GEOSPATIAL CRIME HOTSPOT DETECTION: A ROBUST FRAMEWORK USING BIRCH CLUSTERING OPTIMAL PARAMETER TUNING

Shima Chakraborty[1] , Sadia Sharmin[2] and Fahim Irfan Alam[3]

[1]Department of Computer Science and Engineering, University of Chittagong, Chittagong- 4331, Bangladesh
[2]Software Engineer, Mid Day Dreams, Chittagong, Bangladesh
[3]South Western Sydney Clinical Campus, School of Clinical Medicine, UNSW

ABSTRACT

Crime causes physical and mental damage. Several crime prevention measures have been developed by law enforcement officials since they realized how serious this problem is. These preventative measures are not strong enough to help lower crime rates because they are typically slow-paced and ineffectual. In this regard, machine learning community has started developing automated approaches for detecting crime hotspot, after performing a careful analysis of the crime trend incorporating geospatial, temporal, demographic, or other relevant information. In this research, we look at detecting crime hotspots using geospatial information of prior crime occurrences. We proposed BIRCH algorithm to detect high crime prone areas with four essential aspects: (1) PCA (Principle Component Analysis) has been used to minimize the dimensionality of crime data, (2) Silhouette score Elbow and Calinski Harabaz have been used to find the optimal number of cluster (3) utilized hyper-parameter tuning to choose the best hyper-parameters for the BIRCH algorithm (4) applied BIRCH with the three aspects mentioned above. The results of the suggested framework were then contrasted with those of alternative clustering techniques, such as K-means, DBSCAN, and the agglomerative algorithm. We explored our approaches on the London Crime Dataset and found some fascinating results that can help reducing crime by helping people take the appropriate measures.

KEYWORDS

PCA, K-means, DBSCAN, agglomerative, BIRCH

1. INTRODUCTION

Crime is defined as an unlawful act that results in the loss of money, property, or other assets together with physical or psychological suffering. This may result in human suffering, death to individuals or death on a large scale or major life-threatening injuries. This is a widespread occurrence that follows breaking laws and that, once law enforcement officials fully comprehend the nature of the crimes, leads to convictions. It's a very dynamic phenomenon that varies globally in both quantity and style. The negative impact of crimes does not restrict to personal level only. Additionally, it impacts social values, mental health, childhood trauma, financial development [1] and even a nation's reputation [2] and even reputation of a country. When crime increases the nation's development falls on its face. The role of the police is not only to grasp a perpetrator who has committed a crime or offence, but also to act safely and effectively in high-risk areas where crimes are likely to occur, so that the police can create an environment in which criminals cannot commit the crime or are arrested by police before they do so. A community may be able to concentrate on a particular region and implement efficient measures to deter potential

DOI: 10.5121/ijwest.2025.16101

crimes with the aid of comprehensive investigation and evaluation, which can also offer us insightful information about crime trends. One of the main components of crime mapping is the identification of hot spots, or areas with a high crime rate. Hot spot analysis aids authorities in identifying high-crime regions, crime types and the best path of action. To keep an area secure, law enforcement organizations employ various patrolling tactics based on the information they receive. We could forecast location of a crime before it happened, including the time of the crime and the name of the perpetrator. Even though it might sound like science fiction, social scientists have long understood that past criminal behaviours greatly influence current trends. Crime analysis is the use of analytic and statistical methodologies by which the police identify potentially risky targets for police involvement, crime prevention, or to investigate an alreadycommitted crime. Since the beginning of police practice, statistical or geographical approaches have been utilized, but with the advancement of information technology, the focus has shifted to data connected to crime and its collection, processing, and analysis. Predicting crime hotspot will benefit society in various ways. The major goals are to increase criminology knowledge and to develop tactics that promote more efficient and effective police measures. This will aid law enforcement agencies in reducing crime by allowing them to forecast future crime rates, crime locations, and crime times. It will not only increase the public safety but also decrease the economic loss. With these strategies, police forces should be able to work more effectively with minimal resources. Thus, a sustainable development for society will be maintained. Wim Bernasco et al. [3] state various earlier contributions demonstrating that crime is concentrated in specific micro locations inside the city with high intensity; such locations are referred to as hotspots. The authors also suggested that the use of geographical patterns of crime to predict crime requires the establishment of a theoretical framework. As a result, the spatial characteristics of crime geography can help in specialized police operations such as hotspot policing and predictive policing. k-means is a data clustering method that may be applied to unsupervised machine learning. It can divide unlabeled data into a predefined number of groups based on similarities (k). After calculating centroids, K-means clustering iterates until the optimal centroid is found. The number of clusters should be known. The number of clusters discovered by the algorithm from data is denoted by the letter ’K’ in K- means. Jyoti Agarwal et al. [4] focuses on crime analysis by utilizing the rapid miner tool to execute the k-means clustering method on crime datasets. Mrs. S. Aarthi et al. [5] describes the K-means clustering technique and the streaming algorithm for identifying crime. Unrelated observations can be grouped together using K- Means clustering. Even if the observations are dispersed throughout the dimensional space, they eventually come together to create a cluster. Each data point contributes to the formation of clusters since clusters are produced by the mean value of cluster members. A little change in data points can impact clustering results. This issue is much decreased with DBSCAN due to the manner clusters are generated. DBSCAN is a density-based clustering technique, which means that clusters are dense areas of space separated Data points that are "densely clustered" are combined into a single cluster. It can find clusters in massive geographical datasets by evaluating the local density of data points. DBSCAN clustering’s tolerance to outliers is its most remarkable feature. In this study, Divya G et al. [6] compared three clustering techniques, namely hierarchical clustering, k -means clustering, and DBSCAN clustering, to determine which, one is most suited for crime hotspot research. Each of the clustering methods evaluated here requires inputs such as cluster number, neighbour distance, minimum number of points, and so on. The Euclidean distance is used to calculate cluster similarity. Because of its intrinsic density-driven character, the results show that DBSCAN is significantly better appropriate for crime hotspot analysis. Hierarchical clustering can be used as an alternative to partition clustering because there is no need to specify the number of clusters to be formed. Hierarchical agglomerative clustering is a way of grouping that works from the bottom up which is a popular example [11]. Most clustering methods do not scale well as dataset quantities increase and input/output costs decrease. BIRCH generally takes only a single scan of the database to locate a suitable clustering and increase the quality further with a few further scans. BIRCH’s capacity to incrementally and

dynamically cluster incoming multidimensional metric data points to achieve the best quality clustering with available resources such as memory and time limits is one of its advantages. BIRCH is also the first clustering algorithm established in the field of machine learning that effectively manages noise. Tian et al. [7] proposed the BIRCH clustering algorithm and demonstrated its suitability for very large datasets. They also compared BIRCH's performance to CLARANS, a recently developed clustering approach for huge datasets, and discovered that BIRCH outperforms CLARANS. Borish et al. [8] proposed A-BIRCH, a parameter-free form of BIRCH, is a method for automatically estimating thresholds for the BIRCH clustering algorithm using the Gap Statistic. Du et al. [10] described D-BIRCH cluster's algorithm, a sort of cluster optimizing BIRCH cluster's algorithm that can alter threshold values in real time and operate data. The rest of the paper is organized as follows. Section II outlines the methods for comprehending this paper's intricate framework. In our paper Section III articulates the current geographic detection framework, Section IV concentrates on the specific data preparation process, and Section V outlines our research methodology.

2. METHODOLOGY

In this section, we will look at the approaches that will be used to develop our model for detecting crime hotspots using unsupervised machine learning, as well as discuss the significance of parameter optimization.

2.1. Principal Component Analysis (PCA)

PCA is a commonly used statistical approach for unsupervised dimension reduction. PCA is performed before clustering for efficiency reasons, as clustering methods are more efficient for lower dimensional data. It's utilized when dealing with the dimensionality curse in data with linear lationships, i.e. when there are too many dimensions (features) in data, which generates noise and problems. It decreases the size of a dataset by extracting new characteristics from the existing ones. As a result, it mixes the input variables (or features) in a precise way to produce" new" features while maintaining the most relevant information from all of the original features. After PCA, all the "new" variables are unrelated to one another. Also, PCA reduce the computation cost. Despite not being necessary, this step is strongly advised.

2.2. Hyper Parameter Tuning

Hyper parameter optimization is an important part of man- aging a machine learning model's behaviour. If we don't modify our hyperparameters appropriately, our estimated model parameters produce less-than-ideal results since they don't minimize the loss function.This implies that our model makes more errors.

2.3. Balanced Iterative Reducing and Clustering Hierarchies (BIRCH)

The two most popular methods for clustering are agglomerative clustering and K means. However, BIRCH and DBSCAN are the advanced clustering algorithms that are recommended when accurate clustering on very large datasets is needed. Furthermore, because of its ease of application, BIRCH is very beneficial. huge datasets were a challenge for earlier clustering techniques, as they were unable to manage scenarios in which a dataset was too huge to fit in main memory. Furthermore, for every clustering choice, the majority of earlier iterations of BIRCH analyze every data point (or every cluster that is currently in existence) equally. They don't employ heuristic weighting based on data point distance. Consequently, a significant amount of overhead was needed to maintain an acceptable clustering quality while reducing the

expense of additional IO (input/output) operations. The BIRCH clustering method divides the dataset into brief summaries first, and then groups the short summaries together. It does not cluster the dataset directly. That's the reason BIRCH is often used alongside with other clustering approaches; once the summary is created, it may be further grouped using those other clustering methods. It transforms data to a tree data structure and reads the centroids from the leaf. These centroids can subsequently be utilized as the final cluster centroid or as input to other cluster algorithms such as Agglomerative Clustering. BIRCH is a scalable hierarchical clustering algorithm that requires only a single scan of the dataset, allowing it to deal with big datasets efficiently. This approach is based on the CF (clustering features) tree. Furthermore, this technique creates clusters using a tree structured summary. The BIRCH algorithm, known as the Clustering feature tree, constructs the tree structure of the input data (CF tree). A triple of integers (N, LS, SS) denotes a cluster of data points, where N is the number of elements in the sub-cluster, LS is the linear sum of the points, and SS is the sum of the squares of the points. BIRCH clustering algorithm has four phases. Flow- chart of BIRCH Algorithm is depicted in Fig. 1.

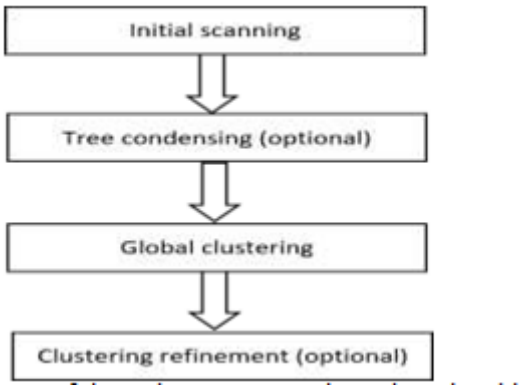


Fig. 1: Flow-chart of BIRCH Algorithm

Initial scanning: Scanning all data and constructing an initial in-memory CF tree.

Optional Condensing: Rebuild the CF-tree to make it smaller and faster to analyze, but at the expense of accuracy.

Global clustering: It passes CF trees to current clustering methods for clustering.

Clustering refinement: The issue with CF trees, where different leaf nodes receive the same valued points, is resolved by refining

2.4. Parameters of BIRCH

This algorithm has three tuning parameters. Unlike K-means, the optimal number of clusters (k) is determined by the algorithm and does not require user input.

Threshold: The most data points that can be stored in a sub-cluster within the CF tree's leaf node.

Branching factor: The maximum number of CF sub- clusters that can exist in a single node is specified by this parameter (internal node).

N clusters: The number of clusters that are returned after the BIRCH algorithm has run through to the end, or the number of clusters following the last clustering step.

2.5. Parameter Optimization

Hyper parameter optimization is an important part of man- aging a machine learning model's behaviour. If we do not properly change our hyper parameters, our predicted model parameters do not minimize the loss function, which results in less-than-ideal results. This implies that our model makes more errors.

Hyper-parameter Tuning of BIRCH: The threshold and branching factor are considered as two important and significant hyper parameters for achieving steady clustering performance. There is an underlying relationship between these two which affect the clustering to a larger extent. In order to determine each of these two hyperparameters' unique optimal values to supply as input to the clustering algorithm, the relationship between them is taken into account in this work.We set specific values for the threshold and branching factor and compute the silhouette score for each possible combination of those two hyper parameters. As we exploit the possible combination among all the values that we primarily set, we obtain silhouette score values that will make optimal impact on the clustering performance. The pseudocode for tuning the hyper parameters is shown in the Alg. 1.

Algorithm 1 An algorithm for tuning the hyper-parameters of BIRCH

Input: X ← dataset

```
for threshold in param grid['threshold'] : do
    for branching factor in param grid['branching factor'] : do
        birch ← Birch(n clusters = 2, threshold ← threshold, branching factor ← branching factor)
        birch.fit(X)
        X all.append(X)
        y pred← birch.predict(X)
        y all.append(y pred)
        SH ← metrics.silhouette score(X, y pred)
        SH all.append(CH)
        paras all.append([threshold, branching factor])
    end for
end for
```

Output: "Threshold:", threshold, "Branching factor:",branching factor, "Silhouette Score:", % SH)

Output: "Threshold:", threshold, "Branching factor:",branching factor, "Silhouette Score:", % SH)

In addition to implement BIRCH, we should follow the steps shown of Fig. 2.

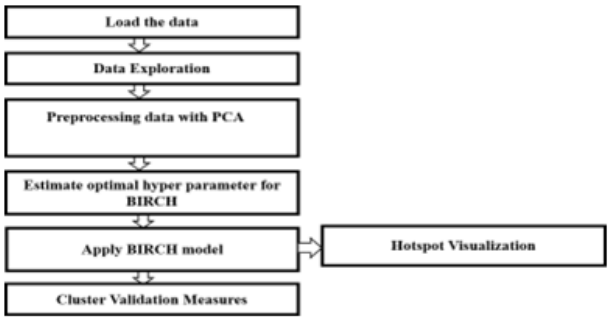


Fig. 2: Steps to Implement BIRCH

We employ the following steps in our BIRCH clustering:

Data exploration: Evaluating the most impacted neighborhoods in the city, determining what types of crime occur where, and analyzing crime hotspots around the city.

Pre-processing data with PCA: After scaling, we apply PCA to reduce dimensionality. Estimate optimal hyper parameter for BIRCH: To get the optimal value for BIRCH, we estimate hyper parameter for BIRCH.

Apply BIRCH model: BIRCH model is applied with optimal hyper parameter value. Hotspot visualization: ArcGIS geoprocessing tool used to visualize significant crime hot spots area.

Cluster Validation Measures: Different cluster validation is used to validate the clustering model.

2.6. Using Folium to Visualize Geospatial Data

Folium is a Python data visualization toolkit that focuses on displaying geographic data. Folium gives the ability to create a map of any location on the world. Folium's maps are interactive as well, allowing users to zoom in and out once the map has been presented, which is a really valuable feature. Folium was built with simplicity, speed, and utility in mind. It performs well, can be expanded with a variety of plugins, and has a user-friendly API.

2.7. Measures of Cluster Validation

Cluster analysis requires cluster validation. The accuracy and performance of the clustered data are then evaluated. External indices and internal indices are the two types of validity indicators used to evaluate accuracy and quality. An external index measures cooperation between two partitions, one of which is the known clustering structure and the other the output of the clustering method [12]. In the absence of external data, internal indices are employed to evaluate a clustering structure's quality [13].

We employed internal indices in our experiment because we didn't have a previous clustering structure and didn't know the ground truth labels. As a result, we employed four internal indices, as listed and explained below.

Silhouette Coefficient

The silhouette coefficient is a validation and interpretation method for analyzing data cluster consistency. Its value reflects how well the data point has been classified. It's a statistic that compares the resemblance of a data sample to its own cluster (cohesion) to that of other clusters (separation). Each data sample's value is calculated using the mean intercluster distance and mean nearest-cluster distance [14].

Dunn Index: The Dunn Index is a statistic used to evaluate clustering methods. It determines the cluster's com- pactness, or the maximum distance between its data points, as well as the cluster's separation (the lowest distance between clusters) [17].

Calinski Harabaz Score

The Calinski Harabaz Score, or Variation Ratio Criterion, looks at the difference between within-class and inter-class dispersion. It is based on clusters that are closely spaced. It's used to figure out how many clus- ters are best [15]. And is derived by dividing the inside cluster distance by the between cluster distance, then computing the clusters' overall average.

Davis Bouldin Score:

The Davis Bouldin (DB) score, like the Dunn Index, silhouette score, and Calinski-Harabasz index, is based on the cluster itself rather than external labels. In comparison to other scores or indexes, it is straight forward to calculate. It ranges from 0 to 1, with a lower Davis Bouldin score being deemed better. It is limited to utilizing the Euclidean distance function since it calculates the distance between cluster centroids [16].

3. EXPERIMENTAL RESULTS

3.1. Dataset Description

We have gathered the London crime dataset, which is accessible to the general public on London police's official website. [9]. This benchmark dataset encompasses a large amount of crime data, covering 14 different categories of crimes in the city of London, such as antisocial behavior, bicycle theft, burglary, criminal-damage and arson, drugs, other crime, other theft, possession of weapons, public order, robbery, shoplifting, theft from the person, vehicle crime, violence, and sexual offenses. Every month, a separate file containing the data from each month's crimes was distributed by the police authority. We set up our experimental setup by combining multiple monthly records into a single dataset.

The dataset we collected was in an unstructured form which is why we use the data processing technique to structure it by computing the number of each crime that occurred per month. For our study, we conduct a separate assessment of crime incidents that occurred in 2019, 2020, and 2021. Within the data, there is a string format column for latitude and longitude. Any of our Machine Learning models (K-means, DBSCAN, Agglomerative, and BIRCH) need numerical input. That is why we use to numeric () function which is one of the general functions in Pandas that is used to convert argument to a numeric type.

3.2. Data Exploration

We explore the city's most affected areas, determining what kind of crimes occur where, and assessing crime hotspots around the city. Investigate which crimes occurred the most in each year. The data set was modified such that the key crime indicators in London were grouped by area.

3.3. Hyper-Parameter Settings

The hyper-parameters of the models, which are essential part of the machine learning models, must be specified. The hyper parameters utilized in the machine learning models that we employed throughout our experiments are described in this section. Table: I present the hyper parameters of our experimented approaches.

TABLE I: Hyper Parameter used in Clustering Models

Clustering	Hyper parameter setting
K-means	n clusters=2, random state= 3425
DBSCAN	Eps=1.5, MinPts =4
Agglomerative	n clusters=2,linkage='ward'
BIRCH	threshold=1, branching factor=50, n clusters=2

3.4. Building the BIRCH Model

Birch is an efficient and convenient unsupervised clustering approach for huge volumes of data. The key challenge with this method is calculating the value of k, which is the number of clusters that must be known before performing the clustering process. Despite the fact that the number of clusters in this problem is evident because there are only two zone types labeled violent and non-violent, Silhouette score Elbow for Birch Clustering and Calinski Harabaz Score Elbow for Birch Clustering study were used to the dataset to arrive at an adequate value of k.

Fig. 3 and Fig. 4 shows the result of Silhouette score Elbow for Birch Clustering and Calinski Harabaz Score Elbow for Birch Clustering where suggest cluster = 2 as the best value of n clusters for BIRCH clustering. PCA is an unsupervised method for preprocessing and reducing the dimensionality of huge datasets while preserving their original structure and relationships. PCA contributes to better clusters and faster running times. This study also attempts to develop and apply PCA on the data analysis refers to clustering in order to improve the display of created clusters over the 2D plane. PCA gathers the characteristics with the highest point of variance and attempts to minimize dimensionality by extracting just these features. To capture the maximum variety in the data, three major components are selected based on the largest principal components. The clustered results are shown along two main components. Fig 5 shows and compares clustering results with and without the use of PCA.

Hyper parameter Tuning

In Table. II shows Silhouette Score of the different combination of threshold and branching factor to detect the optimal hyper-parameter for BIRCH.

TABLE II: Silhouette Score for different value of Threshold and Branching factor

Threshold	Branching factor	Silhouette Score
0.2	50	0.67
0.2	150	0.78
0.2	200	0.79
0.3	50	0.73
0.3	100	0.85
0.3	200	0.79
0.5	50	0.90
0.5	150	0.85
1.0	50	0.92

In Table. II, we can see that among all value, threshold=1 and branching factor =50 has the highest silhouette score. The findings of several internal validation metrics used to the clusters generated using BIRCH clustering are summarized in Table VI.

Result Analysis with PCA

As mentioned above, PCA decreases the dimensionality of the data, which improves the model's efficiency and speeds up algorithms on the dataset because crime data is highly dimensional. Additionally, this effort aims to improve the clustering outcomes by applying PCA to the data prior to clustering. PCA analyzes the characteristics with the highest point of variance and extracts just these features to minimize dimensionality. The clustering results are presented along two main components. Fig. 5 shows and compares clustering results when PCA is used.

3.5. Comparison Between Clustering Techniques

Cluster validation is an essential component of cluster analysis. The important step after clustering all of our data is to validate the clustered data's outcomes in terms of accuracy and performance, as well as to quantify their validity and quality. We employed internal indices to compare clustering techniques because we had no previous clustering structure, i.e. ground truth labels were unknown. We employed four internal indices: the Silhouette score, the Dunn Index, the Calinski Harabaz Score, and the Davis Bouldin Score.

TABLE III: Internal validation measure for k-means clustering

K-means			
Validation	2019	2020	2021
Silhouette	0.825	0.735	0.743
Dunn Index	0.004	0.006	0.008
C-H Score	3290.12	2485.97	2021.5
D-B Score	0.731	0.829	0.8

TABLE IV: Internal validation measure for DBSCAN clustering

DBSCAN			
Validation	2019	2020	2021
Silhouette	0.004	0.61	0.48
Dunn Index	431.1	0.006	0.004
C-H Score	1.97	668.38	431.1
D-B Score	1.58	1.77	1.97

TABLE V: Internal validation measure for Agglomerative clustering

Agglomerave			
Validation	2019	2020	2021
Silhouette	0.84	0.63	0.82
Dunn Index	0.009	0.004	0.014
C-H Score	3127.09	2148.74	1570.56
D-B Score	0.72	0.97	0.69

TABLE VI: Internal validation measure for BIRCH clustering

BIRCH			
Validation	2019	2020	2021
Silhouette	0.946	0.918	0.92
Dunn Index	0.325	0.233	0.22
C-H Score	1065.376	739.317	675.4
D-B Score	0.195	0.299	0.31

We can state that BIRCH clustering is the best acceptable clustering strategy for this dataset when compared to K-means, DBSCAN, and Agglomerative after comparing the validation scores for all metrics of each clustering method.

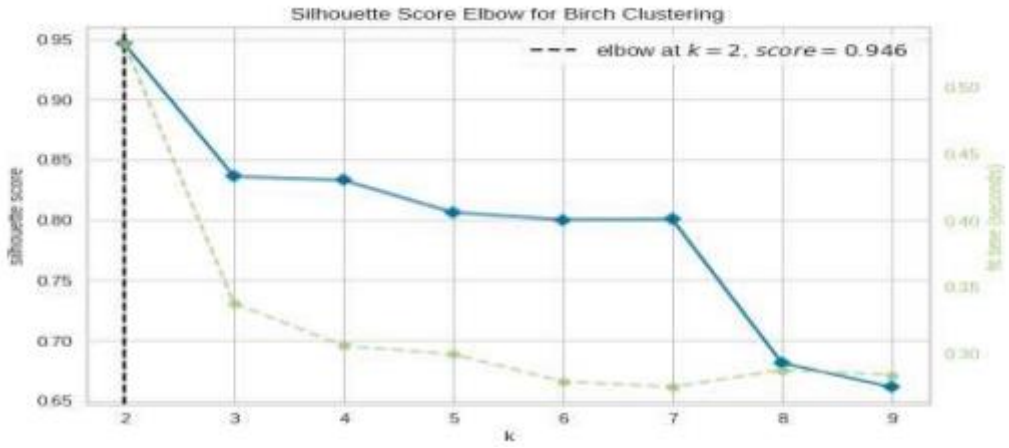


Fig. 3: Silhouette Score index for BIRCH clustering

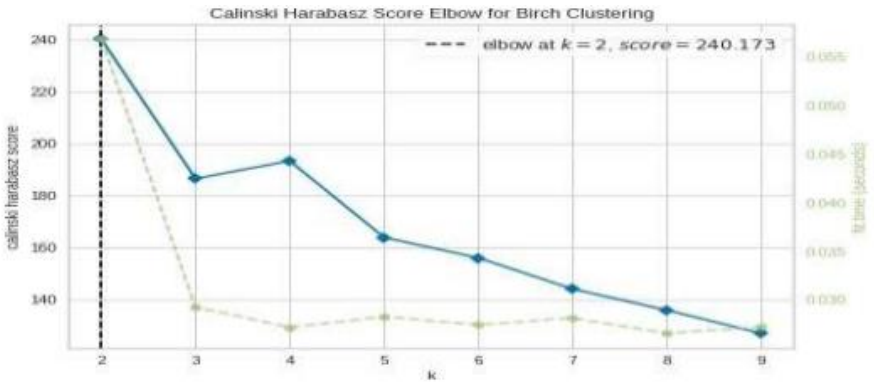


Fig. 4: Calinski Harabasz score for BIRCH clustering.

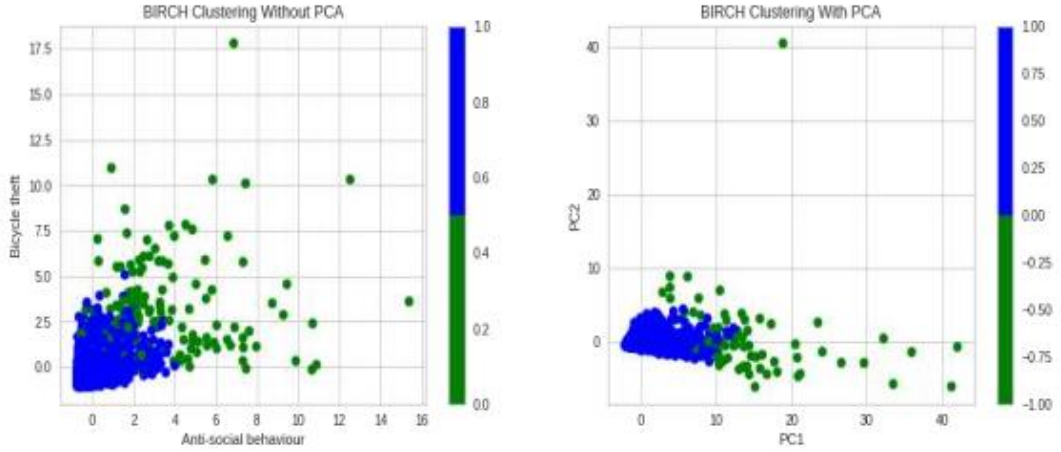
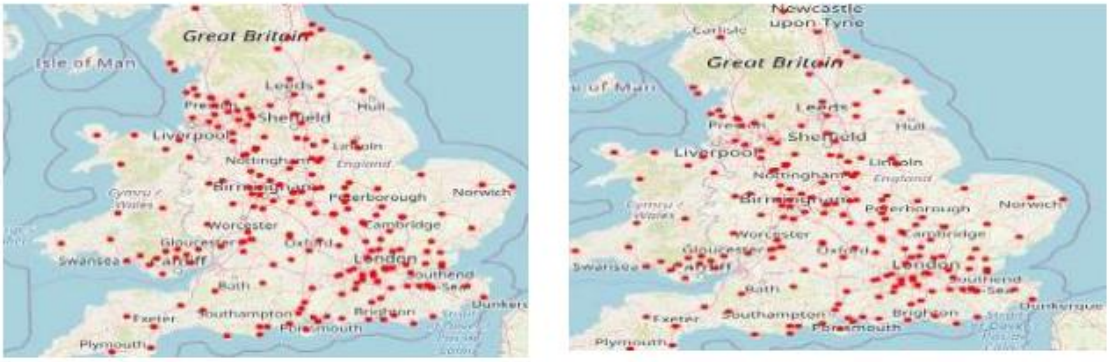


Fig. 5: BIRCH Clustering Without and With PCA

4. VISUALIZATION OF CRIME HOTSPOT AREAS OF LONDON

The BIRCH clustering findings are displayed on Fig. 6a, 6b over a map of London to provide a better visual representation of violent neighbourhoods for both police and the general public.



(a) London's most crime prone areas in 2019

b) London's most crime prone areas in 2020

5. CONCLUSION

In this paper, we proposed integrated solutions for identifying crime hotspot in London with a view to analyze the highly crime zone areas. We used the extended formulations of clustering techniques called BIRCH. Then compare the result of proposed model with K-means, DBSCAN and Agglomerative clustering methods. Using these findings, crime analysts can advise people on the appropriate safety measures to prevent crimes.

REFERENCES

[1] Otranto, Edoardo & Detotto, Claudio. (2010). Does Crime Af-fect Economic Growth?. Kyklos. 63. 330-345. 10.1111/j.1467- 6435.2010.00477.x.

[2] Brewer-Smyth, K., Cornelius, M. E., & Pickelsimer, E. E. (2015). Child- hood adversity, mental health, and violent crime. Journal of forensic nursing, 11(1), 4-14.

[3] Vandeviver, Christophe & Bernasco, Wim. (2017). The geog- raphy of crime and crime control. Applied Geography. 86. 10.1016/j.apgeog.2017.08.012.

[4] Agarwal, Jyoti & Nagpal, Renuka & Sehgal, Rajni. (2013). Crime Analysis using K-Means Clustering. International Journal of Computer Applications. 83. 1-4. 10.5120/14433-2579.

[5] Samyuktha, M. & Sahana, M.. (2019). Crime Hotspot Detec- tion With Clustering Algorithm Using Data Mining. 401-405. 10.1109/ICOEI.2019.8862587.

[6] Divya (2014). Suitability of Clustering Algorithms for Crime Hotspot Analysis.

[7] Zhang, T., Ramakrishnan, R., & Livny, M. (1996). BIRCH: an efficient data clustering method for very large databases. ACM SIGMOD Conference.

[8] Lorbeer, Boris & Kosareva, Ana & Deva, Bersant & Softic´, Dz˘enan & Ruppel, Peter & Ku¨pper, Axel. (2017). A-BIRCH: Automatic Threshold Estimation for the BIRCH Clustering Algorithm. 169-178. 10.1007/978- 3-319-47898-2 18.

[9] https://data.police.uk/data/.

[10] Du, Haizhou & Yong Bin, Li. (2010). An Improved BIRCH Clus- tering Algorithm and Application in Thermal Power. 53 - 56. 10.1109/WISM.2010.123. 10.1243/095440605X8298. A.

[11] K. Jain, R. C. Dubes, Algorithms for clustering data, Prentice-Hall, Inc., 198.

[12] Dudoit, S., & Fridlyand, J. (2002). A prediction-based resampling method for estimating the number of clusters in a dataset. Genome biology, 3, 1-21.

[13] Thalamuthu, A., Mukhopadhyay, I., Zheng, X., & Tseng, G. C. (2006). Evaluation and comparison of gene clustering methods in microarray analysis. Bioinformatics, 22(19), 2405-2412

[14] Aranganayagi, S., & Thangavel, K. (2007, December). Clustering cat- egorical data using silhouette coefficient as a relocating measure. In International conference on computational intelligence and multimedia applications (ICCIMA 2007) (Vol. 2, pp. 13-17). IEEE.

[15] Baarsch, J., & Celebi, M. E. (2012, March). Investigation of internal validity measures for K-means clustering. In Proceedings of the inter- national multiconference of engineers and computer scientists (Vol. 1, pp. 14-16). sn.

[16] Davies, D. L., & Bouldin, D. W. (1979). A cluster separation measure. IEEE transactions on pattern analysis and machine intelligence, (2), 224- 227.

[17] Desgraupes, B. (2013). Clustering indices. University of Paris Ouest-Lab Modal'X, 1(1), 34.

[18] S. Ashraf and T. Ahmed, "Sagacious Intrusion Detection Strategy in Sensor Network," 2020 International Conference on UK-China Emerging Technologies (UCET), Glasgow, UK, 2020, pp. 1-4, doi:10.1109/UCET51115.2020.9205412.

[19] S. Saleem, S. Ashraf and M. K Basit, "CMBA - A Candid Multi-Purpose Biometric Approach," August 2020, ICTACT Journal on Image and Video Processing , Volume: 11, Issue: 1, Pages: 2211-2216, doi: 10.21917/ijivp.2020.0317

AUTHORS

Shima Chakraborty obtained B.Sc. in 2009 and MS (Engg.) in 2012 in Computer Science and Engineering from University of Chittagong. She is presently working as an Assistant Professor in Department of Computer Science and Engineering at University of Chittagong. Her areas of interest in study are machine learning, artificial intelligence, data mining, big data, and the semantic web. She has published research articles in various national and international conferences.

Sadia Sharmin is a Software Engineer at Mid Day Dreams Software Firm. She graduated from the University of Chittagong with a Bachelor of Science in Computer Science and Engineering in 2022 and a Master of Science in Computer Science and Engineering in 2024. Her research primarily focuses on data analysis, crime hotspot detection, and predictive modelling through machine learning techniques.

Dr. Fahim Irfan Alam is a post-doctoral research fellow at the school of medicine & health, University of New South Wales, Australia, leveraging his expertise in machine learning to develop automated solutions to address critical research questions in the radiation oncology domain. With a strong academic foundation that includes a bachelor's degree in computer science and engineering from the University of Chittagong, Bangladesh, a master's from St. Francis Xavier University, Canada, and a PhD from Griffith University, Australia, Fahim focuses on building predictive models and facilitating clinical data integration for multi-centre studies under the Australian Computer-Assisted Theragnostics (AusCAT) platform.