IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

STUDY OF UNSUPERVISED LEARNING TECHNIQUES : K-MEANS AND HIERARCHICAL CLUSTERING ALGORITHM

[1]Pratiksha Usatkar

[2]student,

[1] Bharti Vidyapeeth' Institute of Management and Information Technology, Navi Mumbai,India

Abstract

Cluster analysis separates information into important, useful groups (cluster). Clustering algorithms measure similarity or dissimilarity between data objects. Clustering is used to find meaningful information/patterns from a data set. Cluster analysis is an unsupervised learning algorithm.
Unsupervised learning models the structure or dispersion within the information to memorize more almost the information. There are different types of clustering viz. Hierarchical, Partition, Exclusive, Overlapping, Fuzzy, Complete, Partial, Well-separated, Prototype-based, Graph-based, Contiguity- based, and Density-based Clustering. This paper compares with k-Means Clustering and Hierarchical Clustering Techniques. Clustering method and their technique and process.

Keywords: k-means, Hierarchical, unsupervised learning, Clustering algorithm

I. Introduction

Given a set of data points, ready to utilize a clustering algorithm to classify each information point into a particular group. Data points must have the same properties that are within the comparable cluster and types. i.e. Intercluster distance in the same cluster should be less, whereas data points totally different clusters ought to have exceedingly disconnected properties ie. Intracluster distance between different clusters should be maximum.

Clustering algorithms are an effective strategy for machine learning on unsupervised information. Cluster analysis can be a powerful tool for any organization that needs to identify discrete groups of customers, sales transactions, or other types of behaviors and things For case, insurance suppliers utilize cluster analysis to identify false claims, and banks utilize it for credit scoring These algorithms are mostly used in the identification of fake news, spam filter, marketing sales, classify network traffic, and identifying fraudulent or criminal activity.

in machine learning, the most common algorithms are k-means and Hierarchical. These two algorithms are inconceivably effective when connected to different machine learning issues.

II.Literature Review

[1]K-Means Clustering - Centroid Based Progressive Clustering - Divisive and Agglomerative K-means clustering is one of the well-known sorts of partitioning-based Clustering. Partitioning algorithms are Clustering strategies that subdivide the information set in a set of k bunches, where k is the pre-determined number of clusters for generalization. In k-means clusters are represented by the center or mean of information focuses having a place in the cluster.

[3]There are six sorts of clustering procedures- k-Means Clustering, Hierarchical-Based Clustering, and EM Algorithm, Optics. Performance investigation of k-means with distinctive initialization techniques for high dimensional data

[15] In this paper Khaled Alsabti et al, the creators present the novel algorithm for performing k-means clustering. The most point of the creator here was to think approximately the computational points of view of the k-means procedure. Here the datasets are made misleadingly to induce the scaling properties of the calculation utilized by them. They communicated that their algorithm arranged will inside and out prevalent execution than the facilitate k-means calculation in most cases of their exploratory comes approximately. The plan which they proposed is said to be making strides in the computational speed of the arrange k-means calculation by an organized to two orders of greatness inside the add up to a few remove calculations and the in general time of computation

[6]Trupti M.Kodinariya et al., elaborated six different approaches for the selection of K value for the K-Mean clustering algorithm in a dataset. He concluded that clusters are in a viewing eye and analyzed the situation when clusters, though not definitely typical, are in data.

[8] Evolving restrictions in K-means algorithm in data mining and their removal" K-means is prevalent since it is conceptually basic and is computationally quick and memory-efficient, but distinctive types of restrictions in k-means the algorithm that creates extraction difficult

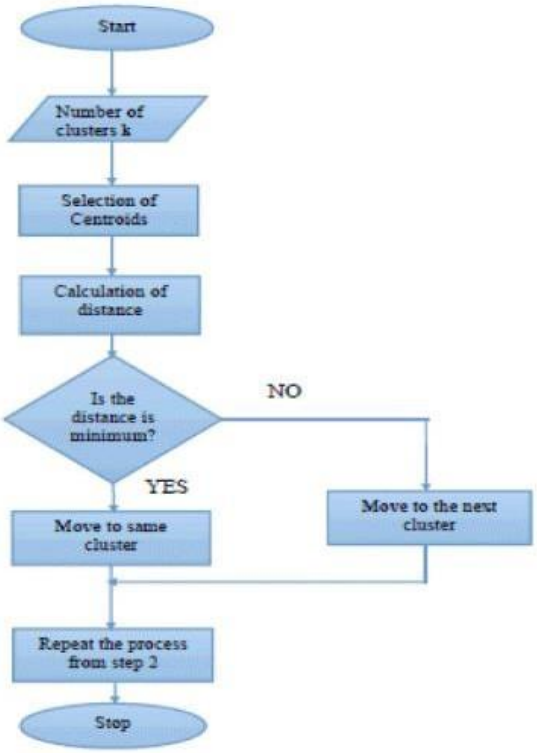
[9]A Hierarchical Latent Variable Model for Data Visualization Researchers introduce a hierarchical visualization algorithm which allows the complete the information set to be visualized at the top level, with clusters and sub-clusters of data points visualized at deeper Levels.

[12] Modified Form of the K-Means Algorithm with a Distance-Based on Cluster Symmetry. In this paper, Researchers propose a modified version of the Kmeans algorithm to cluster data

III.algorithm

I. k-means clustering algorithm

1. K-means clustering could be a type of unsupervised learning which is utilized once you have unlabeled data
2. the objective of this algorithm is to discover groups inside the data with the number of bunches talked to by the variable k 3.
3. the algorithm works iteratively to assign each information point to one of the k bunches based on the the highlights that are provided
4. data points are clustered based on include similarity

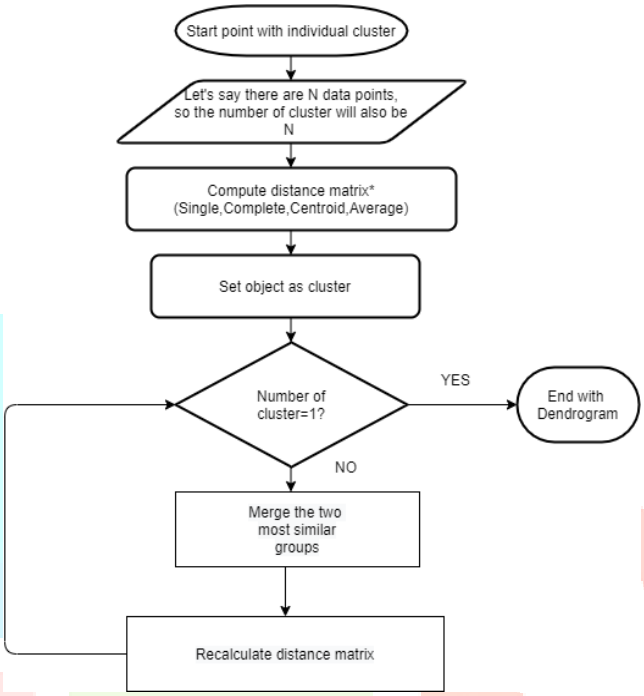


II.Hierarchical clustering algorithm

Hierarchical clustering is an unsupervised machine learning algorithm ,it is utilized bunch the unlabeled datasets for dataset for clusters called Hierarchical cluster examinations in this algorithm in this algorithm ,we make in this calculation,we develop the progression of clusters within the shape tree-shaped structure is called the dandogram

1.clustering is performed based upon dissimilarities between clusters
2.produces an arrangement or tree of clustering
does not require the number of the cluster as input

4.partitions can be visualized employing a tree structure 5possible to see allotments at the diverse level of granularities utilizing distinctive k



I.Agglomerative

1. Start point with the individual cluster In this algorithm If there is an n data point, then the number of clusters would-be n
2. At every step, interface the closest coordinate of clusters till because it was one cluster is clear out
3. Compute the distance matrix(Single, complete, centroid, and average Linkage)
4. all the clusters are merged toward one huge cluster, create the dendrogram to partition the clusters as per the issue

II.Divisive

1.Start with one cluster, all-inclusive cluster

2.At each step, part a cluster until each cluster contains a k cluster

3. Compute the distance matrix(Single, complete, centroid, and average Linkage)

4. all the clusters are merged toward one huge cluster, create the dendrogram to partition the clusters

IV.Customer segmentation

In an organization customer identification and their behaviors plays an important role. We can group the customers who can buy similar products based on their identification and behavior.
An unsupervised clustering in machine learning (ML), will help to identify customers which includes two clustering techniques K means clustering and hierarchical clustering which we are going to implement in Python.

Problem Statement

Suppose there is a mall, where they have recorded 200 details of customers like age, gender, annual income, and spending score through campaigns, spending score is ready based on the investing habits of the buys they have made from the shopping center. Presently, the mall is bringing the new and luxurious products to the mall and wants to reach the customers, we can't go to each customer and ask how is the product, instead, we can separate the customers into groups who can buy the products. This problem can be done using clustering.To represent this we can use Two Dimensional Euclidean space, one is the X-axis which is represented by annual income, and Y-Axis which is represented by spending Score. By representing each customer on the plane, we can use the clustering method to find the customers who buy luxurious products.

I.Python implementation of kmeans algorithm

1. The first step is data processing, we have to import the data first, after that, we have to import the dataset. Then we have to extract the independent variable
2. In the second step, we will use the elbow method and find the optimal number cluster
3. In the third step, We have to train the K-means algorithm from the training data set
4. In the fourth step visualize the cluster

II.Implementation of hierarchical clustering

1. The first step is data processing, we have to import the data first, after that, we have to import the dataset.

   Then we have to extract the matrix of feature
2. In the second step, we will use the Dendrogram and find the optimal number cluster
3. In the third step, we have to train the hierarchical algorithm from the training data set
4. In the fourth step visualize the cluster

V.Python implementation customer segmentation steps:

1. Import the basic libraries to examined the CSV record and visualize the information
2. Read the dataset that's in a CSV record. Characterize the dataset the demonstrate
3. To execute the K-Means clustering, we have to find the perfect number of clusters in which clients will be put. the ideal number of clusters for K-Means, the Elbow procedure is utilized based on Within- Cluster-Sum-of-Squares (WCSS). It'll be plotted as given below: As we will see inside the over the figure, the over plot is visualized and we ought to recognize the region of the elbow on the X-axis. Inside the over plot, the elbow shows up to be on point5 of the X-axis. So, the perfect number of clusters will be 5 for the K-Means calculation.
4. After finding the perfect number of clusters, fit the K-Means clustering appears to the dataset characterized inside the minute step and after that anticipates clusters for each of the information components. It infers it'll expect which of the 5 clusters
5. When the calculation predicts a cluster for each of the data things, we through the plot. For superior representation, we have to be delivered each of the clusters interesting colors and titles.
6. The title of clusters is given based on their wage and contribution. For case, when alluding to a client with low pay and high contributing, we have utilized cyan color. This bunch appears as ‘Careless Customer’ since in spite of having a low compensation, they spend more. To offer an excessive thing, a person with a tall wage and tall contributing propensities ought to center. This bunch of clients is is spoken to in maroon color within over the chart.
7. Presently the same assignment will be executed utilizing Various leveled clustering. The perusing of CSV records and making a dataset for calculations will be common as given inside the primary and moment steps. In K-Means, the number of perfect clusters was found utilizing the elbow strategy. K-Means, the number of ideal clusters was found utilizing the elbow method. In various leveled clustering, the dendrograms are utilized for this reason. Plot and visualized a dendrogram for our dataset.In case you're aware of this method, you'll be able to I see inside the over charts. The combination of 5 lines isn't joined on the Y-axis from 100 to 240, for approximately 140 units. the ideal

number of clusters would be 5 for progressive clustering.

8. Presently we prepare the hierarchical clustering calculation and expect the cluster for each data point.

Once the algorithm predicts the cluster for each of the information focuses, it can be visualized now.

the over plot can be visualized as We'll see inside the over graph that the clustering of clients is nearly comparative to what was done by K-Means clustering. As it were the colour combinations have been changed for separating both the chart.

VI.Comparison between k-means and Hierarchical clustering

As we have seen inside the over an area, the comes about of both the clustering are almost comparable to the similar dataset. It may be possible that small. Be that as it may, at the side numerous likenesses, these two techniques have a few contrasts too. The below table shows up the comparison between K- Means and Hierarchical clustering calculations based on our executions.

K-means clustering	Hierarchical clustering
K-means algorithm performance is better	Hierarchical clustering algorithm performance is Less
K-means algorithm increases its execution Duration	Hierarchical algorithm performance is good
K-means clustering K Implies clustering required advancement data of K i.e. no. of clusters one has to partition your data.	In hierarchical clustering one can stop at any number of clusters, one finds reasonable by translating the dendrogram.
K is mandatory	K is not mandatory as an input
K-means algorithm shows quality very low	The hierarchical algorithm shows quality very High
K-means algorithm is fine for a big dataset of Data	The hierarchical algorithm is fine for the small Dataset
The directional approach in k-means Clustering are not any, as it were centroid is considered to create clusters	The directional approach in Hierarchical Clustering are top-down and bottom-up
The category of the k-means clustering is centroid based and division-based	The category of Hierarchical clustering is Hierarchical Agglomerative
elbow method used in the k-means clustering to find the optimal number of cluster	Dendrogram method used in the Hierarchical clustering to find the optimal number of cluster

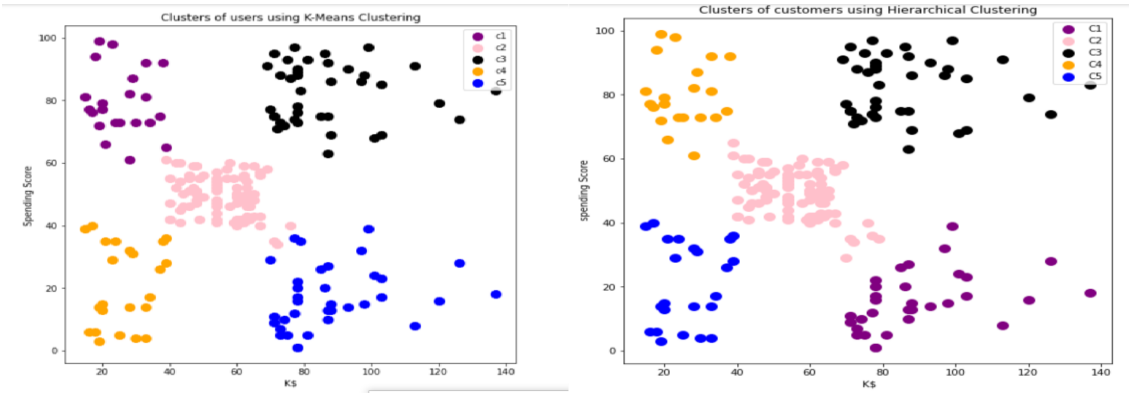
VII.Result:

We have compared the k-means algorithm with the Hierarchical algorithm. The algorithm is implemented by using python.

Algorithm of k-means and Hierarchical algorithm

Algorithm	Time
K-means	9.207941055297852 seconds
Hierarchical	21.182947158813477 seconds

Python implementation of k-means and Hierarchical algorithm



The output image is showing five different colors with different clusters

C1 appears the clients with normal salaries and normal spending so ready to categorize these clients as C2 appears the customer incorporates a high salary but low investing, so able to categorize them as careful.C3 appears the low salary additionally low investing so they can be categorized as sensible.C4 appears the clients with low salary with exceptionally high spending so they can be categorized as careless. C5 shows the clients with high salaries and high investing so they can be categorized as target

Algorithm	Time
K-means	82.34615516662598 seconds
Hierarchical	250.29868912696838 seconds

VIII.Conclusion

K-means clustering is a broadly utilized method for data cluster analysis.The ‘k-means’ inside the K-means refers to averaging of the information; that's, finding the centroid.We also call Hierarchical Clustering a Greedy Algorithm because splits and merges of clusters vary based on Linkage selection.Hierarchical clustering is the most well known and broadly utilized strategy to analyze social network data.When there is a large dataset, the quality of the algorithm is good and performance is also good. Being a Huge dataset, the K-Means algorithm is faster than the rest of the algorithms.k-means algorithm performance is better than Hierarchical clustering.The hierarchical algorithm was received for categorical data, and due to its complexity and a new approach for assigning rank value to each categorical attribute using K- implies can be utilized in which categorical data is,begin with, changed over into numeric by assigning rank.

IX.References

[1].Gurpreet Kaur and Naveen Aggarwal, “Exploiting Hierarchal Structure of XML Data Using Association Rule Analysis”, International Journal of Machine Learning and Computing, Vol. 2, No. 3, June

[3] Verma, M., Srivastava., Chack, N., Diswar, A . K ., Gupta, N.,” A Comparative Consider of Distinctive Clustering algorithm in Data Mining,” International Journal of Engineering Research and Applications (IJERA), Vol. 2, Issue 3, pp. 1379-1384, 2012

[15] Sharma, N., Bajpai, A., “Comparing the diverse clustering calculations of Weka tool”, Worldwide Journal of Developing Technology and Progressed Engineering,2(5),2012.

[2] Khaled Alsabti, Sanjay Ranka and Vineet Singh “A proficient k-means clustering algorithm”, Syracuse College SURFACE, L.C."

[6] Trupti M.Kodinariya and Dr.Prashant R. Makwana 2013, Survey on deciding the number of cluster in K-Means.

[9] Bishop C. M., Michael, E. Tipping, “Hierarchical Latent Variable Model for Data Visualization”, IEEE Trans. Pattern Anal. Mach Intell., 20 (3), 281-293,1998.

[12] Su, M.C, Chou, C.H., A Modified Version of the KMeans Algorithm with a Distance Based on Cluster Symmetry”, IEEE Transactions on Pattern Analysis and Machine, 23 (6), Aug 7, 2002.

[3]"K ., Gupta, N.," A Comparative Consider of Diverse Clustering Algorithms in Data Mining," Universal Journal of Engineering Inquire about and Applications (IJERA), Vol."

[8]Singh, K., Malik D., Sharma, N., "Evolving confinements in k-means algorithm in data mining and their evacuation,"IJCEM Universal Journal of computational Designing

[16] https://www.javatpoint.com

[17]https://analyticsindiamag.com

[18] https://www.geeksforgeeks.org/

[19] https://towardsdatascience.com/