K-Means Clustering Algorithm: Analysis Cyber Crime Data

Dr. Neelam Sahu[1] & Mr. Sagar Darokar[2]

[1]ASSOCIATE PROFESSOR, DEPT OF INFORMATION TECHNOLOGY, DR. C.V. RAMAN UNIVERSITY KOTA, BILASPUR (C.G.), INDIA.

[2]RESEARCH SCHOLOR, DEPT OF INFORMATION TECHNOLOGY, DR. C.V. RAMAN UNIVERSITY KOTA, BILASPUR (C.G.), INDIA.

Received: September 20, 2018 Accepted: October 30, 2018

ABSTRACT: *Cyber Crime is technology based crime committed by technocrats. This paper deals with Variants of cyber crime held in Chhattisgarh between 2005to 2013. Under this, the Age wise Clustering of arrested people has been displayed on basis of cybercrime in Chhattisgarh. Data mining k-Means algorithm is used for clustering. In k-means clustering, we are given a set of n data points in d-dimensional space Rd and an integer k and the problem is to determine a set of k points in Rd, called centers, so as to minimize the mean squared distance from each data point to its nearest center. Python Software has been used to implement the K-Mean Algorithm in cyber crime dataset.*

Key Words: *Cyber Crime, Types of Cyber Crime, k-Mean 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 2013, the clustering has been made through the k-mean algorithm which is based on cyber crime dataset. We can do k-means algorithm using python.

2. Methodology

Cluster analysis or clustering is the process of partitioning aset of data objects into subsets. Each subset is a cluster, such that objects in a cluster resemble one another, yet dissimilar toobjects in other clusters. In this context, k-Means clustering methods may generate different clustering's on the Cyber Crime dataset. The partitioning is not performed by humans, but by the clustering algorithm K-mean clustering algorithms were used for formation of clusters on cyber crime database. The data was collected from the National crime record bureau (2005 to 2013) data set converted into iris dataset using in python. The data set contains the various instances and the 6 attributes. The attributes are year, Crime type (act according), People arrested, age, Crime type. The algorithm is used in following manner:

K-Means: Technique

The k-means algorithm defines the centroid of a cluster as the mean value of the points within the cluster. First, it randomly selects k of the objects in D, each of which initially represents a cluster mean or center. For each of the remaining objects, an object is assigned to the cluster to which it is the most similar, based on the Euclidean distance between the object and the cluster mean. The k-means algorithm then iteratively improves the within-cluster variation. For each cluster, it computes the new mean using the objects assigned to the cluster in the previous iteration clusters formed in the current round are the same as those formed in the previous round. The k-means procedure along with algorithm is given below.

```
Algorithm K-means:
Input = K:
The number of clusters= D: A dataset containing n objects
Output = A set of K clusters
Method:
    (1) Arbitrarily choose K-objects from D as the initial cluster centers
    (2) Repeat
    (3) Re-assign each object to the cluster to which the object is the most similar, based on the mean
        value of the objects in the cluster
```

(4) Update the cluster means, i.e. calculate the mean value of the objects for each clusters
(5) Until no changer

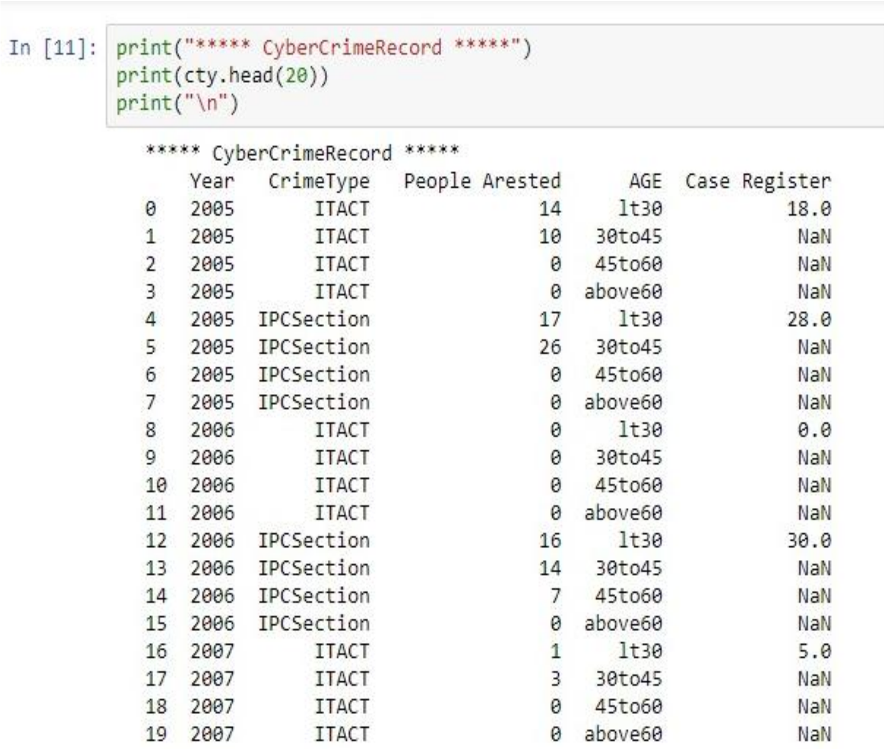
The time complexity of the k-means algorithm is O(nkt),where n is the total number of objects, k is the number of clusters, and t is the number of iterations. Normally, k ≪ n and t ≪ n. Therefore, the method is relatively scalable and efficient in processing large data sets.

3. Technology& Dataset

K-Means Clustering is one of the popular clustering algorithms. The goal of this algorithm is to find groups (clusters) in the given data.We implement K-Means 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. Scikit-learn:Scikit-learn provideK-Mean algorithms via a consistent interface in Python.
4. Seaborn: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.

The data wascollected from the National crime record bureau (2005 to 2013) data set converted into iris dataset using in python. The data set contains the variousinstances and the 6 attributes. The attributes are year, Crime type (act according),People arrested, age, Crime type. Describe in image format below:

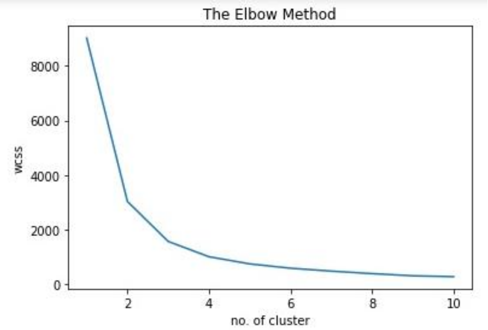


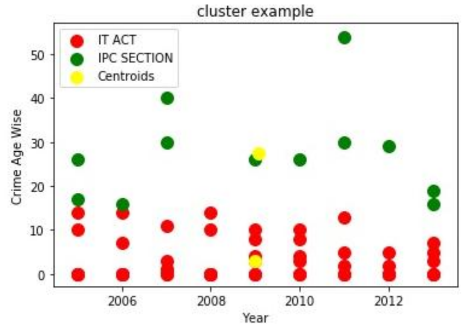
```
In [11]: print("***** CyberCrimeRecord *****")
         print(cty.head(20))
         print("\n")
```

```
***** CyberCrimeRecord *****
    Year   CrimeType  People Arested     AGE  Case Register
0   2005       ITACT              14    lt30           18.0
1   2005       ITACT              10  30to45            NaN
2   2005       ITACT               0  45to60            NaN
3   2005       ITACT               0 above60            NaN
4   2005  IPCSection              17    lt30           28.0
5   2005  IPCSection              26  30to45            NaN
6   2005  IPCSection               0  45to60            NaN
7   2005  IPCSection               0 above60            NaN
8   2006       ITACT               0    lt30            0.0
9   2006       ITACT               0  30to45            NaN
10  2006       ITACT               0  45to60            NaN
11  2006       ITACT               0 above60            NaN
12  2006  IPCSection              16    lt30           30.0
13  2006  IPCSection              14  30to45            NaN
14  2006  IPCSection               7  45to60            NaN
15  2006  IPCSection               0 above60            NaN
16  2007       ITACT               1    lt30            5.0
17  2007       ITACT               3  30to45            NaN
18  2007       ITACT               0  45to60            NaN
19  2007       ITACT               0 above60            NaN
```

4. Result

In every model, the accuracy and the cost analysis plays animportant role in the acceptance of that model for the application. The result of the cyber crime data set is being displayed as a cluster form. The result is displayed o the basis of axis x-axis represent year and y-axis represents age of arrest people and the cluster are represent in form of dot yellow and green ,centroids cluster represent as yellow color.





Conclusion

This paper presents a k-Mean clustering using python. It is taking cyber crime dataset (Chhattisgarh) from (2005 to 2013) and classification of peoples arrested in that year by cluster. it also helpful for other prescribe dataset.

REFERENCES

1. KulwantMalik , “Emergence of Cyber Crime in India” **,** InternationalReferred Research Journal,July,2011,ISSN-0975-3486, RNI: RAJBIL2009/30097, VOL-II *ISSUE 22
2. Hemraj Saini, Yerra Shankar Rao, T.C.Panda, “Cyber-Crimes and theirImpacts: A Review”, International Journal of Engineering Research andApplications(IJERA) ISSN: 2248-9622 www.ijera.com Vol. 2, Issue2,Mar-Apr 2012, pp.202-209
3. Varun Kumar, NishaRathee,”Knowledge Discoveryfrom Database using an Integration of clustering andClassification”, IJACSA, vol 2 No.3,PP. 29-33,March2011.
4. Weka – Data Mining Machine Learning Software, http://www.cs.waikato.ac.nz/ml/.
5. Cheung Y.M. (2003). k-means: A New Generalisedk-means ClusteringAlgorithm. N-H Elsevier Pattern Recognition Letters 24, Vol 24(15), 2883–2893
6. Kanungo T., Mount D.M., Netanyahu N.S., Piatko C.D., Silverman R.and Wu A.Y. (2002). An Efficient k-means Clustering Algorithm: Analysisand Implementation. IEEE Transactions on Pattern Analysis and MachineIntelligence, Vol. 24, 881-892.
7. M. Inaba, N. Katoh, and H. Imai, “Applications of WeightedVoronoi Diagrams and Randomization to Variance-Basedk-clustering,” Proc. 10th Ann. ACM Symp. Computational Geometry,pp. 332-339, June 1994.
8. Tapas Kanungo, Senior Member, IEEE, David M. Mount, Member, IEEE,An Efficient k-Means Clustering Algorithm:Analysis and ImplementationIEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 24, NO. 7, JULY 2002
9. D. Pollard, “A Centeral Limit Theorem for k-means Clustering,”Annals of Probability, vol. 10, pp. 919-926, 1982.
10. Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. From data mining toknowledge discovery: An overview. In Advances in Knowledge Discoveryand Data Mining, pp. 1 --34. AAAI Press, Menlo Park, CA, 1996.
11. Ester M., Kriegel H.-P., Sander J., Xu**X,** “ADensity-Based Algorithm for Discovering Clusters in
Large Spatial Databaseswith Noise”, **In:** Proc. 2nd Int.Conf. on Knowledge Discovery and Data Mining,
12. Han Jiawei, Kamber M. Fan Ming, Meng Xiao-Fenget al. Translated. “Data Mining: Conceots and
Techniques”, Beijing: China Mechine Press, 200l(inChinese)

Webography

1. www.datacamp.com
2. www.kaggle.com
3. www.r-boggers.com
4. www.mubaris.com
5. www.ncrb.gov.in