Crime Data Analysis Using Clustering by Fast Search and Find of Density Peaks

Ahmed Alghamdi

College of Computer Science and Engineering, University of Jeddah, KSA

Summary
Analyzing crime data is becoming a hot area of research because it has a direct connection to human life. This requires to discover hidden patterns in the crime data and group (or classify) them accordingly. Clustering is a discovery process that groups datasets into different categories based on their similarities. Clustering has been used in various areas like computing and IT, business, medicine, and biology. We used clustering by fast search and find of density peaks (CFSFDP) algorithm on crime data set. CFSFDP algorithm is based on two assumptions: (1) cluster center is a higher density data point as compared to other neighboring data points; and (2) cluster centers lies at large distance from each other. Unlike k-mean clustering algorithm, number of clusters is automatically formed by CFSFDP algorithm; however, cluster centers are manually selected by the user from decision graph. We performed experiments on crime dataset and tested different number of clusters to evaluate the performance of our approach.

Key words:
Crime data analysis; Clustering; Decision graph; Information security.

1. Introduction

With the advent of new technology and better facilities for human being, the crime rate is also increasing due to numerous factors like unemployment, thrill, jealousy, frustration and other issues that a human faces in daily life. A recent study in [20] reveals there is a certain correlation between health disparities and crimes. For instance, a human who was a victim of a crime might be more vulnerable to anxiety than a human who have never been a victim to crime. Therefore, there is a need to analyze the various factors related to crime. In the recent past, various efforts have been made to address this important issue [21, 22, 23, 24, 25, 26]. One such issue the grouping (clustering) of the crime data using appropriate techniques.

Clustering is a technique in which data points from a specified data set is grouped into groups called clusters on the basis of their similarities. The role of clustering is important in different fields of study like biology; information retrieval; climate; psychology and medicine; and business [1]. In computer science clustering techniques are widely used in bioinformatics; image processing; social networks and cyber security [2]. Summarization, comparison and efficiently finding nearest neighbors are few applications of clustering.

Clusters can be formed using different techniques, this might be the reason there is no agreed upon definition of clustering. There are different types of cluster such as well-separated; prototype-based; graph-based; and density-based. Some of the techniques/ algorithms that can be used to form clusters are: (1) K-mean; (2) Agglomerative Hierarchical Clustering and (3) DBSCAN. K-mean is prototype-based technique which only works with numeric data. K-mean partitioned n data points into user specified number of classes i.e. k. These points are represented by their centroids. The knowledge about the number of clusters and selection of centroids is essential for the effectiveness of k-mean. Hierarchical clustering technique is a better quality approach as compared to other clustering techniques. However, it has a drawback of its quadratic time complexity. Agglomerative Hierarchical Clustering can be applied using three different techniques i.e. intra-cluster similarity; centroid similarity; and UPGMA [3]. DBSCAN is a density-based clustering technique which produce a partitioned clustering. In contrast to k-mean, DBSCAN algorithm automatically determined number of clusters. Those data points that lies in low-density regions are classified as noise and eliminated.

A clustering algorithm based on density CFSFDP was proposed by Rodriduez and Laio in [4]. CFSFDP algorithm has both similarities with k-mean and DBSCAN algorithms. Its similarity with k-means algorithm is that it works on the distance between data points while it has the ability to detect non-spherical clusters and automatically find the number of clusters like DBSCAN [5]. The advantage of CFSFDP is that there is a minimum human interaction in clusters formation while its disadvantage is human based selection of cluster center.

In this work, we used Washington DC crime data download from DC government website and run a CFSFDP in MATLAB to evaluate the performance of the algorithm.

The rest of the paper is organized as follows. Section 2 presents the related work. Section 3 presents the detail about our approach methodology 3. Section 4 describes experiments and results, and finally, the concluding remarks have been presented in Section 5.

Manuscript received November 5, 2019
Manuscript revised November 20, 2019

2. Related Work

The authors in [10] proposed to use the environmental design for controlling the crime but they do present any crime prevention model. Another work in [11] used Cross-Entropy (CE) algorithm and greedy search techniques for crime data analysis; however, this technique is generally not feasible for real world applications. To overcome some of its shortcomings, in [12], the authors improved the accuracy but again fail to explain how their work is useful for solving real world problems. Another interesting work in [13] used genetic algorithms (GA) for crime prevention but this technique fails to address non-deterministic behavior of criminals.

A technique in [14] used artificial neural networks (ANN) to predict the crime but fail to factor out the parameter that play vital role in crime prediction. Similarly, the technique in [15] suffers to work with large datasets. The work in [16] requires the biography data of the suspect for prediction and analysis of crime data but fails to do so if the biography data is not available. A data-driven approach in [17] analyzes the crime data for the relationship between the time and space to get an in-depth knowledge of the correlation between them. A linear regression-based technique in [18] works for predicting the number of crimes only. The Apriori algorithm was used in [19] for dimension reduction in predicting the crime. In [9], authors use deep learning for predictive analysis of crime data.

Almost all of the above mentioned works do not work for predicting a new type of crime. While on the other hand, the proposed technique uses clustering through CFSFDP algorithm for analyzing the crime data and predicting the new crimes using the existing data.

3. Proposed Work

As mentioned earlier, in order to analyze crime data, we used CFSFDP proposed by Rodrinquez and Laio with the aim to explore the applicability of crime data to the said clustering algorithm.

Clustering by fast search algorithm has similarities with K-means and DBSCAN because it works on distance between data points and also has the ability of detecting non-spherical correct number of clusters. In this algorithm, cluster centers are defined as local maxima in the density of data points like mean-shift method. However, the difference between mean-shift method and clustering by fast search method is that in this method data is not embedded in vector space which actually maximize the density field for each data point [6].

There are two basic assumptions in clustering by CFSFDP. First, cluster centers are surrounded by neighboring data points which have lower local density; secondly, these cluster centers are far from any point with a higher local density. For every single point let's say i, two quantities are computed i.e. local density denoted by pi and its distance from points of higher density di. Both of these calculated quantities depend on distance dij between data points. These quantities need to satisfy the triangular inequality.

To define the local density pi of data point, equation (1) is used as:

$$Pi = \sum_j {}^{x}(dij- dc) \qquad (1)$$

In the above expression, X(x) = 1 if x<0 and X(x) = 0 otherwise, while dc is a cutoff distance and pi shows the number of points that are closer to dc than to point i.

δ is calculated by measuring the minimum distance between points i and any other point with higher density.

$$\delta = \min (dij) \qquad (2)$$

However, for the highest density point, δ will be calculated as δi = maxi (dij). The parameter δi would have larger value than nearest neighbor distance for the points that are local or global maxima in the density. Similarly, δi will have a value greater than the nearest neighbor distance only for local or global maxima in the density. Consequently, the very large value of δi of certain points makes them to appear as cluster center (see Figure 1).

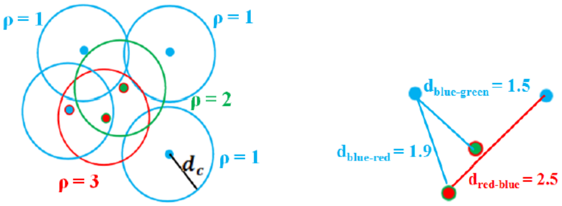

Fig. 1 ρ and δ - Schematic representation of how ρi and δi is calculated [7]

Figure 2 below demonstrates the core of clustering by fast search algorithm. In Figure 2(a) there are 28 pints shown in two dimensional space. It is clear from the Figure that point 1 and 10 are representing density maxima and identify as cluster centers. Figure 2(b) is decision graph that represent the plot of δi as a function of pi for each point. It is clear from the decision graph that point 9 and 10 has similar value of p but the value of $^{\delta}$ is still different for these two points and as a result both are in different clusters.

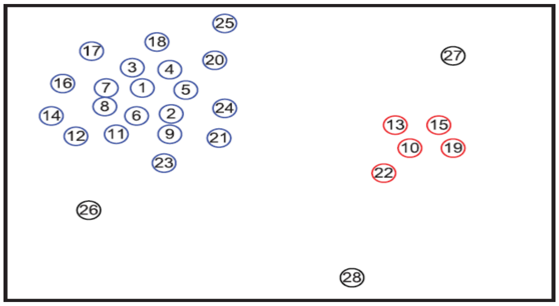

Fig. 2(a) Data points showing the order of decreasing density

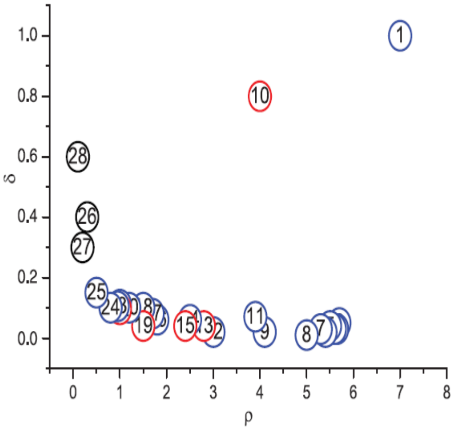

Fig. 2(b) Decision graph

After the identification of cluster centers, the nearest neighbors of highest density are assigned to that particular cluster in one step. In this method, there is no concept of noise-signal cutoff. However, the border region for each cluster is identified as a set of points assigned to particular cluster within a distance dc from data point of other clusters. The next step is the identification of highest density within its border region rb. Those points in clusters whose density is more than rb form the cluster core (robust assignment), while other points are considered to be the part of cluster halo.

4. Experiments and Results

This section provides the details about the design of experiments and discussion on the results.

4.1 Dataset

To evaluate clustering by CFSFDP, we used Washington DC crime dataset downloaded from [8]. From the available data we choose 2016 crime data set which contain 69330 crime records. We used five variables i.e. Offense, Method, Wards, Districts and Neighborhoods from the dataset. Nine types of offenses are recorded in total. These offenses are burglary, theft/other, theft f/other, assault with dangerous weapon, sex abuse, robbery, motor vehicle theft, homicide and arson. Three methods were used in crimes i.e. others, gun and knife. All these crime records are from 8 wards, seven districts and 39 neighborhoods of Washington DC. The data was preprocessed first in order to make it suitable to run on our clustering algorithm. It is worth mentioning here that we chose first 3000 data points from 2016 crime dataset that represents the availability of different density maxima.

4.2 Results and Discussion

We run the algorithm CFSFDP on our dataset with the aim of evaluating its performance and to study its weakness if found. After running the algorithm, two plots are created: (i) standard plot and (ii) distance matrix plot. The standard plot function produce a decision graph which is used for selecting number of clusters manually. From the given decision graph, we can easily see that the expected cluster centers and other data points that are assigned by the algorithm to nearest cluster center based on their value of δ in one step. Fig. 3 shown the decision graph

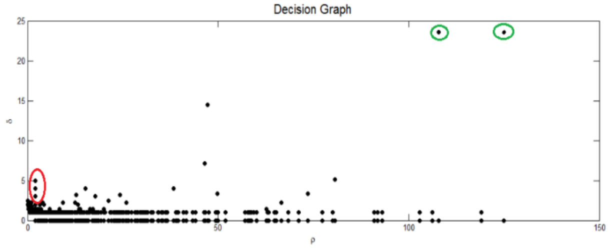

Fig. 3 Decision graph of crime data set

In Figure 3, the data points encircled in green having highest densities with maximum values of δ and also having highest value of ρ as compared to other data points in the same data set are identified as cluster centers. The data points encircled in red having low values of ρ are isolated and considered as outlier or noise. Also it is to be noted that number of clusters depends on selection of data points with high values of δ manually from decision graph. Different clusters are shown in different shape, size, and density using different colors.

From our decision graph, we have selected different number of clusters i.e. 2, 4 and 6 in order to differentiate between numbers of clusters. The reason for choosing different number of cluster is that the accuracy of this algorithm depends on two things: (i) firstly, on accurate estimation of densities of data set; and (ii) secondly on the cutoff distance (dc). The cut off distance is used for calculating the density of each and every data point within crimes data set which identified the border points in clusters. Figure 4 shows the 2D Non-classical multidimensional scaling to represent clusters. Similarly, Figure 5 shows the result with four clusters.

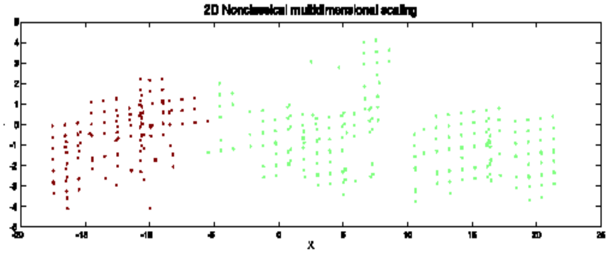

Fig. 4 Two Clusters

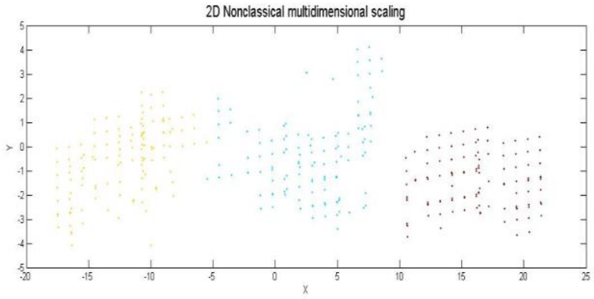

Fig. 5 Four Clusters

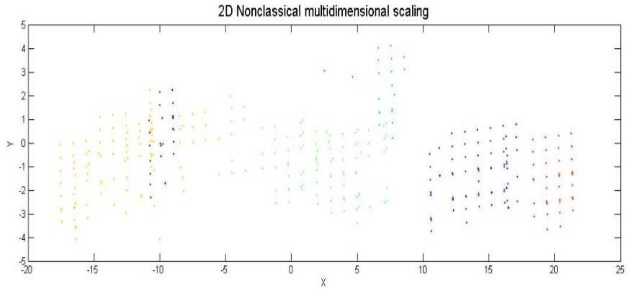

Fig. 6 Six Clusters

CFSFDP algorithm is proved to be efficient because it provide good results in cluster analysis on crime data set. The key idea in using this algorithm is the representation of crime data into 2D space with two axes i.e. ρ and δ. In crime data set ρ is calculated for each and every data point as a local density which is total number of data points around that point while δ is measured as a distance for each point which is the minimal distance between two points with higher density. For a new cluster data points with highest ρ and δ will be selected as cluster centers. Therefore we can argue that CFSFDP is an efficient clustering algorithm in which density is calculated based on cutoff distance set by user. In this way, it found fast and accurate clusters of crime data set by easily finding cluster centers employing density peaks. It has also been noted that density ρ is dependent on cutoff distance dc that appeared as weakness of this algorithm because different choice of cutoff distance could lead to completely different clustering results. Also in decision graph the user has to choose average number of neighbors (in percentage) that could form a cluster where this percentage is strictly related to dc which is not a good option at all because more neighbors the greater dc will be. This fact is clear in Figure 6 in which six points with greater dc were chosen in decision graph where we can see the effect in six clusters in which data points of fifth and sixth clusters overlapped.

5. Conclusion

This study presented the results of experimental study of crime data set using CFSFDP. CFSFDP is a density based clustering algorithm developed to overcome the weaknesses of k-mean and DBSCAN clustering algorithms. CFSFDP automatically identified clusters without any intervention of human; however, the selection of cluster center from decision graphs is important for the user. Cluster centers depend on density peaks and also on cutoff distance (dc) therefore when we increased the cutoff distance, some crime data points appeared split into two different clusters.

References

[1] S.Goswami, and A. Chakrabarti. "Quartile Clustering: A quartile based technique for Generating Meaningful Clusters." Journal of Computing , pp 48-57, 2012.

[2] Bie, Rongfang, Rashid Mehmood, Shanshan Ruan, Yunchuan Sun, and Hussain Dawood. "Adaptive fuzzy clustering by fast search and find of density peaks." Personal and Ubiquitous Computing 20, no. 5, pp. 785-793, 2016.

[3] Karypis, Michael Steinbach George, Vipin Kumar, and Michael Steinbach. "A comparison of document clustering techniques." In TextMining Workshop at KDD2000 (May 2000). 2000.

[4] Rodriguez, Alex, and Alessandro Laio. "Clustering by fast search and find of density peaks." Science 344, no. 6191, pp. 1492-1496, 2016.

[5] Cheng, Qing, Xin Lu, Zhong Liu, Jincai Huang, and Guangquan Cheng. "Spatial clustering with density-ordered tree." Physica A: Statistical Mechanics and its Applications, pp.188-200, 2016.

[6] Lu, Kaiyue, Siyu Xia, Junkang Zhang, and A. Kai Qin. "Robust road detection in shadow conditions." Journal of Electronic Imaging 25, no. 4, pp. 043027, 2016.

[7] Rayan Albarakati, Desnity based data clustering (Master Thesis), 2015.

[8] DC Gov. 2017. Crime Incidents. [ONLINE] Available at: http://opendata.dc.gov/datasets?q=crimes&sort_by=relevance. [Last Accessed 6 May 2019].

[9] Krishnan, Anish, Aditya Sarguru, and AC Shantha Sheela. "PREDICITVE ANALYSIS OF CRIME DATA USING

DEEP LEARNING." International Journal of Pure and Applied Mathematics 118, no. 20, pp. 4023-4031, 2018.

[10] D. W. Sohn, “Residential crimes and neighborhood built environment: Assessing the effectiveness of crime prevention through environmental design (CPTED),” Cities, pp.86-93, 2016.

[11] C. Xu and T. S. P. Yum, "Cross Entropy approach for patrol route planning in dynamic environments," in 2010 IEEE International Conference on Intelligence and Security Informatics, Vancouver, BC, pp. 114-119, 2010.

[12] L. Li, Z. Jiang, N. Duan, W. Dong, K. Hu, and W. Sun "Police patrol service optimization based on the spatial pattern of hotspots," in Service Operations Logistics and Informatics (SOLI) 2011 IEEE International Conference, on, pp. 45-50, 2011.

[13] R. Danilo, M. Adriano, L. V. C. André and F. Vasco, “Towards Optimal Police Patrol Routes with Genetic Algorithms,” in S.Mehrotra et al. (Eds.): ISI 2006, 2006.

[14] Soumya and A. S. Baghel, “A Predictive Model for Mapping Crime using Big Data Analytics,” vol. 04, no. 04, pp. 344–348, 2015.

[15] [P. Chen, H. Yuan, and X. Shu, “Forecasting Crime Using the ARIMA Model,” in Fifth International Conference on Fuzzy Systems and Knowledge Discovery, Nov. 2008.

[16] A. Ghazvini, M. Z. B. A. Nazri, S. N. H. S. Abdullah, M. N. Junoh and Z. Abidin bin Kasim, "Biography commercial serial crime analysis using enhanced dynamic neural network," 2015 7th International Conference of Soft Computing and Pattern Recognition (SoCPaR), Fukuoka, pp. 334-339, 2015.

[17] J. Xu, “Predict Future Events with Point-process Modeling,” IBM Big Data & Analytics Hub, 18-Sep-2015. [Online]. Available: http://www.ibmbigdatahub.com/blog/predict-future-events-pointprocess-modeling

[18] M. A. Awal, J. Rabbi, S. I. Hossain, and M. M. A. Hashem, “Using Linear Regression to Forecast Future Trends in Crime of Bangladesh,” in 5th International Conference on Informatics,Electronics and Vision (ICIEV), 2016.

[19] A. Agarwal, D. Chougule, A. Agarwal, and D. Chimote, “Application for Analysis and Prediction of Crime Data using Data Mining,” in International Journal of Advanced ComputationalEngineering and Networking, ISSN: 2320-2106., vol. 4, no. 5, pp. 9–12, 2016.

[20] Weisburd, David, and Clair White. "Hot spots of crime are not just hot spots of crime: Examining health outcomes at street segments." Journal of Contemporary Criminal Justice, 1043986219832132, 2019.

[21] Chan, Janet, and Lyria Bennett Moses. "Is big data challenging criminology?." Theoretical criminology, vol. 20, no. 1, pp. 21-39, 2016.

[22] Yu, Ya, C. Nicholas Mckinney, Steven B. Caudill, and Franklin G. Mixon Jr. "Athletic contests and individual robberies: An analysis based on hourly crime data." Applied Economics, vol. 48, no. 8 pp. 723-730, 2016.

[23] Williams, Matthew L., Pete Burnap, and Luke Sloan. "Crime sensing with big data: The affordances and limitations of using open-source communications to estimate crime patterns." The British Journal of Criminology, vol. 57, no. 2, pp. 320-340, 2017.

[24] Sanders, Carrie, and Camie Condon. "Crime analysis and cognitive effects: the practice of policing t hrough flows of data." Global crime, vol. 18, no. 3, pp. 237-255, 2017.

[25] Wang, Hongjian, Daniel Kifer, Corina Graif, and Zhenhui Li. "Crime rate inference with big data." In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, pp. 635-644. ACM, 2016.

[26] Kim, Suhong, Param Joshi, Parminder Singh Kalsi, and Pooya Taheri. "Crime Analysis Through Machine Learning." In 2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), pp. 415-420. IEEE, 2018.

Ahmed Alghamdi is Chairman and Assistant Professor in the Cybersecurity Department, College of Computer Science and Engineering, University of Jeddah, KSA. He graduated from the Department of Electrical Engineering and Computer Science of the School of Engineering at The Catholic University of America in Washington, DC with an emphasis in network security. He earned his master of science in formation System from DePaul University in Chicago in 2010.

Prior to entering academia, Dr. Ahmed practiced IT for over six years. He started his technical career as a platform specialist and eventually became a network administrator at Batelco Jeraisy Limited (Atheer) in 2003. He then practiced teaching at the Royal Commission in Yanbu for two Years before he became an IT manager at Al Musahim Gate Company in 2006.