Crime Analysis Based on K-Means Clustering

Nithin Joseph

August 1, 2021

Crime Analysis Based on K-Means Clustering

Nithin Joseph

ME – CSE (AI & ML)

20MAI1005@cuchd.in

Punjab

***Abstract*— In today's world, criminals and terrorists are technologically sophisticated. All the government gives higher priority to prevent and reduce crimes. Crime analysis is a collection of strategies that allow the police forces to become more effective through better knowledge. The basic objective of any clustering algorithm is to cluster or group similar data points into a single cluster. Our proposed framework aims to forecast the probability of crime occurring in a city by analyzing the crime dataset and visualizing the findings for better comprehension. This research is achieved by using a clustering algorithm of k means that group related objects into clusters, the proposed research work mainly focuses on predicting the region with higher crime rates.**

Keywords—K-Means, Crime analysis, Clustering, Unsupervised algorithm

I. Introduction

A. Introduction About Technology

In today's world security is an aspect that gives higher priority by all politicians and governments and tries to reduce crimes. In the present scenario, criminals are technologically enthusiastic and it is one of the biggest challenges faced by intelligence and law enforcement agencies. Behind every organized crime, there might be some small crime or chain of crimes that we ignored as random or coincidental. So these enforcement agencies need technologies to extract patterns and trends in crimes in different areas and thereby possibly prevent or reduce the crimes. So we should choose an appropriate field to extract the relevant information from the crime dataset. Datamining is referred to extracting the relevant or necessary information from a high-volume dataset. In this project, data mining is used on a large dataset of crime to extract patterns and it will support law agencies to reduce crime rates by identifying crime-prone areas and the type of crime that possibly occurs. This project uses K – Means clustering algorithm of data mining to clustering the data points in such a way that the distance between points in the same groups is as minimum as possible and the distance between points in different groups is higher.

B. Steps To Perform Clustering

a) *Pre-processing and feature selection:* it involves choosing appropriate pre-processing and feature selection. Pre-processing ensures that the data is good to perform operations. Feature selection chooses a subset of relevant features from the dataset to reduce the dimensionality of the large dataset.

b) *Similarity measures:* one of the important steps in the processing of clustering. Data points are clustered into different groups/clusters based on the similarity measures.

c) *Clustering algorithms:* clustering algorithms uses particular similarity measures to naturally groups the data points into different clusters. Most of the algorithms use the distance between cluster centroids and data points.

d) *Result validation:* check whether the results are valid. If not, iterate back to some previous stages. try to improve the performance of the algorithm and verify the result again.

The objectives are

- To cluster the data points based on the distance between crime patterns.
- To apply pre-processing techniques.
- To analyse the data and identify the crime patterns, perform the prediction techniques based on the identified crime areas that are obtained after clustering.

II. Related Works

K-means clustering algorithm of data mining is commonly used for predicting the patterns in the crime from a given crime data set. The dataset consists of facts on crime prevalence and attributes which includes description, month, day, hour, type of crime and the no of arrest made etc. [1]. Implementing a clustering algorithm on crime datasets enables analysis of crimes by clustering similar crimes into a single cluster and then extracting the patterns from it [2]. It makes identification and analysis of various criminality trends over the years through their conclusion.

K-means clustering algorithm is highly sensitive to the initial random centroids. The random initial starting points produced by K-means give results in the form of the cluster that helps in reaching the local optima. Careful selection of random initial cluster centers may help to improve the overall performance of the K- Means clustering algorithm [3]. Fuzzy C- Means clustering algorithm can also use to predict the crimes. The crime patterns can be analyzed and prevented based on the crime data. After collecting and pre-processing the data, apply clustering techniques to cluster the data of crime and analyze it. To analyze the data and identify the crime patterns, perform the prediction techniques based on the identified crime areas that are obtained after clustering. The main scope of the project is to analyze the crime patterns and identify the crime-predicted areas based on the crime rate using the crime data set. Using fuzzy clustering, obtain the crime patterns to know in which area the crime will occur frequently [4].

Efficacious clustering algorithm for the extraction and interpretation of data to predict the crime analysis results. K-means is one of the best algorithms that resolve a familiar clustering problem. An optimized K-means method for data clustering will reduce changing the cluster values after a certain amount of repetitions. The cluster number depends on the k value and finding the optimum k value will lead to picking a better centroid value [5].

Data is classified through the collection, classification, pattern identification, prediction, and visualization for analyzing crime rates of each state [6]. They use Naive Bayes classifiers for creating a model for classifying each crime. It is a supervised learning problem where the class for a set of training data points and needs to propose the class for any other given data point. Apriori algorithm can be used to identify crime patterns and to make predictions with the aid of decision trees [7]. It uses a heat map for the identification of the level of each activity. For example, dark color is representing low activities.

A decision tree algorithm can be used to detect suspicious emails using an enhanced Iterative Dichotomiser3 (ID 3) algorithm using improved feature selection methods [8]. The application of these important factors produces more desirable and quicker decision trees. Prediction of crime occurrence can also be made through the data fusion method with deep neural networks [9].

III. Methodology

The type of crime and rate of crime in different areas are used to find the pattern and analyze the crime data. if we could cluster the dataset into different clusters based on these factors with better accuracy, then it is easy to analyze dataset and predict future occurrence of crimes.

In the crime analysis process, the very first step is to load the data, and then remove noisy and missing data based on preprocessing techniques. After preprocessing the data set, cluster the data using the clustering techniques. K-Means is the best among various unsupervised clustering algorithms. K-Means algorithm is used to cluster the data based on the distance measures. Finally visualize the clustered data for further analysis.

A. Steps for Implementation

The system mainly contains four modules.

1. Preprocessing the crime data.
2. Identifying the number of clusters.
3. Cluster the data set using K-Means.
4. Visualize the clustered data points.

In this methodology, the US Arrests data set is used. The data set can be collected from the “KAGGLE” website. It maintains the district-wise number of arrests made in an year with various types of crimes such as Assault, Murder, Urban Pop, and Rape etc.

Preprocessing the data

Crime-prone area identification and pattern extraction are done based on the dataset. The prediction of the crime uses different types of crimes available in the dataset. Detection of crime, predicting the behaviour or pattern are depending on the dataset. Hence, the data fed into the system should be clean and accurate.

Identifying no of Clusters

In the K-Means clustering algorithm, K represents the number of clusters. It also used to finalize the cluster centers. The performance of the entire system depends on the value of K. one of the main challenges is to choose the correct value for K. for the same, an ELBOW method is used. The ELBOW method is a heuristic method used to find the number of clusters in the K-Means clustering algorithm. Elbow method plots a graph with a number of clusters in the X-axis and a distribution score in the Y-axis. The distribution score for the dataset decreases exponentially until a specific number and then starts to decrease linearly. This point is taken as the number of the cluster for the dataset.

Cluster The Data Set Using K-Means

K-Means clustering algorithm is an unsupervised algorithm which widely used for clustering purposes. The algorithm clusters the data points into a user-defined number of clusters. The algorithm uses distance measures to cluster the data. Euclidian distance is the most widely used distance measurement. The algorithm first randomly chooses K points in the dataset as cluster centers. Then using the Euclidian distance measure cluster each data points into associated clusters. Once all the data points are clustered into any of the clusters, calculate the cluster centers again. This process iterated K times and the final cluster centers are fixed. These values are further used for processing.

Visualize Clustered Data Points

After the K-Means clustering algorithm performed, data points will be cluster into associated cluster centers. With the help of a scatter diagram, these data points are visualized for better understanding. The same is represented in Fig 5.6. once the data points are clustered, the cluster number of each data point is added into the data set and the updated dataset is stored in the cloud for the analysis purpose.

B. *Flow Chart*

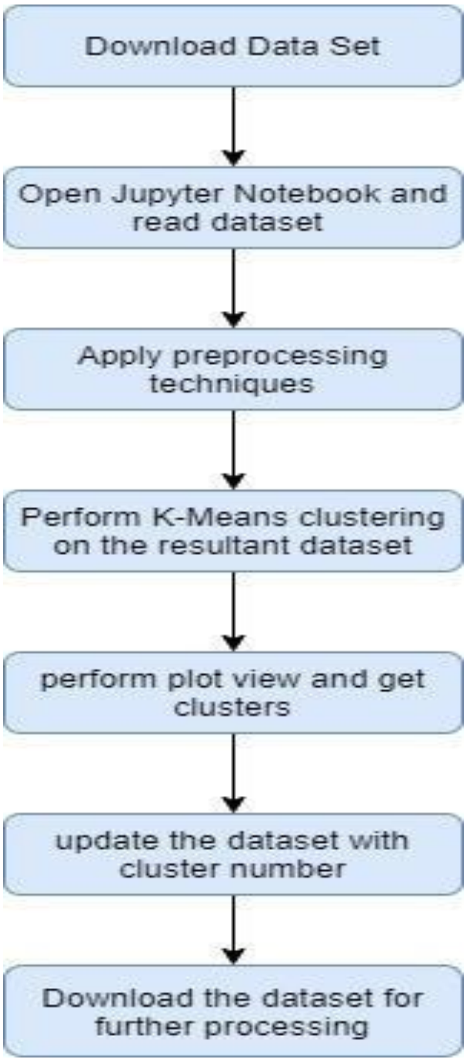

Download Data Set
Open Jupyter Notebook and read dataset
Apply preprocessing techniques
Perform K-Means clustering on the resultant dataset
perform plot view and get clusters
update the dataset with cluster number
Download the dataset for further processing

Fig 1 : crime analysis steps

IV. IMPLEMENTATION

In this phase clustering is done based on the crime data. Using K-Means clustering algorithm the data points are clustered into different clusters. Compared to other clustering techniques, K-Means cluster the data points more accurately with less time. Finally, the clustered data set is updated for further analysis. After loading the data set, noisy and duplicate datas are removed and missing values are properly handled. This data is used for further processing.

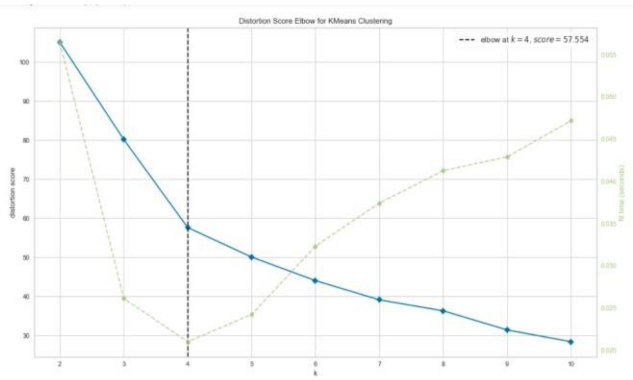


Fig 2: the Elbow curve

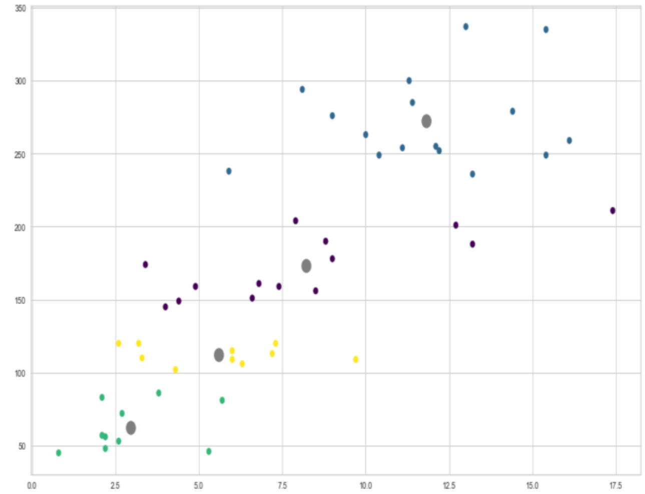


Fig 3 : Clustered data points.

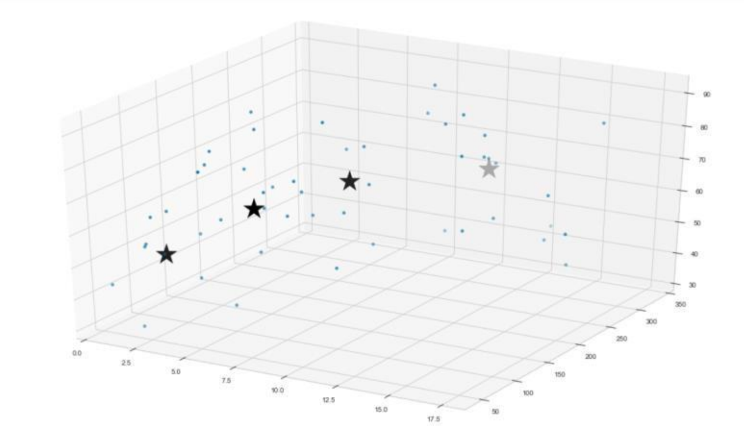
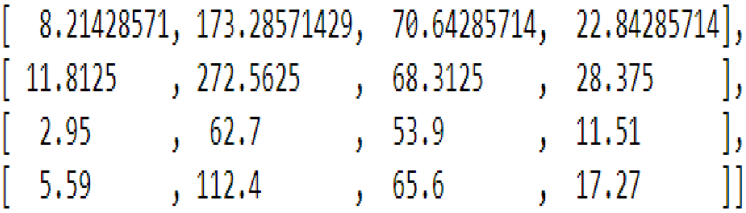


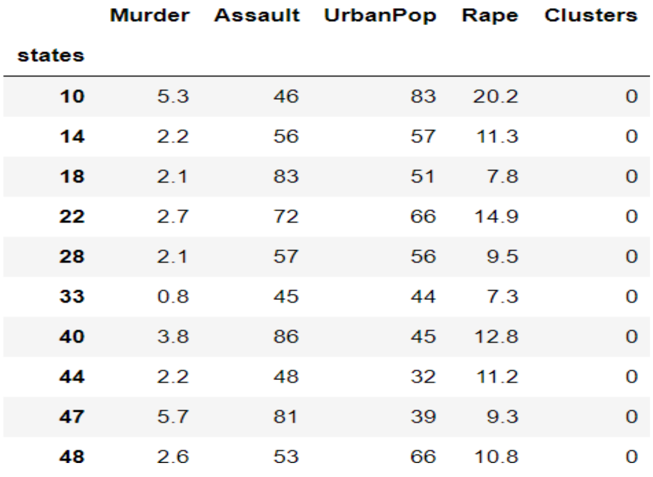
Fig 4 : 3-D visualization of cluster

V. RESULTS



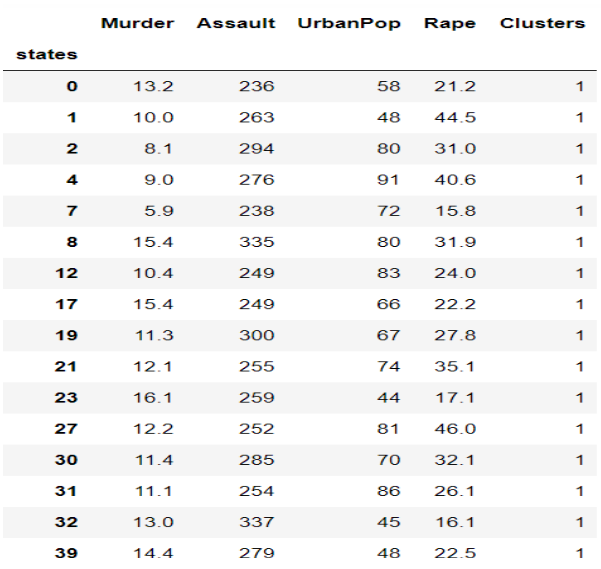
```
[  8.21428571, 173.28571429,  70.64285714,  22.84285714],
[ 11.8125    , 272.5625    ,  68.3125    ,  28.375     ],
[  2.95      ,  62.7       ,  53.9       ,  11.51      ],
[  5.59      , 112.4       ,  65.6       ,  17.27      ]]
```

Fig 5 : Calculated cluster centers



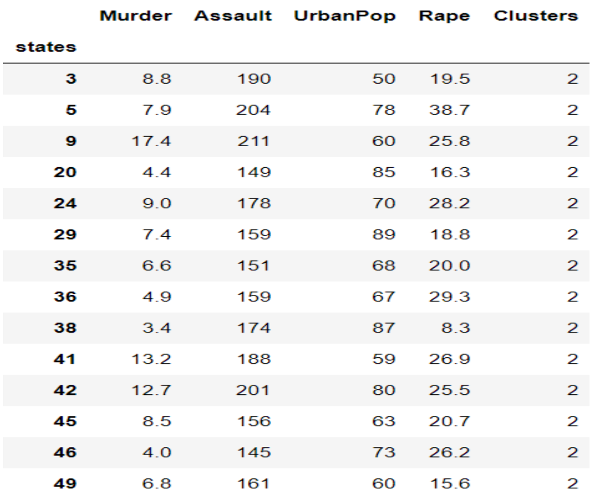
states	Murder	Assault	UrbanPop	Rape	Clusters
10	5.3	46	83	20.2	0
14	2.2	56	57	11.3	0
18	2.1	83	51	7.8	0
22	2.7	72	66	14.9	0
28	2.1	57	56	9.5	0
33	0.8	45	44	7.3	0
40	3.8	86	45	12.8	0
44	2.2	48	32	11.2	0
47	5.7	81	39	9.3	0
48	2.6	53	66	10.8	0

Fig 6 : Cluster 0



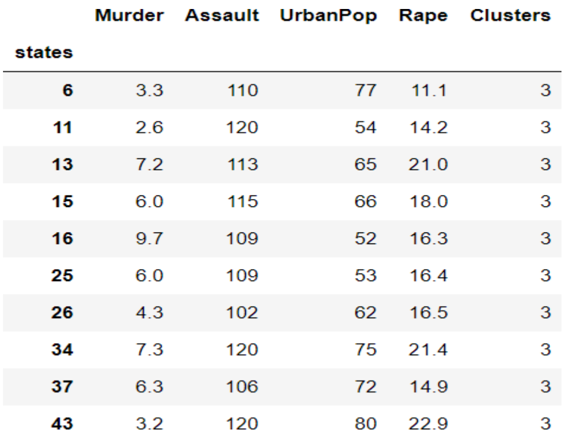
	Murder	Assault	UrbanPop	Rape	Clusters
states					
0	13.2	236	58	21.2	1
1	10.0	263	48	44.5	1
2	8.1	294	80	31.0	1
4	9.0	276	91	40.6	1
7	5.9	238	72	15.8	1
8	15.4	335	80	31.9	1
12	10.4	249	83	24.0	1
17	15.4	249	66	22.2	1
19	11.3	300	67	27.8	1
21	12.1	255	74	35.1	1
23	16.1	259	44	17.1	1
27	12.2	252	81	46.0	1
30	11.4	285	70	32.1	1
31	11.1	254	86	26.1	1
32	13.0	337	45	16.1	1
39	14.4	279	48	22.5	1

Fig 7 : Cluster 1



	Murder	Assault	UrbanPop	Rape	Clusters
states					
3	8.8	190	50	19.5	2
5	7.9	204	78	38.7	2
9	17.4	211	60	25.8	2
20	4.4	149	85	16.3	2
24	9.0	178	70	28.2	2
29	7.4	159	89	18.8	2
35	6.6	151	68	20.0	2
36	4.9	159	67	29.3	2
38	3.4	174	87	8.3	2
41	13.2	188	59	26.9	2
42	12.7	201	80	25.5	2
45	8.5	156	63	20.7	2
46	4.0	145	73	26.2	2
49	6.8	161	60	15.6	2

Fig 8 : Cluster 2



	Murder	Assault	UrbanPop	Rape	Clusters
states					
6	3.3	110	77	11.1	3
11	2.6	120	54	14.2	3
13	7.2	113	65	21.0	3
15	6.0	115	66	18.0	3
16	9.7	109	52	16.3	3
25	6.0	109	53	16.4	3
26	4.3	102	62	16.5	3
34	7.3	120	75	21.4	3
37	6.3	106	72	14.9	3
43	3.2	120	80	22.9	3

Fig 9 : Cluster 3

CONCLUSION

The system is primarily used to cluster the dataset and apply the K-Means algorithm on the crime dataset for the purpose of crime analysis. experimental results show K-Means clustering algorithm clusters the data points much better than any other clustering algorithm. From the clustered results it is easy to identify crime-prone areas. The developed application has promising value in the current complex crime scenario and can be used as an effective tool for the government.

Crime prediction have a promising future. Effective use of crime analysis will help the police force not only to identify the pattern, but it may also help to prevent a crime even before it happens. Advanced K-Means algorithm. Many other algorithms can be used instead of the K-Means clustering algorithm to improve the performance in the future.

REFERENCES

[1] 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.

[2] Chetan G. Wadhai , Tiksha P. Kakade , Khushabu A. Bokde , Dnyaneshwari S. Tumsare, 2018, Crime Analysis Using K-Means Clustering, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) Volume 07, Issue 04 (April 2018),

[3] Sujatha, S., & Sona, A. (2013). New Fast K-Means Clustering Algorithm using Modified Centroid Selection Method. *International journal of engineering research and technology,*

[4] B. Sivanagaleela and S. Rajesh, "Crime Analysis and Prediction Using Fuzzy C-Means Algorithm," 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI), 2019, pp. 595-599, doi: 10.1109/ICOEI.2019.8862691.

[5] S. G. Krishnendu, P. P. Lakshmi and L. Nitha, "Crime Analysis and Prediction using Optimized K-Means Algorithm," 2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC), 2020, pp. 915-918, doi: 10.1109/ICCMC48092.2020.ICCMC-000169.

[6] Sathyadevan, S. et al. "Crime analysis and prediction using data mining." *2014 First International Conference on Networks & Soft Computing (ICNSC2014)* (2014): 406-412.

[7] David. H, Benjamin & Suruliandi, A.. (2017). SURVEY ON CRIME ANALYSIS AND PREDICTION USING DATA MINING TECHNIQUES. ICTACT Journal on Soft Computing. 7. 1459-1466. 10.21917/ijsc.2017.0202.

[8] Yuki, Jesia & Sakib, Md. Mahfil & Zamal, Zaisha & Habibullah, Khan & Das, Amit. (2019). Predicting Crime Using Time and Location Data. 124-128. 10.1145/3348445.3348483.

[9] Chen, Peng & Kurland, Justin. (2018). Time, Place, and Modus Operandi: A Simple Apriori Algorithm Experiment for Crime Pattern Detection. 1-3. 10.1109/IISA.2018.8633657.