Crime Analysis and Prediction using Optimized K-Means Algorithm

Conference Paper · March 2020

DOI: 10.1109/ICCMC48092.2020.ICCMC-000169

Nitha Leeladharan
Amrita Vishwa Vidyapeetham
5 PUBLICATIONS **8** CITATIONS

Proceedings of the Fourth International Conference on Computing Methodologies and Communication (ICCMC 2020)
DVD Part Number: CFP20K25-DVD; ISBN: 978-1-7281-4888-5 pp. 915-918, 9076403

Crime Analysis and Prediction using Optimized K-Means Algorithm

[1]Krishnendu S.G, [2]Lakshmi P.P, [3]Nitha L

[1]PG Student, [2]PG Student, [3]Assistant Professor
Department of Computer Science and IT,
Amrita School of Arts and Sciences,Kochi
Amrita Vishwa Vidhyapeetham, India
[1] krishnendusg1@gmail.com, [2] pplakshmi14@gmail.com, [3]nitha.leeladharan@gmail.com

Abstract-In India, the crime rate is rising each day. In the current situation, recent technological influence, effects of social media and modern approaches help the offenders to achieve their crimes. Both analysis, prediction of crime is a systematized method that classifies and examines the crime patterns. There exist various clustering algorithms for crime analysis and pattern prediction but they do not reveal all the requirements. Among these, K means algorithm provides a better way for predicting the results. Our paper mainly focuses on predicting the region with higher crime rates and age groups with more or less criminal tendencies. We propose an optimized K means algorithm to lower the time complexity and to improve efficiency in the result.

Keywords- Crime, Clustering, Optimized K-Means algorithm.

I. INTRODUCTION

The crime rate continues to grow across all regions in India. Crime analysis methods focus on identifying and predicting patterns and trends in criminology. We use a data mining method to forecast the crime-prone areas which helps the police forces to find various age groups that are more susceptible to crime activities. To accomplish analysis of crimes, an efficient clustering algorithm must be determined. We use the K means clustering method in data mining that categorize a set of objects where objects belonging the same category are similar than other category. There exist many research papers that propose various clustering algorithms that deal with the prediction of criminal incidents.

In this paper, we use an efficacious clustering algorithm for the extraction and interpretation of data to predict the results. K-means is one of the best algorithm 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. As of now, crimes in India are increasing in a rapid rate. Our analysis considers a larger dataset and manually selecting the k value from this huge data set is a very difficult process and takes high time and complexity. To overcome this problem, we choose elbow method to find the k value. To implement optimized K-means clustering in python, we use the spider software.

II. LITERATURE REVIEW

According to Shiju Sathyadevan et al., classification of data is carried out through collection, classification, pattern identification, prediction, and visualization for analyzing crime rates of each state[1]. They used Naive Bayes classifiers for creating a model for classifying each crime. It is a supervised learning problem where you know 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[2]. It uses a heap map for identification of the level of each activities. For example, dark color is represents low activities. A decision tree algorithm can be used to detect suspicious emails using enhanced Iterative Dichotomiser3 algorithm using improved feature selection methods [3]. Application of these important factors produce more desirable and quicker decision trees. The genetic algorithm was used for the optimization of decision tree parameters. Crimes prediction is done by Machine Learning and supervised learning methods and NLP methods are adopted for data classification. Prediction of crime occurrence can also be made through data fusion method with deep neural networks [4]. K-means clustering is used to predict crimes using time and location datasets obtained from the CLEAR system. The dataset consists of facts on crime prevalence and attributes which includes description, month, day, hour, etc. [5]. Mainly, it focuses on the algorithm which can work on combination with categorical and numerical values.

Crime analysis tool is developed using various distinct data mining methods. It supports the police officers for investigating crimes [6]. Implementing a clustering algorithm on crime datasets enables analysis of crimes [7]. It makes identification and analysis of various criminality trends over the years through their conclusion. The random initial starting points produced by K-means which gives results in the form of cluster that helps in reaching the local optima [8]. So to overcome this problem, the partitioned data along with the data axis with the highest variance for assigning the initial centroid for K-Means clustering was applied. So it is observed that the proposed technique uses a lesser number of iteration thereby reducing the clustering time. Using merge sort, K-means algorithm can be improved for clustering the Hidden Markov Model (HMM) [9].

III. PROPOSED SYSTEM

K-MEANS ALGORITHM

In this algorithm, the entire process is divided into two phases.

Phase-I: Initial centroids are calculated in this phase. This is done by using Elbow method algorithm for finding optimum k value

Algorithm 1: The k-means clustering algorithm [4]

Input:

T = {T1, T2... T10}

Output:

Set of 8 clusters.

K=8

Steps:

1. Randomly choose 8 data items from set T and set it as starting centroids;

2. Repeat

3. Compute the distance between centroids and each data item Ti and assign Ti to the cluster which has the nearest centroid.

Phase-II: In phase-II the distance between the starting centroids and each data item is computed. The data points are then grouped with the clusters that has closest centroids. Cluster to which each data point is associated and the distance to it will be recorded. Then the cluster centroid will be recalculated. The above procedure is repeated until the newly obtained cluster value is equal to the previous cluster value.

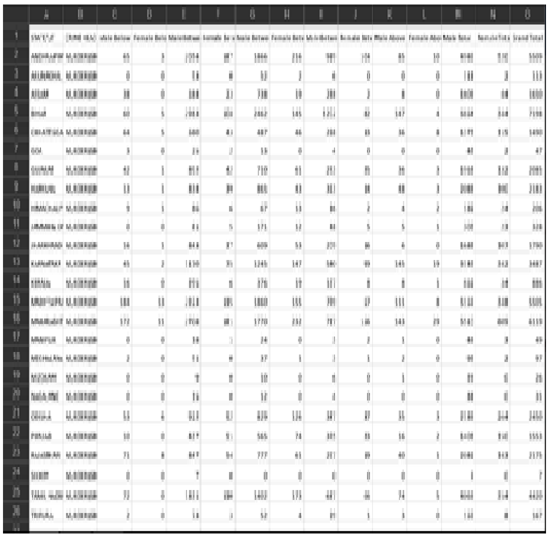


Fig 1 Dataset

We are working on Spyder for implementation. Here we use a Spyder 3.7 version. Spyder is an integrated development environment for systematic programming in Python. Here we implemented different packages like matplotlib,numpy,sklearn, pandas,etc. Which helps to plot elbow graph and data frame table using a K-means clustering algorithm.
Dataset is collected from Kaggle datasets and import datasets into Spyder in CSV format as shown in Fig 1. We perform normalization for finding the accurate number of clusters (k) using the elbow method. The elbow method performs k-means clustering on the obtained dataset for a range of values of k (2-15) and calculates the SSE. A line chart of the SSE is plotted for each value of k. Since there is a slight variation in the elbow graph (Fig 2) at the cluster value 8, it was recognized as the optimized k value. Our goal is to identify a small value of k that still has a low sum of squared errors and the remaining clusters. In Fig. 4, it clearly shows elbow at k=8, indicating that 8 is the best number of clusters. So based on the value of k (that is,8), we are grouping dataset into 8 different clusters.

IV. RESULT & DISCUSSION

The dataset is shown in Fig 1 consists of crimes in I ndia. This dataset contains broad evidence about va rious crimes that took place in India during 2001-20

10. The dataset contain 1053 values and normalization is performed to rescale the dataset.

The resultant pre-processed dataset goes through k-means clustering technique. An initial stage for an unsupervised algorithm should be finding out the optimal number of clusters into which the data can be clustered Plot the curve of total_within_ss according to the number of clusters k.To find out the best and accurate k value we randomly give numbers from 2 to 15.Here from this graph we analyze the changes that happened in
8th value so we take the value as k=8.

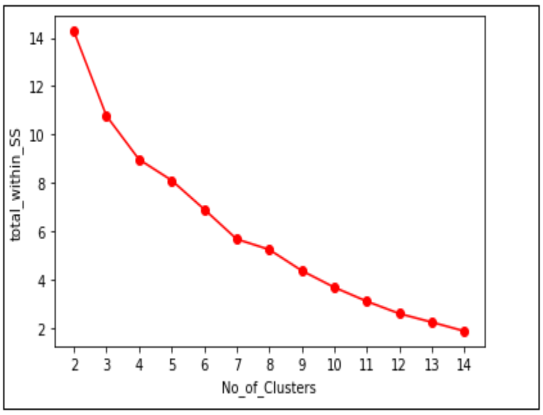
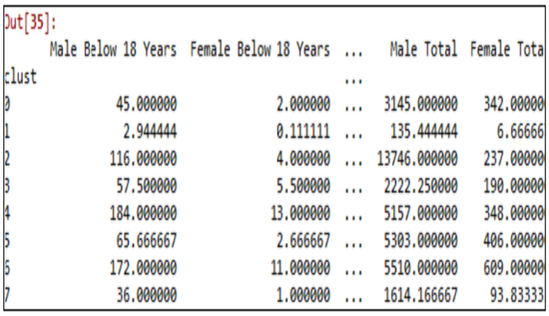


Fig 2 Elbow graph

In the elbow graph, X-axis contains no_of_clusters and y-axis contains total_within_ss which means, it is the variable for storing total within sum of squares for each k-means. The plot contains a bend (knee) location which is generally indicate the number of clusters. ie k=8.From the Fig 2 graph, here we choose k from 2 to 8 for iteration.
After finding the optimum k value the dataset is formed to k-means clustering



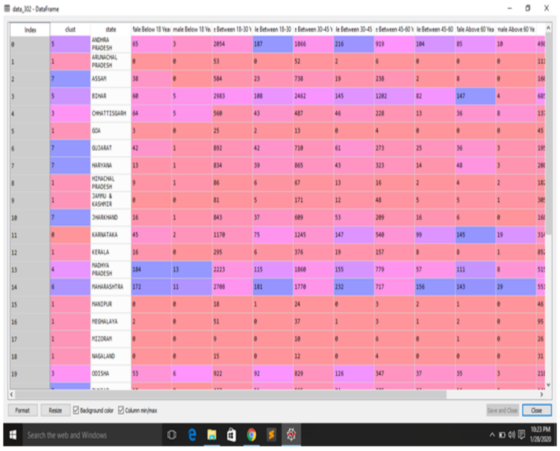
```
Out[35]:
      Male Below 18 Years  Female Below 18 Years  ...   Male Total  Female Tota
clust                                              ...
0               45.000000               2.000000  ...  3145.000000  342.00000
1                2.944444               0.111111  ...   135.444444    6.66666
2              116.000000               4.000000  ... 13746.000000  237.00000
3               57.500000               5.500000  ...  2222.250000  190.00000
4              184.000000              13.000000  ...  5157.000000  348.00000
5               65.666667               2.666667  ...  5303.000000  406.00000
6              172.000000              11.000000  ...  5510.000000  609.00000
7               36.000000               1.000000  ...  1614.166667   93.83333
```

Fig 3 Clustering result

The clustering result are demonstrate below,

Total male criminals are more in cluster 2 ,i.e. value 13746.
Female criminals are more in cluster 6
Male below 18 years age are more in cluster4.
Female below 18 years age are more in cluster 4.

After clustering fig 4 shows the added clusters as the second attribute to the original dataset and it is shown below



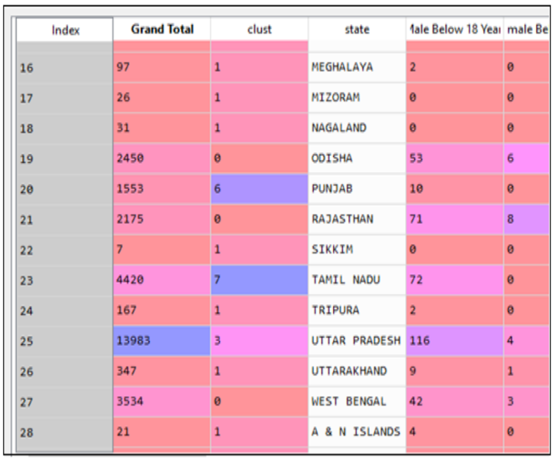
data_302 - DataFrame

Index	clust	state	lale Below 18 Year	male Below 18 Ye	e Between 18-30 Y	le Between 18-30	e Between 30-45 Y	le Between 30-45	e Between 45-60 Y	le Between 45-60	lale Above 60 Yea	male Above 60 Ye	
0	5	ANDHRA PRADESH	65	3	2054	187	1866	216	919	104	85	10	49
1	1	ARUNACHAL PRADESH	0	0	53	0	52	2	6	0	0	0	11
2	7	ASSAM	38	0	584	23	738	19	238	2	8	0	16
3	5	BIHAR	60	5	2983	108	2462	145	1202	82	147	4	68
4	3	CHHATTISGARH	64	5	560	43	487	46	228	13	36	8	13
5	1	GOA	3	0	25	2	13	0	4	0	0	0	45
6	7	GUJARAT	42	1	892	42	710	61	273	25	36	3	19
7	7	HARYANA	13	1	834	39	865	43	323	14	48	3	20
8	1	HIMACHAL PRADESH	9	1	86	6	67	13	16	2	4	2	18
9	1	JAMMU & KASHMIR	0	0	81	5	171	12	48	5	5	1	30
10	7	JHARKHAND	16	1	843	37	609	53	209	16	6	0	16
11	0	KARNATAKA	45	2	1170	75	1245	147	540	99	145	19	31
12	1	KERALA	16	0	295	6	376	19	157	8	8	1	85
13	4	MADHYA PRADESH	184	13	2223	115	1860	155	779	57	111	8	51
14	6	MAHARASHTRA	172	11	2708	181	1770	232	717	156	143	29	55
15	1	MANIPUR	0	0	18	1	24	0	3	2	1	0	46
16	1	MEGHALAYA	2	0	51	0	37	1	3	1	2	0	95
17	1	MIZORAM	0	0	9	0	10	0	6	0	1	0	26
18	1	NAGALAND	0	0	15	0	12	0	4	0	0	0	31
19	3	ODISHA	53	6	922	92	829	126	347	37	35	3	21

Fig 4 Data frame table based on clustering

This implementation result is the case study of the year 2001-2010.It shows the criminal behavior of all-state. From the case
study, it is clearly understood and reached into the conclusion that most of the male criminals are from MADHYA PRADESH and female criminals are from MAHARASHTRA.

The male and female criminals below age 18 are more in cluster 4 i.e. from CHHATTISGARH.



Index	Grand Total	clust	state	lale Below 18 Year	male Be
16	97	1	MEGHALAYA	2	0
17	26	1	MIZORAM	0	0
18	31	1	NAGALAND	0	0
19	2450	0	ODISHA	53	6
20	1553	6	PUNJAB	10	0
21	2175	0	RAJASTHAN	71	8
22	7	1	SIKKIM	0	0
23	4420	7	TAMIL NADU	72	0
24	167	1	TRIPURA	2	0
25	13983	3	UTTAR PRADESH	116	4
26	347	1	UTTARAKHAND	9	1
27	3534	0	WEST BENGAL	42	3
28	21	1	A & N ISLANDS	4	0

Fig 5 Data frame table based on clustering

Fig 5 clearly shows the total number of criminals is more in UP i.e. 13983 criminals. This result will help the police departments and other authorities to be more aware of the criminal behaviors and also to plan different strategies according to this result to prevent future crimes.

CONCLUSION

The implementation has been done in Python language. Here we find out which state has more or fewer criminals occur based on each cluster's values. It is really helpful for the authorities to be more aware. Clustering is the process of creating different groups consisting of similar data points. The points in one cluster are as similar as possible are there are no similarities between different cluster points. The result of the optimized k-means algorithm is efficient and provides improved accuracy of the final cluster reduced the number of iterations. In the future, the result of crime analysis can be used to make various strategies for crime control and the optimal deployment of resources in crime avoidance.

REFERENCES

1. Shiju Sathyadevan M.S, Surya Gangadharan: Crime Analysis and Prediction Using Data Mining,in NetworksSoft Computing(ICNSC),(2014) First International Conference. https://ieeexplore.ieee.org/document/6906719.
2. H. Benjamin Fredrick David1,A. Suruliandi: Survey on crime analysis and prediction using data mining techniques.Department of Computer Science and Engineering,Manonmaniam Sundaranar University, India. Ictact journal on soft computing, april(2017). https://www.researchgate.net/publication/3222541877_SURVEY_ON_CRIME_ANALAYSIS_AND_PREDICTION_USING_DATA_MINING_TECHNIQUES.
3. JesiaQuader Yuki, Md.MahfilQuaderSakib, ZaishaZamal, Khan Mohammad Habibullah, Amit Kumar Das: Predicting Crime Using Time and Location Data(2019). https://www.researchgate.net/publication/335854157_Predicting_Crime_Using_Time_and_Location_Data.
4. Peng Chen,Justin Kurland, Modus Operandi: Time,Place,A Simple Apriori Algorithm Experiment for Crime Pattern Detection(2018).9th International Conference on IISA. https://www.researchgate.net/publication/330877862_Time_Place_and_Modus_Operandi_A_Simple_Apriori_Algorithm_Experiment_for_Crime_Pattern_Detection.
5. JyotiAgarwal,RenukaNagpal, RajniSehgal: Crime Analysis using K-Means Clustering(2013). https://www.researchgate.net/publication/269667894_Crime_Analysis_using_K-Means_Clustering
6. Malathi. A & Dr. S. SanthoshBaboo: An Enhanced Algorithm to Predict a Future Crime using Data Mining. International Journal of Computer Applications (0975 – 8887) Volume 21– No.1, May 2011
7. Khushabu A. Bokde,Tiksha P. Kakade, Dnyaneshwari S. Tumsare, ChetanG.Wadhai: Crime Analysis Using K-Means Clustering(2018).https://www.ijert.org/crime-analysis-using-k-means-clustering.
8. Mrs.S. Sujatha, Mrs. A. Shanthi Sona: New Fast K-Means Clustering Algorithm using Modified Centroid Selection Method(2013).
9. Keerthi A, Remya M S, Nitha L: Detection of Credit Card Frauds Using Hidden Markov Model With Improved K-Means Clustering Algorithm (2015). https://www.ijarcs/article/download/1239/1227.