ISSN: 2229-6956 (ONLINE)
DOI: 10.21917/ijsc.2019.0270

A VIOLENT CRIME ANALYSIS USING FUZZY C-MEANS CLUSTERING APPROACH

M. Premasundari and C. Yamini
Department of Computer Science, Sri Ramakrishna College of Arts and Science for Women, India

Abstract

Clustering Techniques are the most significant method of grouping data points based on certain similarity. There are two ways in clustering techniques, namely hard and soft clustering. Traditional clustering approaches include grouping of each object to only one cluster. However, there are some cases that each object may belong to multiple partitions. Normally, healthcare and educational data encompass multiple clustering. Such multiple partitioning can be accomplished using overlapping clustering and soft or fuzzy clustering approaches. In this work, Fuzzy C-Means clustering model is applied for multiple clustering based on crime rates. The proposed multiple clustering model is evaluated using USArrests dataset and the results are useful to predict the high possibility of crime incidence by visualizing the crime analysis in various states in US.

Keywords:

Crime Analysis, Hard and Soft Clustering, Fuzzy C-Means Clustering, Overlapping Clustering.

1. INTRODUCTION

A crime is a punishable illegal act. Crimes like murder, assault, rape, robbery, etc. are growing extensively nowadays. The crime analysis is now very significant to prevent its occurrence. The crime rate is grouped based on their classes of offence so that it can be prevented. With the rapid growth in usage of automated systems to track crimes and recognize criminals, analysts join hands with law enforcement officers and detectives in solving crimes quickly. Criminology is the process of identifying crimes and criminal activities. The criminology technique aids to judge the criminals and assess the possibility of crime occurrence. The criminology department has been used in the proceedings of crime tracking ever since 1800 [9]. The Indian Government has taken an initiative to develop tools and applications for State and Central Police in relation with the National Crime Records Bureau (NCRB) [18]. Any research that can help in solving crimes faster will pay for itself and about 10% of the criminals commit about 50% of the crimes [22]. Crime analysis is a law enforcement task that involves the systematic analysis for detecting and exploring patterns and trends in crime and disorder. Information on patterns can help law enforcement organizations deploy resources in a more effective manner, and assist detectives in identifying and arresting suspects. Crime analysis also plays a role in contriving solutions to crime problems, and formulating crime prevention strategies. It can occur at various levels, including tactical, operational, and strategic. Crime analysts study crime reports, arrest reports, and police calls for service to identify emerging patterns, series, and trends as quickly as possible. They analyze these phenomena for all relevant factors, sometimes predict or forecast future occurrences, and issue bulletins, reports, and alerts to their agencies. Data mining came into existence in the middle of 1990’s as a powerful tool that is suitable for extracting useful information from huge datasets and find relationship between the attributes of data [23]. Data mining is the process of analyzing the large datasets to identify patterns and gain insights for decision making using various methods at the intersection of machine learning and statistics. Data mining tools allow to predict the future trends. Machine learning is the set of algorithms which are useful to learn from and make predictions on the data. Machine learning techniques include supervised, unsupervised and semi-supervised learning methods. The applications of machine learning incorporate medical diagnosis, fraud detection, information retrieval, market basket analysis, sentiment analysis, etc. The analysis of data can be performed through these machine learning or data mining techniques. Similarly crime analysis can also be done using the machine learning approaches. Moreover, several machine learning methods can be integrated to solve complex analytical problems.

In this study, the paper is organize as follows: section 2 describes the literature review. Section 2 gives the study of theoretical background about various clustering approaches. Section 4 illustrates the data and section 5 explains the proposed system architecture with its techniques. Section 6 shows the experimental results. Finally conclusion and future work is drawn in section 7.

2. LITERATURE REVIEW

In [1], crime cannot be predicted since it is neither systematic nor random and also predicted crime prone regions in India on a particular day by building a model using Bayes, Apriori and Decision trees.

Bezdek et al. [2] implemented the coding of fuzzy c-means algorithm in FORTRAN-IV, which generated fuzzy partitions and prototypes for any kind of numerical data and this is applicable to a wide variety of geostatistical data analysis problems.

Lu et al. [3] developed an intelligent diagnosis system to examine the defects of a solder bump using an improved fuzzy c-means clustering algorithm based on entropy weights by enhancing the thermal contrast between defective and good bumps.

Rashedi et al. [4] proposed the boosted hierarchical clustering ensemble method based on boosting which iteratively choose a new training set using a weighted random sampling to perform hierarchical clustering that results to a final aggregated cluster.

Zheng et al. [5] proposed a feasible, multiple clustering approach for text documents based on a frequent term model and also introduced WordNet as external knowledge to remove redundant results.

Sreedevi et al. [6] developed a data mining review procedure that can solve crimes faster and also explained the prediction of crime and various types of criminals based on crime data using several data mining methods.

Bharati et al. [7] proposed a predictive model using Chicago crime dataset with various classification techniques such as *k*-Nearest Neighbour, Support Vector Machine, decision trees and Bayesian methods etc. and KNN model predicts the type of crime with accuracy of 78.9%.

Aarathi et al. [8] proposed a schema that intends to raise a systematic opportunity for combining different elements or characteristics of crime and dealt with high throughput and low maintenance cost of providing database using Hadoop tools and graphically presented using R tool.

David et al. [9] aimed to work on a survey towards crime analysis, crime prediction and criminal identification using various supervised and unsupervised techniques and the survey focused on crime analysis using NLP based methods, Evidence-based methods, spatial Geo-location based methods, prisoner based methods, communication based methods and also quantitative analysis of crime analysis and prediction approaches.

Taha et al. [17] proposed a framework with a forensic analysis system named SIIMCO which can identify the influential members of a criminal organization and the immediate leaders of a given list of lower-level criminals. In this framework, a network is constructed to represent a criminal organization or crime incident reports and in the constructed network, a vertex represents the individual criminal and a link represents the relation between two criminals. The techniques proposed by the SIIMCO system overcome the incomplete and inconsistent limitations of most current methods. The datasets used in this system are Enron email corpus, DBLP and Nodobo mobile phone record datasets and SIIMCO system is evaluated by experimental comparison of CrimeNet Explorer and Log Analysis using precision, recall and Euclidean distance measurements.

An interactive interface with crime analysis tool for Crime Criminal Information System (CCIS) based on the decision support system and data mining techniques has been proposed [18] in order to carry out police activities efficiently. The crime analysis tool will identify crime hot spots and zones of a particular region with certain crime types based on the query and will also display the plot for particular type of crime specified in the query. The proposed tool can be integrated with latest visualization techniques such as Geographical Information System for better understanding of results and patterns.

In this paper [19], the authors presented a new framework with a theoretical model which applied clustering and classification techniques on real crime data collected by police in England and Wales from 1990 to 2011. This proposed work involves genetic algorithm for optimizing outlier detection operator parameters for crime prediction based on the spatial distribution of existing data and crime identification.

Bogahawatte et al. [20] developed an Intelligent Crime Investigation System (ICIS) which can identify criminal based on the evidence collected from the crime location by the use of clustering and classification techniques for effective crime investigation and criminal identification. Clustering is used to segment the crime patterns into groups based on available evidences and Naïve Bayes classification is used to predict possible suspects from the criminal records. The system is a multi-agent system managed with Java Beans for crime prediction and it makes easy to encapsulate the requested entities and returns it to the bean for exposing properties.

Agarwal et al. [21] used the rapid miner tool for analysing the crime rates and expectation of crime rate using K-Means Clustering algorithm. The crime analysis work includes the extraction of crime patterns, prediction of crime based on the spatial distribution of existing data and detection of crime. This analysis tracks homicide crime rates according to year-wise variations.

Donald [24] presented a Regional Crime Analysis Program (ReCAP) system which was designed as a computer application to help local police forces in the crime analysis and prevention of crime. ReCAP works in cooperation with the Pistol 2000 records management system which consists of all the crime information from a particular region. The research was mainly focused on the individual components of the system, including database, Geographical Information System (GIS), and data mining tools that can produce spatial mining results over the crime hotspots.

Rizwan et al. [25] performed the classification of crime data to predict the type of crime for different states of the United States of America. The real crime data collected from socio-economic data from 1990 US Census, law enforcement data from the 1990 US LEMAS survey and crime data from the 1995 FBI UCR is used in this research. The two different classification algorithms such as Naïve Bayesian and Decision Tree is used for predicting Crime type for different states in USA. The experimental results showed that Decision Tree outperformed Bayesian algorithm and achieved 84% accuracy approximately in predicting the type of crime for different states of USA.

Sharma [26] described a zero crime concept in the society by designing an enterprise application. Data mining concepts have been utilized for detecting the suspects and criminal activities. The main objective of the paper was to detect the suspicious e-mails about the criminal activities by applying the ID3 algorithm with enhanced feature selection method and attribute-importance factor to produce a better and faster decision tree based on explicit information entropy. The author has used text processing techniques to classify email messages or posts into suspicious or non-suspicious records.

Chen et al. [27] proposed a general framework for crime data mining that pulls on experience procured with the Coplink project with the researchers at Arizona. The author used a concept space approach which will extract criminal from the incident abstracts. The work mainly emphases on displaying the relationships between crime types and the link between the criminal organizations.

3. CLUSTERING APPROACHES

Clustering is the task of segregating data objects into a number of partitions such that data objects in the same partitions are more similar. In simple words, the aim is to isolate groups with similar behaviors and assign them into clusters. The different methods of clustering approaches include:

- Partitioning methods

- Hierarchical clustering
- Fuzzy clustering
- Density-based clustering
- Model-based clustering

It is a common task of exploratory data mining for statistical data analysis used in many fields including machine learning, pattern recognition, image analysis, information retrieval, Bioinformatics, data compression, and computer graphics. The fine distinctions of clustering is hard, soft and overlapping clusters.

3.1 HARD CLUSTERING

Each data point must belong to only one cluster. In non-fuzzy clustering (also known as hard clustering), data is divided into distinct clusters, where each data point can only belong to exactly one cluster.

3.2 SOFT CLUSTERING

Each data point belongs to more than one cluster to a certain degree. This is also known as fuzzy clustering. Fuzzy clustering (also referred to as soft *k*-means) is a form of clustering in which each data point can belong to more than one cluster. In fuzzy clustering, data points can potentially belong to multiple clusters.

3.3 OVERLAPPING CLUSTERS

Each data point belongs to more than one cluster usually involving hard clusters. This is also called alternative or multi-view clustering.

4. DATA DESCRIPTION

The US Arrests is an R's own dataset [10] - [12] which contains statistics, in arrests per 100,000 residents for assault, murder, and rape in each of the 50 US states in 1973. The percent of the population living in urban areas is also given. The data frame consists of 50 observations on 5 variables. The Table.1 shows the data frame.

Table.1. shows the data frame

Variables	Datatype	Description
States	Character	States in US
Murder	Numeric	Murder arrests (per 100,000)
Assault	Numeric	Assault arrests (per 100,000)
UrbanPop	Numeric	Percent urban population
Rape	Numeric	Rape arrests (per 100,000)

5. PROPOSED ARCHITECTURE

In this study, a systematic approach of fuzzy clustering has been proposed for the analysis of crime rates. The proposed system architecture is shown in Fig.1. Fuzzy c-means clustering technique have been used to group the crime into multiple clusters. These multiple clusters can be useful to predict the places or states with high crime rates so that the identified US states can be given additional civil force for the prevention of crime through which the protection of people and public order can be maintained.

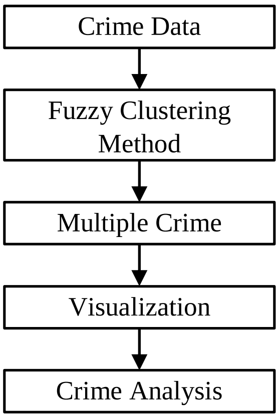

Fig.1. Proposed System Architecture

Fuzzy C-Means clustering is applied on crime data to create multiple clusters. Those multiple clusters can be visualized using factoextra visualization R package. This cluster visualization helps in analysing the crime prone states in US.

5.1 FUZZY C-MEANS CLUSTERING

Fuzzy C-Means (FCM) is a method of clustering which allows one piece of data to belong to two or more clusters. This method is developed by Dunn [13] and improved by Bezdek [14]. It is frequently used in pattern recognition. It is based on minimization of the following objective function:

$$J_m = \sum_{i=1}^{N}\sum_{j=1}^{C} u_{ij}^{m} \left\| x_i - c_j \right\|^2 ,\ 1 \le m < \infty \tag{1}$$

where, m is any real number greater than 1, u_{ij} is the degree of membership of x_i in the cluster j, x_i is the ith of d-dimensional measured data, c_j is the d-dimension center of the cluster, and ||*|| is any norm expressing the similarity between any measured data and the center.

This algorithm works by assigning membership to each data point corresponding to each cluster center on the basis of distance between the cluster center and the data point. More the data is near to the cluster center more is its membership towards the particular cluster center.

The algorithm is composed of the following steps [15]:

Step 1: Initialize $U = [u_{ij}]$ matrix, $U^{(0)}$

Step 2: At k-step, calculate the centers vectors $C^{(k)}=[c_j]$ with $U^{(k)}$

$$c_j = \frac{\sum_{i=1}^{N} u_{ij}^{m} x_i}{\sum_{i=1}^{N} u_{ij}^{m}} \tag{2}$$

Step 3: Update $U^{(k)}$, $U^{(k+1)}$

$$u_{ij} = \frac{1}{\sum_{k=1}^{C} \left(\frac{\left\| x_i - c_j \right\|}{\left\| x_i - c_k \right\|} \right)^{\frac{2}{m-1}}} \tag{3}$$

Step 4: *If* $||U^{(k+1)} - U^{(k)}|| < \varepsilon$, then STOP; otherwise return to step 2.

Step 5: The Fuzzy partitioning [15] is carried out through an iterative optimization of the objective function shown in Eq.(1), with the update of membership u_{ij} and the cluster centers c_j by using Eq.(3) and Eq.(2)

Step 6: This iteration will stop when

$$\max_{ij}\left\{\left|u_{ij}^{(k+1)}-u_{ij}^{(k)}\right|\right\}<\varepsilon,$$

where, ε is a termination criterion between 0 and 1, whereas k is the iteration steps. This procedure converges to a local minimum or a saddle point of J_m.

5.2 MEMBERSHIP

Membership grades are assigned to each of the data points. These membership grades indicate the degree to which data points belong to each cluster. Thus, points on the edge of a cluster, with lower membership grades, may be in the cluster to a lesser degree than points in the center of cluster.

5.2.1 Advantages:

- It provides the best result for overlapped dataset and comparatively better than *k*-means algorithm.
- Unlike *k*-means, each data point is assigned membership to each cluster center as a result of which data point may belong to more than one cluster [16].

5.2.2 Disadvantages:

- Predictive description of the number of clusters.
- It provide the better result, however at the cost of the number of iterations.
- Euclidean distance measures can unequally weight underlying factors [16].

6. EXPERIMENTAL RESULTS

To evaluate the proposed approach, the experiments were done on the platform with AMD A8-6410 APU with AMD Radeon R5 Graphics at 2GHz, 4GB RAM, Windows 8.1 64-bit OS. The approach is tested using the USArrests data in RStudio1.0.143. This dataset consists of 3 clusters, namely Murder, Assault and Rape. Fuzzy c-means clustering model is performed on the data to yield three multiple partitions. After the model is built using above proposed process, prediction of US states that has most and least murder, assault and rape arrests are done. The Table.2 represents the states with maximum and minimum crime rates.

Table.2. States that has most and least murder, assault and rape arrests.

Crime	States with most arrests	States with least arrests
Murder	Georgia	North Dakota
Assault	North Carolina	North Dakota
Rape	Nevada	North Dakota

The US states which are in the top 25% of the murder, assault and rape crimes are shown in the Table.3-Table.5 respectively.

Table.3. Top 25% States with Murder case

Murder	UrbanPop	Murder
Alabama	58	13.2
Florida	80	15.4
Georgia	60	17.4
Louisiana	66	15.4
Maryland	67	11.3
Michigan	74	12.1
Mississippi	44	16.1
Nevada	81	12.2
New Mexico	70	11.4
North Carolina	45	13
South Carolina	48	14.4
Tennessee	59	13.2
Texas	80	12.7

Table.4. Top 25% States with Assault case

States	UrbanPop	Assault
Alaska	48	263
Arizona	80	294
California	91	276
Florida	80	335
Maryland	67	300
Michigan	74	255
Mississippi	44	259
Nevada	81	252
New Mexico	70	285
New York	86	254
North Carolina	45	337
South Carolina	48	279

Table.6. Top 25% States with Rape case

States	UrbanPop	Rape
Alaska	48	44.5
Arizona	80	31
California	91	40.6
Colorado	78	38.7
Florida	80	31.9
Maryland	67	27.8
Michigan	74	35.1
Missouri	70	28.2
Nevada	81	46
New Mexico	70	32.1
Oregon	67	29.3
Tennessee	59	26.9
Washington	73	26.2

6.1 CRIME VISUALIZATION

This visualization section deals with the analysis of USArrests data and plotting them into various graphs like scatter plot, histogram and bar. The graph analysis is as follows:

- The cluster correlation plot is given in Fig.2.
- A Multiple cluster visualization graph is shown in Fig.3.
- Murder rate in US states.
- Assault crime rate in various states.
- Details of Rape crimes committed in states.
- Correlation between crime variables.

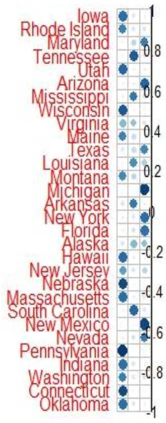

Fig.2. Correlation Plot for the Clusters

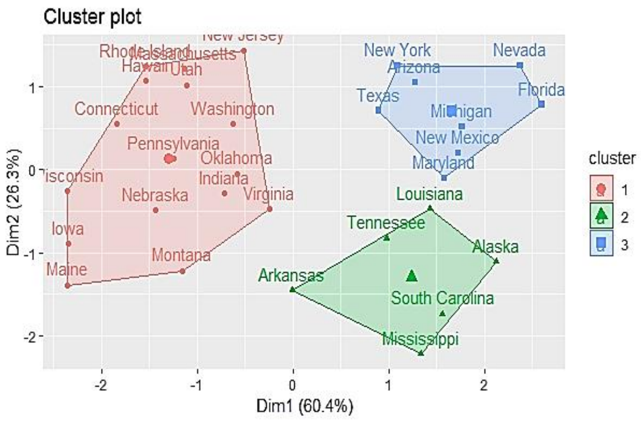

Fig.3. Cluster Visualization Graph

This graph in Fig.4 shows the states in which murder crimes have occurred most. The *x*-coordinate denotes the states in US and *y*-coordinate denotes the murder rate.

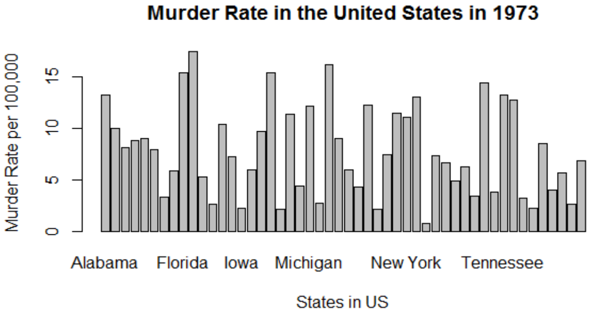

Fig.4. States vs. Murder Rate

The Fig.5 shows the states in which assault crimes have occurred most. The *x*-coordinate denotes the states in the US and *y*-coordinate denotes the assault rate per 100000.

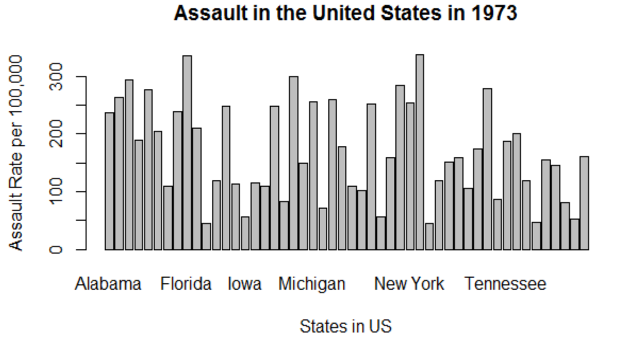

Fig.5. US States vs. Assault Rate

The Fig.6 shows the number of rape crimes have occurred. The *x*-coordinate denotes the rape and *y*-coordinate denotes the frequency of crime with high rates.

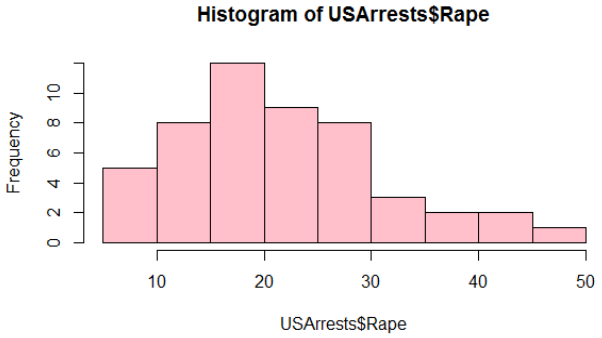

Fig.6. Frequency of rape with high rates

The scatter plot in Fig.7 shows the correlation between crime variables in USArrests data. The various plots in this graph represents the correlation between Murder Vs Assault, Murder Vs UrbanPop and Murder Vs Rape.

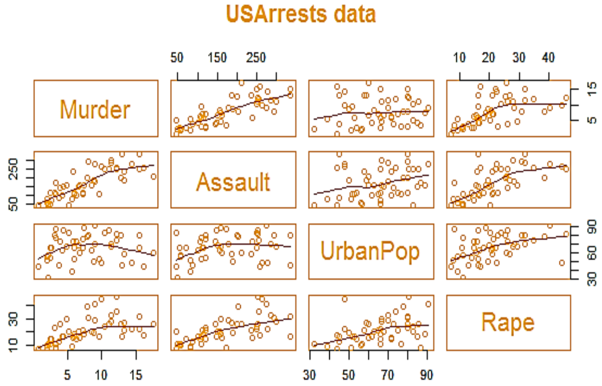

Fig.7. Correlation between crime variables

7. CONCLUSION

In this paper, a multiple clustering approach is proposed based on fuzzy clustering theory. The FCM algorithm works how an individual data point has been grouped in the multiple clusters. The fuzzy membership value is assigned for reweighting samples.

Completing all iterations, an aggregation of all created multiple clustering forms the final result. The final results are used to analyze the crime prone states in US so that it can be stopped by enhancing the security level in those regions. The results are only helpful for crime analysis but there is a requisite of analyzing the crime patterns that can occur in future. The prediction of crimes is impossible, but it can be prevented if the time in which the crime is going to happen is known. In future, the pattern analysis of imminent crime can be performed using association rule mining along with proposed system. Moreover, the work can be extended to predict the time in which crime may happen.

REFERENCES

[1] S. Sathyadevan, M.S. Devan and S.S. Gangadharan, "Crime Analysis and Prediction Using Data Mining", *Proceedings of IEEE 1st International Conference on Networks and Soft Computing*, pp. 406-412, 2014.

[2] J.C. Bezdek, R. Ehrlich and W. Full, "FCM: The Fuzzy C-Means Clustering Algorithm", *Computer and Geosciences*, Vol. 10, No. 2-3, pp. 191-203, 1984.

[3] L. Xiang Ning, S. Tie Lin , W. Su Ya, L. Li Yi , S. Lei and L. Guang Lan, "Intelligent Diagnosis of the Solder Bumps Defects using Fuzzy C-Means Algorithm with the Weighted Coefficients", *Science China Technological Sciences*, Vol. 58, No. 10, pp. 1689-1695, 2015.

[4] E. Rashedi and A. Mirzaei, "A Novel Multi-Clustering method for Hierarchical Clusterings, Based on Boosting", *Proceedings of 19th IEEE Iranian Conference on Electrical Engineering*, pp. 1-4, 2011.

[5] H.T. Zheng, H. Chen, S.Q. Gong, "A Frequent Term-Based Multiple Clustering Approach for Text Documents", *Proceedings of Asia-Pacific Web Conference on Web Technologies and Applications*, pp. 602-609, 2014.

[6] M. Sreedevi, A. Harsha Vardhan Reddy and C.H. Venakata Sai Krishna Reddy, "Review on Crime Analysis and Prediction using Data Mining Techniques", *International Journal of Innovative Research in Science, Engineering and Technology*, Vol. 7, No. 4, pp. 3360-3369, 2018.

[7] A. Bharati and R.A.K. Sarvanaguru, "Crime Prediction and Analysis using Machine Learning", *International Research Journal of Engineering and Technology*, Vol. 5, No. 9, pp. 1037-1042, 2018.

[8] S.N. Aarathi, N. Gayathri, R. Indraja , S. Srividhya and J. Kayalvizhi, "Crime Analysis and Prediction using Big Data", *International Journal of Pure and Applied Mathematics*, Vol. 119, No. 12, pp. 207-211, 2018.

[9] H.B.F. David and A. Suruliandi, "Survey on Crime Analysis and Prediction using Data Mining Techniques", *ICTACT Journal on Soft Computing*, Vol. 7, No. 3, pp. 1459-1466, 2017.

[10] USArrests, Available at: https://www.kaggle.com/deepakg/usarrests

[11] Datasets distributed with R Git Source Tree, Available at: https://stat.ethz.ch/R-manual/R-devel/library/datasets/html/USArrests

[12] J.C. Dunn, "A Fuzzy Relative of the ISODATA Process and its use IN Detecting Compact Well-Separated Clusters", *Journal of Cybernetics*, Vol. 3, No. 3, pp. 32-57, 1973.

[13] J.C. Bezdek, "*Pattern Recognition with Fuzzy Objective Function Algoritms*", Plenum Press, 1981.

[14] A Tutorial on Clustering Algorithms, Available at: https://sites.google.com/site/dataclusteringalgorithms/fuzzy-c-means-clustering-algorithm.

[15] K. Taha and P.D. Yoo, "SIIMCO: A Forensic Investigation Tool for Identifying the Influential Members of a Criminal Organization", *IEEE Transactions on Information Forensics and Security*, Vol. 11, No. 4, pp. 811-822, 2016.

[16] M. Gupta, B. Chandra and M.P. Gupta, "Crime Data Mining for Indian Police Information System", *Journal of Crime*, Vol. 2, No. 6, pp. 43-54, 2006.

[17] R. Kiani, S. Mahdavi and A. Keshavarzi, "Analysis and Prediction of Crimes by Clustering and Classification", *International Journal of Advanced Research in Artificial Intelligence*, Vol. 4, No. 8, pp. 11-17, 2015.

[18] K. Bogahawatte and S. Adikari, "Intelligent Criminal Identification System", *Proceedings of 8th IEEE International Conference on Computer Science and Education*, pp. 633-638, 2013.

[19] J. Agarwal, R. Nagpal and R. Sehgal, "Crime Analysis using K-Means Clustering", *International Journal of Computer Applications*, Vol. 83, No. 4, pp. 1-4, 2013.

[20] S.V. Nath, "Crime Pattern Detection using Data Mining", *Proceedings of IEEE International Conference on Web Intelligence and Intelligent Agent Technology*, pp. 1-4, 2006.

[21] D. J. Hand, H. Mannila and P. Smyth, "*Principles of Data Mining*", MIT Press, 2001.

[22] D. E. Brown, "The Regional Crime Analysis Program (RECAP): A Framework for Mining Data to Catch Criminals", *Proceedings of IEEE International Conference on Systems, Man, and Cybernetics*, pp. 2848-2853, 1998.

[23] R. Iqbal, M.A.A. Murad, A. Mustapha, P.H.S. Panahy and N. Khanahmadliravi, "An Experimental Study of Classification Algorithms for Crime Prediction", *Indian Journal of Science and Technology*, Vol. 6, No. 3, pp. 4219-4225, 2013.

[24] M. Sharma, "Z-Crime: A Data Mining Tool for the Detection of Suspicious Criminal Activities based on the Decision Tree", *Proceedings of International Conference on Data Mining and Intelligent Computing*, pp. 1-6, 2014.

[25] H. Chen, W. Chung, J.J. Xu, G. Wang, Y. Qin and M. Chau, "Crime Data Mining: A General Framework and Some Examples", *Computer*, Vol. 37, No. 4, pp. 50-56, 2004.