The mantel index is calculated as follows:

$$r = \frac{1}{(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{\left(x_{ij} - \bar{x}\right)}{s_x} \cdot \frac{\left(y_{ij} - \bar{y}\right)}{s_y} \tag{4}$$

where,
x, y = variables measured at locations i, j
n = number of elements in the distance matrices
s_x, s_y = standard deviation of variable x and y
$\bar{x}, \bar{y}$ = mean value of variables x and y

The actual values obtained for the mantel index is given in the following Table 1.

Table 1. Mantel index values.

Number of data points	Mantel index	
	SAKDESD	MGCC
20	0.75	0.48
40	0.79	0.63
60	0.84	0.76
80	0.87	0.78
100	0.91	0.82
120	0.94	0.85
140	0.94	0.88
160	0.94	0.90
180	0.98	0.92
200	1	0.94

The graphical representation of the comparison of the proposed research work with the existing research work for the above mentioned actual values are depicted in the following Fig. 2.

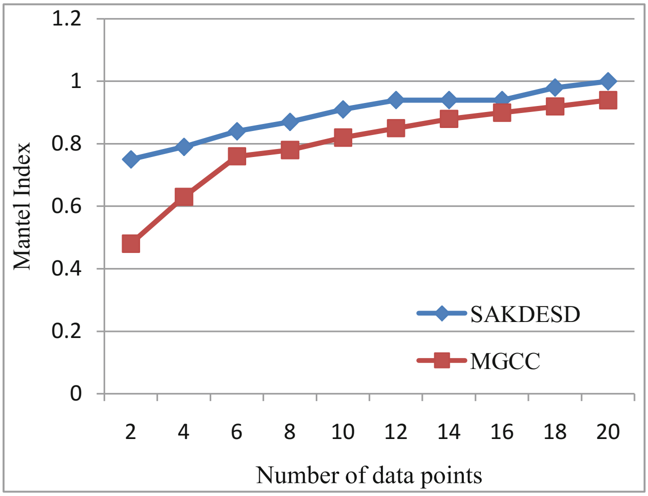

Fig. 2. Mantel index comparison.

In the above graph, mantel index values are evaluated and compared between the existing and proposed research scenarios. In the x axis, number of data points are taken where in the y axis mantel index values are taken. The analysis is done for varying number of data points which is taken in the range of 2 to 20. The mantel index value is increased linearly for increasing number of data points in both existing and proposed methodologies. From this graph it can be proved that the proposed research approach called SAKDESD provides better result than the existing approach called MGCC.

4.3 Jaccard Index

The jaccard index is used to represent the similarity between the data points. The jaccard index value is used to denote the number of crimes happened in different locations that are matched with each other. The jaccard coefficient measures similarity between finite sample sets, and is defined as the size of the intersection divided by the size of the union of the sample sets:

$$\mathrm{J(A, B)} = \frac{|\mathrm{A} \cap \mathrm{B}|}{|\mathrm{A} \cup \mathrm{B}|} \tag{5}$$

where

A, B = Data points

The actual values of Jaccard index obtained for both existing and proposed approach is indicated in the following Table 2.

Table 2. Jaccard index values.

Number of data points	Jaccard index	
	SAKDESD	MGCC
20	0.77	0.55
40	0.84	0.57
60	0.86	0.61
80	0.93	0.61
100	0.95	0.78
120	0.99	0.79
140	1	0.84
160	1	0.89
180	1	0.94
200	1	0.96

The graphical representation of the comparison of the proposed research work for the above mentioned actual values are depicted in the following Fig. 3.

In the above graph, jaccard index values are evaluated and compared between the existing and proposed research scenarios. In the x axis, number of data points are taken where in the y axis jaccard index values are taken. The analysis is done for varying number of data points which is taken in the range of 2 to 20. The Jaccard index value is increased linearly for increasing number of data points in both existing and proposed methodologies. From this graph it can be proved that the proposed research approach called SAKDESD provides better result than the existing approach called MGCC.

The GIS representation of clustered results of crimes which were happened in the various crime locations are depicted using MGCC and SAKDESD as follows:

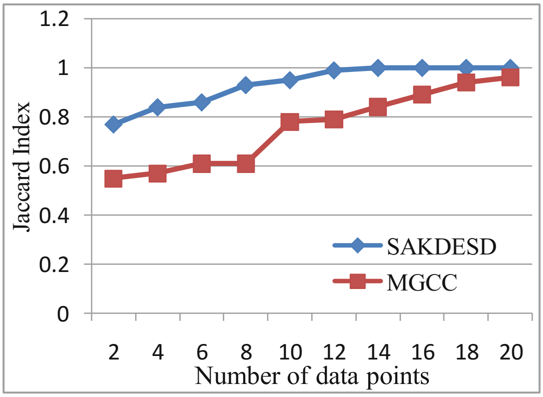

Fig. 3. Jaccard index comparison.

Figure 4 shows the clustering of crime spots performed using MGCC. The MGCC clusters crime spots effectively but the overlapping problem reduces the overall performance.

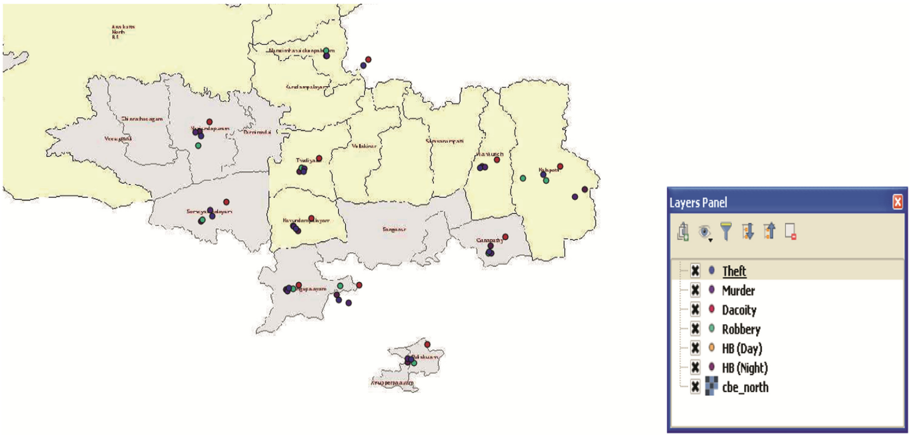

Fig. 4. Clustering crime spots using MGCC.

Figure 5 shows the clustering of crime spots performed using SAKDESD. The SAKDESD clusters crime spots effectively and has better performance than MGCC which is also proved in the graphical representation in terms of performance parameters.

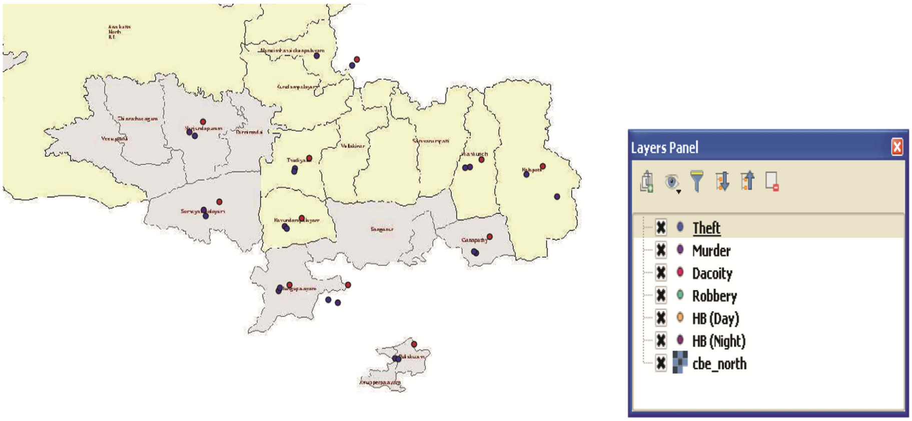

Fig. 5. Clustering crime spots using SAKDESD.

5 Conclusion

Serial crime detection is a major task in the police investigation department where it cannot be performed well manually. The proposed research attempts to find the location where the serial crimes are happening most in terms of the similarity with the crime types. This is done by using the novel approach called the Social crime data aware kernel density estimation based serial crime detection. Hence this approach would find the spatial points in which density of the similar features in terms of crime type T is more. The experimental tests conducted in the matlab simulation environment proves that the proposed research provides better result than the existing approach in terms of improved performance measures called the mantel index and the Jaccard index values.

References

1. Colleen, M.C.: Data mining and predictive analytics in public safety and security. IEEE Comput. Soc. **8**, 12–18 (2006)
2. Sivaranjani, S., Sivakumari, S.: A novel approach for serial crime detection with the consideration of class imbalance problem. Indian J. Sci. Technol. **8**, 1–9 (2015)
3. Susmita, R., Sharmistha, B.: Application of fuzzy-rough oscillation on the field of data mining (special attention to the crime against women at Tripura). Procedia Comput. Sci. **45**, 790–799 (2015). International Conference on Advanced Computing Technologies and Applications
4. Clare, S.A., Helen, M., Lucy, T., Philip, W., Christopher, G.: Neurodevelopmental and psychosocial risk factors in serial killers and mass murderers. Aggress. Violent Behav. **19**, 288–301 (2014)
5. Duygu, S., Murat, T.: The perception analysis of cyber crimes in view of computer science students. Procedia – Soc. Behav. Sci. **182**, 590–595 (2015)
6. Nabeela, K., Junaid, A., Muhammad, N., Khalid, Z.: The socio-economic determinants of crime in Pakistan: new evidence on an old debate. Arab Econ. Bus. J. **10**, 73–81 (2015)

7. Omowunmi, I., Antoine, B., Sonia, B.: A revised frequent pattern model for crime situation recognition based on floor-ceil quartile function. Procedia Comput. Sci. **55**, 251–260 (2015)
8. Tomoki, N., Keiji, Y.: Visualising crime clusters in a space-time cube: an exploratory data-analysis approach using space-time kernel density estimation and scan statistics. Trans. GIS **14**, 223–239 (2010)
9. Wang, D., Ding, W., Lo, H., Stepinski, T., Salazar, J., Morabito, M.: Crime hotspot mapping using the crime related factors-a spatial data mining approach. Appl. Intell. J. **4**, 772–781 (2013)
10. Devendra Kumar, T., Arti, J., Surbhi, A., Surbhi, A., Tushar, G., Nikhil, T.: Crime detection and criminal identification in India using data mining techniques. AI Soc. **30**, 117–127 (2015)
11. Matthew, S.G.: Predicting crime using Twitter and kernel density estimation. Decis. Support Syst. **61**, 115–125 (2014)
12. Christopher, R.H.: The dynamics of robbery and violence hot spots. Herrmann Crime Sci. **4**, 33 (2015)
13. Sivaranjani, S., Sivakumari, S.: Mitigating serial hot spots on crime data using interpolation method and graph measures. Int. J. Comput. Appl. **126**, 17–25 (2015)
14. Fitterer, J., Nelson, T.A., Nathoo, F.: Predictive crime mapping. Police Pract. Res. **16**(2), 121–135 (2015)