2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO) Amity University, Noida, India. June 4-5, 2020

Criminal Behaviour Analysis and Segmentation using K-Means Clustering

Gouri Jha AIIT, Amity University Amity, Sector 125, Noida, Uttar Pradesh gouri.jha1@student.amity.edu Laxmi Ahuja AIIT, Amity University Amity, Sector 125, Noida, Uttar Pradesh lahuja@amity.edu Ajay Rana AIIT, Amity University Amity, Sector 125, Noida, Uttar Pradesh ajay_rana@amity.edu

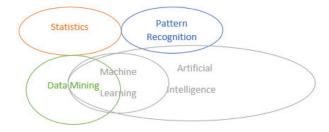
Abstract— Crimes are increasing with a high frequency rates in this new era of world and hence it's a devastating issue that everyone has been experiencing. For finding a patter that can be used for prediction is necessary. The objective of this paper is to understand the concept of data mining and machine learning which can be used for finding criminal patterns and behaviours. The paper is further divided by providing basic differentiation of the clustering techniques used in unsupervised learning. And then after the crime dataset of India that contains record of serious fraud of property in all states and we will apply k-means clustering to find generic patterns. The main reason for this paper is to give a quip thought of how machine learning can be utilized by the law authorization to distinguish, anticipate and illuminate violations by a lot quicker rate.

Keywords— Machine Learning, kmeans clustering, centroids, fuzzy, C Means Algorithm, Expectation-Maximisation Algorithm.

I. INTRODUCTION

Crimes are something that has been experienced by everyone at some point of time in their lives. There is not a particular definition of what crime is but obviously, it Is a harm and threat to the mankind. There can be different types of crimes like murders, rape, child abuse, kidnapping, robbery, physical attack, imprisonment, buggery, homicide and many more. Crimes are growing rapidly in every states of every country with a much faster frequency rate and police or FBI departments are struggling to obtain some patterns and solving crimes with a vast number of datasets. And therefore, there is a need to identify these crimes based on the pattern recognition and making predictions by applying some data science techniques and methods. There are two terms data mining and machine learning that are somehow related with each other which could be useful in crime analysis and segmentation. With the help of machine learning we can evaluate the type of crimes which could happen in a particular area. And therefore, we are going to use K-Means algorithm for criminal behaviour analysis and segmentation in India.

Data mining is basically a process of extracting information from the database or any set of data by applying some knowledge to produce a series of output. It is nothing but a procedure of ascertaining knowledge and finding patterns from the raw data to obtain or make a helpful decision. It involves some basic steps- First, understanding your work objective i.e., defining brief plan and set data mining goals. Second, collection of data and envision, asking questions, and inspection of data comes under highlight. After examining the data, preparation of fata takes place in which the data scientist tends to explore and clean the data and keep only the essential data which could be helpful for future process. And after that, modelling techniques needs to be selected and here comes some tools and techniques that comes under machine learning algorithm. After applying data modelling the data is evaluated and then deployed. And, on the other hand machine learning involves some algorithms and techniques applied on data to increase experience. So, Machine Learning is a technique or approach of data mining to obtain the properties of the raw data and acquiring knowledge and therefore improving experience and based on that experience it generates prediction. There are many machine learning algorithms are available but the main two are categorized as supervised and unsupervised learning. In supervised learning, the data is trained with a right and wrong answer, it maps an input to an output. Whereas, in unsupervised learning, there is no prior-knowledge and it has to learn on its own. Google News is a perfect example of unsupervised learning as it is based on clustering technique.





II. TYPES OF MACHINE LEARNING:

One of the methodologies in Machine Learning calculation is prepared utilizing an input as labelled or unlabelled data then train on the basis of that to produce a desired output. Basically, it is categorized on the factor that training of dataset. Supervised and unsupervised learning are the two most important type of machine learning techniques which can be used in different given dataset and different problems/scenarios.

A. Supervised Learning

As the name suggests it works under supervision, that is its learning in which machine is trained by dataset which is labelled. Data in which you know the target answer as the dataset which is given to you by some predefined leaning. is called as labelled data. The machine is provided by some labelled data and on the basis of that dataset, it associates, understands and learn different pattern and relationship between the dataset and hence when in future new information is provided it produces the output based on its past learning experience. Supervised Learning is used in risk assessment like in financial companies or in an insurance company to minimise the risk domain. Image classification is also one of the most used applicational area of supervised learning. For example, Facebook recognizes your friends based on tagged photos. Further, Supervised Machine Learning (SML) is classified into two categories, i.e., Classification and Regression. Classification is basically predicting the desired learning based on a TRUE/FALSE or YES/NO condition in which the output is discrete. For example- either a disease is under this further category or not. Whereas, in Regression technique, the output variable is continuous, it focuses more on labelled data to produce the quantity. For example- it can be used to predict age, salary, amount, weight or any numerical data which is termed as continuous.

B. Unsupervised Learning

In Unsupervised Learning, no supervision is granted to the machine and allowing it to work on the data which is not labelled and learn itself. Hence, machine tries to identify patterns on its own and give response. Further, Unsupervised Learning can be classified into two types- Clustering and Association. In clustering, Machine forms the group based on their patterns and put data with similarities in one cluster and those with dissimilarities in another. Whereas, in association learning techniques there is a discovery-based rules are followed to find the association in a large dataset.



Fig. 2. Types of Machine Learning

III. CLUSTERING ALGORITHMS IN MACHINE LEARNING

Clustering is a technique that is used in unsupervised machine learning algorithm, it's based-on grouping and putting together similar kind of entities together in such a way that they share some similarities in data points. As, we know that in unsupervised machine learning, the machine has to learn on its own and hence clustering helps us to insight different pattern that is present in different groups. By this definition, it is clear that the data points that resides on the different clusters or groups are very dissimilar with each other whereas the data point that underlies on the same cluster generally shares same properties within each other. This is the main idea behind Unsupervised Learning, grouping the data with respect to their similarities including them and based on the requirement given by the user. There are different clustering techniques available in machine learning, some of them are as follows:

A. K-Means Clustering

K-Means clustering algorithm comes under unsupervised machine learning and it is by far the most widely used as well as popular algorithm.

- i. The process follows easy and yet effective method of dividing the whole data sets into a number of clusters k.
- After defining number of clusters, we need to randomly assign k centroids for each and every cluster.

- iii. The next step is to assign every data point to the nearest centroid and we compute the mean of all data points in one cluster and move the centroid to computed distance and form a new centroid for each cluster.
- iv. Repeat step 3 and recompute the centroid and do move centroid until cluster centroid do not change any further.
- v. If the result is coming same even after doing multiple iterations then we must stop the process and conclude the algorithm is not discovering any new pattern and process can be terminated.

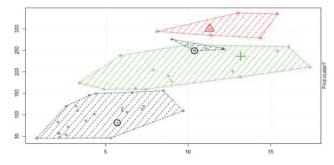


Fig. 3. K-Means Clustering

1) Algorithm for K-means Clustering: Assume that we have z sample feature vectors a1, a2,, az and we already have the knowledge that these vectors lie onto z clusters, z < a. Assume that the m' is the mean of the vectors in cluster *i*.

- Make initial estimation for the means m1, m2,, mz
- Repeat this process till there are no changes in any mean
- Use the estimated means to classify the samples into clusters

For i from 1 to z

Put back m' with the mean of all of the samples for cluster i

end for end until

B. FUZZY, C Means Algorithm (FCM)

As, the name suggests Fuzzy C Means Clustering is a logic based on confusion of selecting and assigning clusters to the data point. In thus method, one data point can belong to two or more clusters The difference is that in case of Kmeans, each element is assigned to only a single cluster, while in case if C-means, being a Fuzzy clustering technique, each element is assigned to all the available clusters with a different membership degree for each cluster. It is an advance version of K-Means clustering algorithm. The centroids are discovered are dependent on the fuzzy coefficient which surveys the quality of participation of information in a group. FCM implies the process to be followed as a comparable way to deal with that of k-implies aside from that it varies in the figuring of fuzzy coefficients and gives out a likelihood appropriation result. The grouping strategy is useful to explore the criminal occasions that are connected with the same criminals and furthermore helps to find Serial killings related crime that has been unknown.

C. Expectation-Maximisation (EM) Algorithm

This algorithm basically solves the data samples with the help of the principle of Gaussian distribution in Statistics. It is not much useful when your datasets have some missing data points, hidden samples or any uncertainty and hence has its biggest limitations as it works better when your data set doesn't have big dimensions. It uses an iteration of finding likelihood estimation based on Gaussian distribution method. It works by picking arbitrary values for the missing information data-points, and utilizing those theories to gauge a second arrangement of information. The new values are utilized to make a superior prediction for the primary set, and the procedure proceeds until the calculation meets on a fixed point. Useful for analysis of crimes in which tracking of location is essential.

D. Hierarchical Clustering Algorithms

And coming to the hierarchical clustering algorithm, it basically works on the fact that it sorts the clusters depend on the similarities observation and then put that in sorted hierarchy. Its fundamental process is performed by the approach that it assigns every data point in one cluster then it divides the cluster depended on the dissimilarities among each other. And then it after dividing it aggregates the cluster based on likewise similarity. It generally has two types agglomerative which is also known as bottom-up clustering and divisive which is referred as top-down clustering.

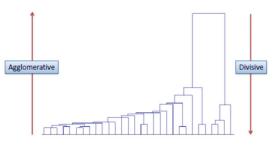


Fig. 4. Fig. 4 Hierarchical Clustering

IV. CRIME CLASSIFICATION IN INDIA

As, it has stated before that this paper is going to focus on the criminal behaviour prediction and segmentation in India. In this chapter, the main problem, procedure and the technique will be explained and thereafter the results that are obtained will be discussed.

A. Description of the problem

Crime is something which affects not only locally but also when it is accelerating it becomes a worldwide phenomenon. In any case, what concerns us today and make it the significant open issue is that its frequency has been going up in all piece of the world and among all sections of society. The ongoing development pace of wrongdoing has not exclusively been uncommon yet additionally has been quickening. There are so many research papers about different countries and their crime data analysis and there are so few that generally focuses on India. So, in this paper we will going to analysis and apply k-means clustering for segmentation of different states of India and their respective crime reports.

B. Dataset Analysis

In this report, the dataset consist data about various aspect of crimes that has been happened in India from the year 2001. Crime consist – serious fraud of property in all states of India that has been recorded and it contains columns such as Area Name(states in which those crimes happened), Years(from 2001), Group Name(cheating or criminal Breach of trust), subgroup name and then further four columns are divided on the basis of loss of property in crores.

A Area_Name 🖓	# Year 🖓	A Group_N_ T	A Sub_Grou F	A Loss_of_ =	A Loss_of =	A Loss_of	A Loss_of_ =	A Loss_of 7
Andhra Pradesh	2001	Serious Fraud - Cheating	2. Cheating	4	0	0	0	0
Arunachal Pradesh	2001	Serious Fraud - Cheating	2. Chesting		0	0	*	
Assan	2001	Serious Fraud - Cheating	2. Chesting	e	9	0	0	e
Bihar	2001	Serious Fraud - Cheating	2. Cheating	8	0	0	1	0
Chandigarh	2001	Serious Fraud - Cheating	2. Cheating	8	0	9	*	*
Chhattisgarh	2001	Serious Fraud - Chesting	2. Cheating	e.	9	0	0	0
Delhi	2001	Serious Fraud - Chesting	2. Cheating	2	NULL	NULL	NULL	NULL
Goa	2001	Serious Fraud - Cheating	2. Cheating	2	NULL	NULL	NULL	NULL
Gujarat	2001	Serious Fraud - Cheating	2. Chesting	14	0	1	1	0
Haryana	2001	Serious Fraud - Cheating	2. Cheating	8	0	0	1	8
Himachal Pradesh	2001	Serious Fraud - Cheating	2. Cheating	£.	0	0		8
Jammu & Kashmir	2001	Serious Fraud - Dheating	2. Cheating	8.	0	0	8	0
Jharkhand	2001	Serious Fraud - Cheating	2. Cheating	£	0	2	1	¥
Karnataka	2001	Serious Fraud - Cheating	2. Chesting	52	0	0		*
Maharashtra	2001	Serious Fraud - Cheating	2. Chesting	E.	1	.0	*	R
Meghalaya	2001	Serious Fraud - Chesting	2. Chesting	0	0	0		0

Fig. 5. Serious Fraud Dataset

If we draw various designs and graphs, we will be able see more obviously how the information of the given dataset is circulated.

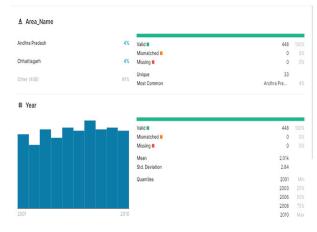


Fig. 6. Graph of the Dataset

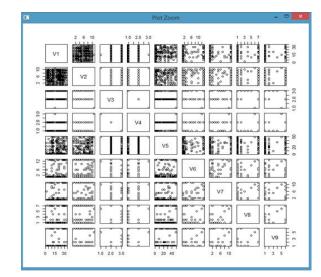


Fig. 7. Dataset Plot

V. IMPLEMENTATION

The dataset that has been used in this paper is taken from Kaggle.com, under the dataset names as crime in India. The steps that are used in the implementation are further divided in the following steps:

A. Collection of data

The data that is used in analysis is in the csv format.

B. Pre-Processing of the data

About 1469 columns are present and thousands of entries are present in the dataset. Data pre-processing involves converting any string variable to the numerical one so that it gets easy for evaluation.

1) Missing value Imputation

So, first thing that we need to do is to assign new values for null values because Kmeans clustering is not able to deal with null information. If there is even missing data then it'll be difficult to deal with that. There are so many methods and commands are available to deal with missing value yet there must be taken care that misleading values should not be provided as it can distort the computation. This formula for imputing missing value is used in this paper.

Dataset_name\$column_name=ifelse(is.na(Dataset_name\$ column_name),ave(Dataset_name\$column_name, FUN = function(x) mean(x, na.rm = 'TRUE')),Dataset_name\$column_name)

2) Dealing with categorical variables

K-Means utilizes separation calculation at centre of its calculation, and consequently can't deal with categorical variable legitimately. There are two possible methods are there in which categorical variables can be handled – First, if cardinal variables are there then it must be converted to binary numerical values of all the classes of categorical values that are there. Second- if the ordinals variables are there then it should convert to arithmetic sequence of proper difference.

3) Arbitrary initialization of centroids

In kmeans clustering, initialization plays an important role as it is a major part to start and give better iteration later. Bad grouping of centroids may result in bad clusters and therefore will not give a better outcome.

C. Feature Selection

Feature selection is required to done to build a model. Group Name (cheating or criminal Breach of trust), and then further four columns are divided on the basis of loss of property in crores are selected as attribute for feature selection.

After pre-processing of data is performed and the null values has been removed, the number of clusters is chosen as 4. Kmeans() work takes the info information and the quantity of groups in which the information is to be bunched. The punctuation is: kmeans(information, k) where k is the quantity of bunch focuses.

> cl <- kmeans(crime, 4)

> class(cl)

^	V1 ⁰	V2 °	V3 °	V4 ⁰	V5 [©]	V6 °	V7 ⁰	V8 °	V9
1	15	11	1	3	49	12	10	6	5
2	2	1	2	2	32	1	1	1	1
3	3	1	2	2	1	1	1	1	1
4	4	1	2	2	1	1	1	1	1
5	5	1	2	2	1	1	1	1	1
6	6	1	2	2	1	1	1	1	1
7	7	1	2	2	1	1	1	1	1
8	10	1	2	2	13	13	11	7	6
9	11	1	2	2	13	13	11	7	6
10	12	1	2	2	7	1	2	2	1
11	13	1	2	2	1	1	1	1	1
12	14	1	2	2	1	1	1	1	1
13	16	1	2	2	1	1	1	1	1
14	17	1	2	2	1	1	5	1	1
15	18	1	2	2	39	1	1	1	1
16	21	1	2	2	47	2	1	1	1
17	23	1	2	2	1	1	1	1	1
18	25	1	2	2	43	1	1	1	1
19	26	1	2	2	1	1	1	1	1
20	28	1	2	2	16	13	11	7	6
21	30	1	2	2	38	13	11	7	6

Fig. 8. Dataset after Pre-processing

	thributes		9	
iame	Type	Value		
O d	list (9) (S3: kmeans)	List of length 9		
O cluster	integer [449]	341111_		
centers	double [5 x 9]	8.20 24.79 23.89 13.02 24.02 5.46 5.56 6.49 6.08 5.48 2.51 2.55 2.29 2		
totss	double [1]	150414.4		
withinss	double [5]	9930 7846 5313 5440 8180		
totwithinss	double [1]	36709.02		
betweenss	double [1]	113705.3		
size	integer [5]	173 139 35 48 54		
iter	integer [1]	3		
6				
attr(*, "na \$ centers :	Named int [1:449] 3 4 1 mes")= chr [1:449] "1" num [1:5, 1:9] 8.2 24.8 mnames")=List of 2 :5] "1" "2" "3" "4"	"2" "3" "4" 5 23.9 13 24		

Fig. 9. clusters summary

The str() work gives the structure of the kmeans which incorporates different parameters like withinss, betweenss, and so on, breaking down which you can discover the presentation of kmeans. Where, totwithinss referred as sum of all the withinss of the considerable number of groups i.e. Total intra-bunch comparability, withinss can be described as within aggregate of square for example Inter cluster comparability and betweenss is the between aggregate of squares for example Intracluster comparability.

Elbow curve: This is the plot between 'k', the quantity of bunches and the 'totwithinss' for each estimation of k. You can see when the quantity of bunch is less, there is a progressive decline in twisting yet as we continue expanding the estimation of k, the pace of decrease of mutilation esteems gets consistent.

To get the clustered results, we applied this formula

> library(animation)

> cl<- kmeans.ani(crime, 4)</p>

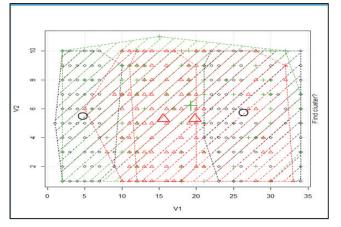


Fig. 10. Output cluster

And thus, we have determined a pattern that the crime of loss of property in less than 10 crores is more and greater in each state and it is increasing per year in every states of India. And we found a similarity between different crimes in each state with their ratio happening per year. We can also assign each state a cluster and then predict each one of them by ranking of their crimes. Each cluster with their sizes and details are represented in the image given below:

```
> LL
Error: object 'CL' not found
> c1
K-means clustering with 5 clusters of sizes 35, 53, 157, 156, 48
Cluster means:
                               V3
                     V2
                                          V4
                                                     V5
                                                                V6
                                                                           V7
          V1
                                                                                     V8
1 23.885714 6.485714 2.285714 1.714286 42.942857 5.914286 4.914286 3.600000
2 24.150943 5.471698 2.528302 1.471698 17.000000 6.547170 5.547170 3.547170
3 23.789809 5.566879 2.541401 1.458599 1.585987 2.101911 1.891720 1.515924
4 7.346154 5.448718 2.506410 1.493590 1.916667 1.961538 1.878205 1.519231
5 13.020833 6.083333 2.416667 1.583333 31.520833 3.895833 3.312500 2.270833
         V9
1 3.171429
2 3.094340
3 1.433121
4 1.461538
5 2.083333
```

Fig. 11. Cluster Data

VI. CONCLUSIONS

With the advancement in technologies that are coming recently in data science and especially in machine learning, it becomes easy and efficient to discover patterns and information which might get useful for future prediction in crime analysis and behaviour segmentation. Clustering is the process of grouping similarities in a dataset so that It can get useful for analysis, discovering patterns and prediction. In this paper, different clustering techniques has been discussed alongside with basis introduction of machine learning and its types. The main work in this paper mainly revolves around the crime dataset of India and basic analysis and segmentation has been performed. Different algorithm for analysis has been reviewed and kmeans clustering algorithm is used to discover general patterns and groups in the theft record of Indian crime dataset. In future, we tend to focus to examine fuzzy clustering algorithm to apply on serial killing related dataset and examine the analysis of the same.

ACKNOWLEDGMENT

Hereby, we would like to sincere acknowledge the institution for their immense support. As, well as for all the papers that has been referenced in here gave us all the immense knowledge that web needed for this paper.

REFERENCES

- Yadav, S., Timbadia, M., Yadav, A., Vishwakarma, R., & Yadav, N. (2017, April). Crime pattern detection, analysis & prediction. In Electronics, Communication and Aerospace Technology (ICECA), 2017 International conference of (Vol. 1, pp. 225-230). IEEE.
- [2] Suhong Kim ; Param Joshi ; Parminder Singh Kalsi ; Pooya Taheri. Crime Analysis Through Machine Learning, 2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON).
- [3] Shamsuddin, N. H. M., Ali, N. A., & Alwee, R. (2017, May). An overview on crime prediction methods. In Student Project Conference (ICT-ISPC), 2017 6th ICT International (pp. 1-5). IEEE.
- [4] Nath, S. V. (2006, December). Crime pattern detection using data mining. In Web intelligence and intelligent agent technology workshops, 2006. wi-iat 2006 workshops. 2006 ieee/wic/acm international conference on (pp. 41-44). IEEE.
- [5] Suhong Kim, Param Joshi, Parminder Singh Kalsi, and Pooya Taheri Fraser. Crime Analysis Through Machine Learning, November 2018, International College, Simon Fraser University.
- [6] Alkesh Bharati, Dr Sarvanaguru RA.K, Crime Prediction and Analysis Using Machine Learning, International Research Journal of Engineering and Technology (IRJET), Volume: 05 Issue: 09 | Sep 2018.
- [7] Arpita Nagpal, Aman Jatain, Deepti Gaur. Review based on Data Clustering Algorithms, Proceedings of 2013 IEEE International Conference on Information and Communication Technologies (ICT 2013).
- [8] Rasoul Kiani, Siamak Mahdavi, Amin Keshavarzi. Analysis and Prediction of Crimes by Clustering and Classification, (IJARAI) International Journal of Advanced Research in Artificial Intelligence, vol 4, No.8, 2015.
- [9] Lawrence McClendon and Natarajan Meghanathan*. USING MACHINE LEARNING ALGORITHMS TO ANALYZE CRIME DATA, Machine Learning and Applications: An International Journal (MLAIJ) Vol.2, No.1, March 2015.
- [10] Sathyadevan, S., & Gangadharan, S. (2014, August). Crime analysis and prediction using data mining. In Networks & Soft Computing (ICNSC), 2014 First International Conference on (pp. 406-412). IEEE.
- [11] N. Tyagi, A. Rana, "Fuel your growth with integration: Hybrid cloud computing", in International Journal of Applied Engineering Research, Vol. 10, Isuue 13, pp 32761-32762 (2015).
- [12] A. Singh, A. Rana, J. Ranjan, "Proposed analytical customer centric model for an automobile industry", in International Journal of Data Mining, Modelling and Management, Vol. 7, Isuue 4, pp 314-330 (2015).
- [13] A. Singh, A. Rana, J. Ranjan, "Data mining techniques and its effect in customer relationship management", in International Journal of Data Analysis Techniques and Strategies, Vol. 7, Isuue 4, pp 406-427 (2015).
- [14] G. Dubey, A. Rana, N. K. Shukla, "User reviews data analysis using opinion mining on web", in 2015 1st International Conference on Futuristic Trends in Computational Analysis and Knowledge Management, ABLAZE 2015, pp 603- 612 (2015).
- [15] A. Rana, S. P. Singh, R. Soni, A. Jolly, "Incremental software version release planning for adaptive maintenance", in International Journal of Software Engineering and its Applications, Vol. 9, Isuue 8, pp 217-228 (2015).
- [16] A. Singh, A. Rana, "Generate frequent queries for views in a data warehouse using data mining techniques", in ACM International Conference Proceeding Series (2014).