An Analysis of Cyber Crime Data

Jyoti Singh, Department of Computer Science & Engineering, RDEC, Ghaziabad

Abstract

Cyber Crime is technology based crime committed by technocrats. This paper deals with Variants of cyber crime held in Chhattisgarh between 2005to 2019. Under this, the Age wise Clustering of arrested people has been displayed on basis of cybercrime in Chhattisgarh .data mining DBSCAN and Hierarchical algorithm is used for clustering. A DBSCAN algorithm is based on this intuitive notion of "clusters". The key idea is that for each point of a cluster, the neighborhood of a given radius has to contain at least a minimum number of points. A Hierarchical clustering method works via grouping data into a tree of clusters. Hierarchical clustering begins by treating every data points as a separate cluster. Python Software has been used to implement the DBSCAN and Hierarchical Clustering Algorithm in cyber crime dataset.

Keywords: Cyber Crime, Types of Cyber Crime, DBSCAN And Hierarchical Clustering Algorithm, Python, Cyber Crime Dataset, Result Analysis.

1. Introduction

Cyber crime always involves some degree of infringement on the privacy of others or damage to computer-based property such as files, web pages or software. This paper is completely focused on cyber crime case register and number of person arrested in Chhattisgarh. The paper also includes Chhattisgarh cybercrime Statistics according age wise people arrested. According to the age of the arrested person based on cyber crime in Chhattisgarh from 2005 to 2019, the clustering has been made through the DBSCAN and Hierarchical Clustering algorithm which is based on cyber crime dataset. We can do DBSCAN and Hierarchical Clustering algorithm using python.

2. Methodology

Cluster analysis or clustering is the process of partitioning a set of data objects into subsets. Each subset is a cluster, such that objects in a cluster resemble one another, yet dissimilar to objects in other clusters. In this context DBSCAN, Hierarchical clustering methods may generate different clustering's on the Cyber Crime dataset. The partitioning is not performed by humans, but by the clustering algorithm DBSCAN, Hierarchical clustering algorithms were used for formation of clusters on cyber crime database. The data was collected from the National crime record bureau (2005 to 2019) data set converted into iris dataset using in python. The data set contains the various instances and the 4 attributes. The attributes are year, Crime type (act according), People arrested, Crime type. The algorithm is used in following manner:

DBSCAN Technique:

DBSCAN (Density-Based Spatial Clustering of Application
with Noise) is a density based clustering algorithm. The algorithm grows regions with sufficiently high density into clusters and discovers clusters of arbitrary shape in spatial database with noise. It defines a cluster as a maximal set of density-connected points. Clusters are dense regions in the data space, separated by regions of the lower density of points.

Algorithmic steps for DBSCAN clustering

Let X = {x1, x2, x3, ..., xn} be the set of data points. DBSCAN requires two parameters: ε (eps) and the minimum number of points required to form a cluster (minPts).

1) Start with an arbitrary starting point that has not been visited.
2) Extract the neighborhood of this point using ε (All points which are within the ε distance are neighborhood).
3) If there are sufficient neighborhood around this point then clustering process starts and point is marked as visited else this point is labeled as noise (Later this point can become the part of the cluster).
4) If a point is found to be a part of the cluster then its ε neighborhood is also the part of the cluster and the above procedure from step 2 is repeated for all ε neighborhood points. This is repeated until all points in the cluster is determined.

5) A new unvisited point is retrieved and processed, leading to the discovery of a further cluster or noise.

6) This process continues until all points are marked as visited.

Hierarchical Clustering Technique:

A **Hierarchical clustering** method works via grouping data into a tree of clusters. Hierarchical clustering begins by treating every data points as a separate cluster. Then, it repeatedly executes
the subsequent steps:

1. Identify the 2 clusters which can be closest together, and
2. Merge the 2 maximum comparable clusters. We need to continue these steps until all the clusters are merged together.

In Hierarchical Clustering, the aim is to produce a hierarchical series of nested clusters. A diagram called **Dendrogram** (A Dendrogram is a tree-like diagram that statistics the sequences of merges or splits) graphically represents this hierarchy and is an inverted tree that describes the order in which factors are merged (bottom-up view) or cluster are break up (top- down view).

The basic methods to generate hierarchical clustering are:

1. Agglomerative:

Initially consider every data point as an **individual** Cluster and at every step, **merge** the nearest pairs of the cluster. (It is a bottom-up method). At first every data set is considered as individual entity or cluster. At every iteration, the clusters merge with different clusters until one cluster is formed.

Algorithm for Agglomerative Hierarchical Clustering is:

- Calculate the similarity of one cluster with all the other clusters (calculate proximity matrix)
- Consider every data point as a individual cluster
- Merge the clusters which are highly similar or close to each other.
- Recalculate the proximity matrix for each cluster
- Repeat Step 3 and 4 until only a single cluster remains

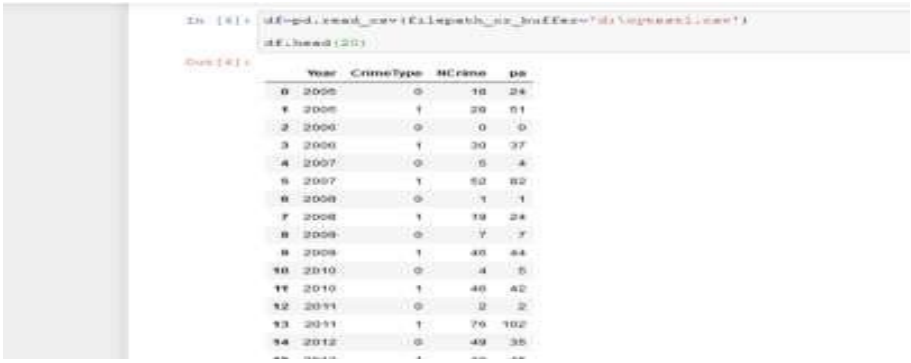
3. Technology & Dataset

DBSCAN and Hierarchical Clustering is one of the popular clustering algorithms. The goal of this algorithm is to find groups (clusters) in the given data. We implement DBSCAN and Hierarchical algorithm using Python packages: pandas, NumPy, scikit-learn, Seaborn and Matplotlib.

1. **Pandas**: Pandas is used to working with “relational” or “labeled” data both easy and intuitive. It aims to be the fundamental high-level building block for doing practical, real world data analysis in Python.
2. **Numpy:** Numpy is used for N-dimensional array object and sophisticated (broadcasting) functions.
3. **SKlearn:** Is a free software machine learning library for the python programming language.
4. **Seaborn:** is use for data visualization and a high-level interface for drawing attractive and informative statistical graphics.
5. **Matplotlib:** Matplotlib is used for 2D plotting library which produces publication quality figures in a variety of hardcopy formats and interactive environments across platforms.
6. **Pylab:** Pylab is a procedural interface to the Matplotlib object oriented plotting library.
7. **Collection:** Collection in python are containers that are used to store collection of datafor example list,dict,set,tuple etc.
8. **Scipy:** Scipy contains varieties of sub packages which help to solve the most common issue related to Scientific Computation.
9. **Dendrogram:** Dendrogram is a diagram representing a tree using hierarchical clustering.

10. **Agglomerative:** Agglomerative is a hierarchical clustering method that applies the "bottom-up" approach to group a element in a dataset.

The data was collected from the National crime record bureau (2005 to 2019) data set converted into iris dataset using in python. The data set contains the various instances and the 4 attributes. The attributes are year, Crime type (act according IT ACT-0,IPC ACT-1), No.of crime (0,1) Act wise,People arrested(0,1) Act wise. Describe in image format below:



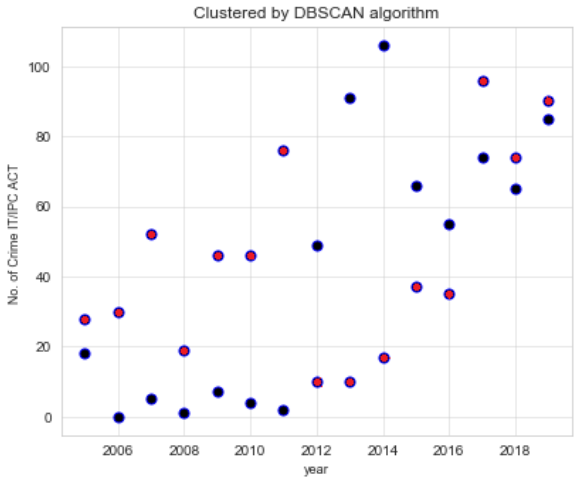
	Year	CrimeType	NCrime	pa
0	2005	0	18	24
1	2005	1	29	51
2	2006	0	0	0
3	2006	1	30	37
4	2007	0	5	4
5	2007	1	52	82
6	2008	0	1	1
7	2008	1	19	24
8	2009	0	7	7
9	2009	1	46	44
10	2010	0	4	5
11	2010	1	46	42
12	2011	0	2	2
13	2011	1	79	102
14	2012	0	49	35

4. Result

In every model, the accuracy and the cost analysis plays an important role in the acceptance of that model
for the application.

(i) DBSCAN Algorithm RESULT :

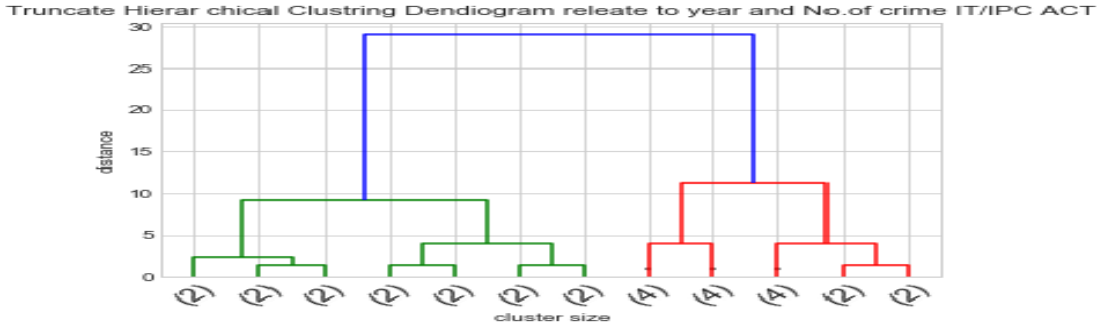
In this result year wise no. of crime according IT/IPC Act Show X and Y axis and year wise no. of person arrested according IT/IPC Act Show X and Y axis.



No. of Person Arrested Based on IPC ACT Crime No. of Person Arrested Based on IT ACT Crime

(ii) Hierarchical Alogorithm RESULT:

In this result year wise no. of crime according IT/IPC Act based cluster ant distance Show X and Y axis using dendrogram and year wise no. of person arrested according IT/IPC Act based cluster ant distance Show Show X and Y axis



using dendrogram.

Conclusion

This paper presents a DBSCAN and Hierarchical clustering using python. It is taking cyber crime dataset (Chhattisgarh) from 2005 to 2019) and classification of peoples arrested in that year by cluster. it also helpful for other prescribe dataset.

REFERENCES

[1] Kulwant Malik , "Emergence of Cyber Crime in India" **,** International Referred ResearchJournal,July,2011,ISSN-0975-3486, RNI: RAJBIL2009/30097, VOL-II*ISSUE 22

[2] Hemraj Saini, Yerra Shankar Rao, T.C.Panda, "Cyber-Crimes and their Impacts: A Review", International Journal of Engineering Research and Applications(IJERA) ISSN: 2248-9622 www.ijera.com Vol. 2, Issue2,Mar-Apr 2012, pp.202-209

[3] Varun Kumar, Nisha Rathee,"Knowledge Discovery from Database using an Integration of clustering and Classification", IJACSA, vol 2 No.3,PP. 29-33,March2011.

[4] Sanjay Chakraborty, Prof. N.K.Nagwani "Analysis and Study of Incremental DBSCAN Clustering Algorithm"International journal of Enterprise computing and business system.

[5] Adriano Morira, Marible Y.Santos, Sofia Carneiro," Density based clustering algorithms DBSCAN and SNN", University of Minho – Portugal, Version 1.0, 25.07.2005.

[6] CHEN Ning , CHEN An, ZHOU Long-xiang,"An Incremental Grid Density-Based Clustering Algorithm", Journal of Software, Vol.13, No.1,2002.

[7] Eshref Januzaj Hans-Peter Kriegel Martin Pfeifle," Towards Effective and Efficient Distributed Clustering",Workshop on Clustering Large Data Sets (ICDM2003), Melbourne, FL, 2003

[8] Chris ding and Xiaofeng He(2002), Cluster Merging And Splitting In Hierarchical Clustering Algorithms.

[9] MarjanKuchaki Rafsanjani,Zahra AsghariVarzaneh,Nasibeh Emami Chukanlo (2012), A survey of hierarchical clustering algorithms, TheJournal of Mathematics and Computer Science, 5,.3,pp.229-240

[10] G.Karypis, E.H.Han and V.Kumar(1999), CHAMELEON: Hierarchical clustering using dynamic modeling,IEEE Computer, 32, pp. 68-75.

[11] J.A.S. Almeida, L.M.S. Barbosa, A.A.C.C. Pais and S.J. Formosinho(2007), Improving Hierarchical Cluster Analysis: A new method with outlier detection and automatic clustering, Chemometrics and IntelligentLaboratory Systems,87,pp.208-217.

Webography

[1] www.datacamp.com
[2] www.kaggle.com
[3] www.r-boggers.com
[4] www.mubaris.com
[5] https://uc-r.github.io/hc_clustering
[6] www.ncrb.gov.in
[7] https://scikit-learn.org