Multi Clustering Technique in Data Mining for Crime Scene Investigation

Carlos D. Correa and Chih-Hsing Chu

***Abstract*---** The retrieval of the informative knowledge is quite interesting, from the huge data and it is also a challenging task. Data mining is the powerful technology to analyse the data skilfully from different perspectives and summarizes it to useful information. The finding of the cold spot and hot spot is the challenging task in crime analysis. Crime is the act that harms the public, increases the violence, demolishes the assets and denies the respect to people. A major challenge facing all law-enforcement and intelligence-gathering organizations is accurately and efficiently analyzing the growing volumes of crime data. Another major challenge faced by the Nigerian law enforcement agencies is the lack of a central repository where all data collections concerning crimes and criminals are stored which possess a bigger problem, as there are many cases of data repetition, and as such it is difficult for law enforcers to see patterns in crimes during analysis. So In this paper crime analysis is done by performing clustering on crime dataset using rapid miner tool.

***Keywords*---**Crime Scene Analysis, High Dimensional Data Clustering.

I. 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.

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, arrests 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.

Manuscript received on August 09, 2019, review completed on August 09, 2019 and revised on August 13, 2019.

Carlos D. Correa is with the Department of Electrical Engineering, Graduate Institute of Communication Engineering, National Chung Hsing University, Taichung 402, Taiwan.

Chih-Hsing Chu is with the Department of Electrical Engineering, Graduate Institute of Communication Engineering, National Chung Hsing University, Taichung 402, Taiwan.

Digital Object Identifier: AIML082019003.

In the research article [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 clusters. 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.

II. Clustering Techniques

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 behaviours 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 clustering.

A. 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.

B. 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.

C. Overlapping Clusters

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

III. DATA DESCRIPTION

The USArrests is a R's own dataset [6,7,8] 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. Table 1 shows the data frame.

S No	Variables	Datatype	Description
1	States	Character	States in US
2	Murder	Numeric	Murder arrests (per 100,000)
3	Assault	Numeric	Assault arrests (per 100,000)
4	UrbanPop	Numeric	Percent urban population
5	Rape	Numeric	Rape arrests (per 100,000)

IV. 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 in 1973 [9] and improved by Bezdek in 1981 [10]. It is frequently used in pattern recognition. It is based on minimization of the following objective function:

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

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. Fuzzy partitioning [11] is carried out through an iterative optimization of the objective function shown above, with the update of membership u_{ij} and the cluster centers c_j by:

$$u_{ij} = \frac{1}{\sum_{k=1}^{C}\left(\frac{\left\|x_i - c_j\right\|}{\left\|x_i - c_k\right\|}\right)^{\frac{2}{m-1}}}, \quad c_j = \frac{\sum_{i=1}^{N} u_{ij}^{m}\cdot x_i}{\sum_{i=1}^{N} u_{ij}^{m}}$$

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.
The algorithm is composed of the following steps [11]:

1. *Initialize $U=[u_{ij}]$ matrix, $U^{(0)}$*
2. *At k-step: calculate the centers vectors $C^{(k)}=[c_j]$ with $U^{(k)}$*

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

3. *Update $U^{(k)}$, $U^{(k+1)}$*

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

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

Advantages

1. Gives the best result for overlapped data set and comparatively better than k-means algorithm.
2. 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 [12].

Disadvantages

1. Predictive description of the number of clusters.
2. Provide the better result however at the cost of more number of iteration.
3. Euclidean distance measures can unequally weight underlying factors [12].

V. 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 Figure1.

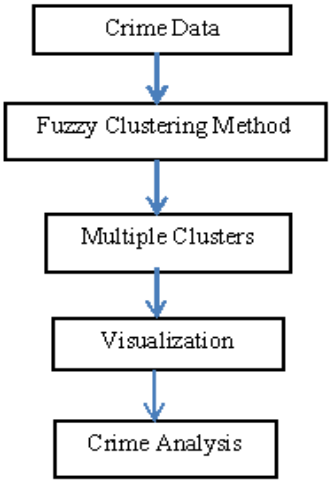

Fig 1. System Architecture

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.

VI. 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 is performed on the data to yield three multiple partitions. The cluster correlation plot and visualization graph is given in Figure2 and 3 respectively.

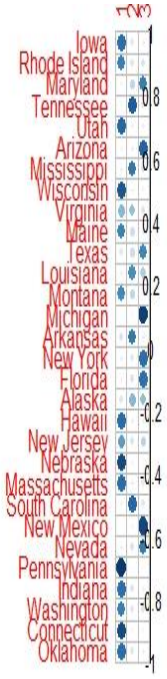

Fig 2. Correlation Plot for the Clusters

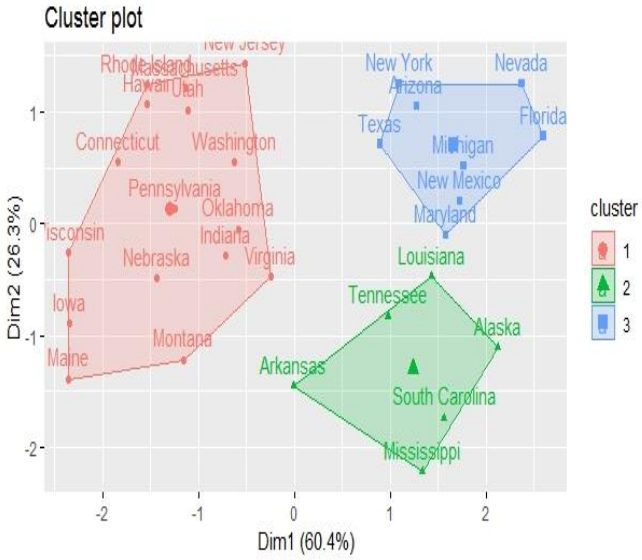

Fig 3. Cluster Visualization Graph

VII. 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 clusterings 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] Cannon, Robert L., Jitendra V. Dave, and James C. Bezdek. "Efficient implementation of the fuzzy c-means clustering algorithms." IEEE transactions on pattern analysis and machine intelligence 2 (1986): 248-255.

[2] Chen, Hsinchun, et al. "Crime data mining: a general framework and some examples." computer 4 (2004): 50-56.

[3] Deshpande, Anjali R., and L. M. R. J. Lobo. "Text summarization using Clustering technique." International Journal of Engineering Trends and Technology 4.8 (2013): 3348-3351.

[4] Dou, Dejing, Hao Wang, and Haishan Liu. "Semantic data mining: A survey of ontology-based approaches." Proceedings of the 2015 IEEE 9th international conference on semantic computing (IEEE ICSC 2015). IEEE, 2015.

[5] Horng, Shi-Jinn, et al. "A novel intrusion detection system based on hierarchical clustering and support vector machines." Expert systems with Applications 38.1 (2011): 306-313.

[6] Hu, Zhengbing, et al. "Fuzzy clustering data given on the ordinal scale based on membership and likelihood functions sharing." arXiv preprint arXiv:1702.01200 (2017).

[7] Jain, Anil K. "Data clustering: 50 years beyond K-means." Pattern recognition letters 31.8 (2010): 651-666..

[8] Kriegel, Hans-Peter, Peer Kröger, and Arthur Zimek. "Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering." ACM Transactions on Knowledge Discovery from Data (TKDD) 3.1 (2009): 1.

[9] McCue, Colleen. Data mining and predictive analysis: Intelligence gathering and crime analysis. Butterworth-Heinemann, 2014.

[10] Pal, Nikhil R., et al. "A possibilistic fuzzy c-means clustering algorithm." IEEE transactions on fuzzy systems 13.4 (2005): 517-530.

[11] Peizhuang, Wang. "Pattern recognition with fuzzy objective function algorithms (James C. Bezdek)." SIAM Review 25.3 (1983): 442.

[12] Su, Lei, et al. "Intelligent diagnosis of flip chip solder bumps using high-frequency ultrasound and a naive Bayes classifier." Insight-Non-Destructive Testing and Condition Monitoring 60.5 (2018): 264-269.

[13] Wang, Yang, et al. "Robust subspace clustering for multi-view data by exploiting correlation consensus." IEEE Transactions on Image Processing 24.11 (2015): 3939-3949.

[14] Wang, Yangtao, Lihui Chen, and Jian-Ping Mei. "Incremental fuzzy clustering with multiple medoids for large data." IEEE transactions on fuzzy systems 22.6 (2014): 1557-1568.

[15] Wei, Wei, et al. "Using active thermography and modified SVM for intelligent diagnosis of solder bumps." Infrared Physics & Technology 72 (2015): 163-169.