OORJA
ISSN - 0974-7869 (Print)
ISSN - 2395-6771 (Online)

COMPARISON OF CLUSTERING ALGORITHMS: A PRACTICAL APPROACH

Ms Meenal Sharma, Assistant Professor, International School of Informatics & Management, Jaipur

Abstract:

Data mining is a technique used for retrieving information and data from the database. This paper focuses on clustering, a technique which is used to divide the data into interdependent clusters that are of similar type. Clustering is a way of classifying and breaking down the data, and putting the similar type of meaningful data/objects into a single group or cluster. The technique follows the most common approach of finding centre of clusters and input vectors, which help to identify which cluster centre belongs to which input vector. Hierarchical clustering is a technique of clustering the data using predetermined way of clustering, either it is top to bottom or it is bottom up approach. It is further divided into two ways: Agglomerative (Bottom-up approach) and Divisive(Top-down approach). K-Means clustering is a technique which makes various partitions of the given data from n observations of objects into k clusters, where each of the object belongs to a particular cluster of nearest mean(also known as center of clusters).

In this paper road accidents dataset is used that contains the records of road accidents occurred in the India on day to day basis. The dataset holds the records of the accidents happened in each state of India. This dataset is used because the number of accidents happening daily in the India is very big in count and the accidents happen are of different types like fatal accidents, major accidents, non-injury accidents, etc. This paper uses clustering algorithm to differentiate the categories of accidents in the dataset.

Keywords: Data Clustering, Unsupervised learning technique, Hierarchical clustering, K-means, Statistical tool R, Clustering techniques.

Introduction

Clustering is an un-supervised classification process of breaking down of big data. Data clustering means dividing the large or big data into the small parts or clusters, which is a particular cluster contains the data which is of same type or related to each other.

It is an un-supervised way of learning with no target fields, and also considered that it is an approach of managing the data in bottom-up way, It was originated in anthropology by Driver and Kroeber in 1932 and later by Zubin in 1938, who introduced it to psychology and further on in 1939 Robert Tryon and in 1943 It was observed that clustering is popularly used by Cattell Beginning in personality psychology for classification of trail theory.

The main benefit of clustering the data is that, it breaks down the large and huge data sets into small groups which is very helpful in processing and extracting from the data. In the data analysis clustering plays an important role, as it is also defined as a practice of extracting the information and set of patterns from the huge data set.

Clustering performs grouping of the data on different data sets in such a way that it establishes the maximum similarity within the data of a particular cluster and along with that it also minimizes the similarity between different clusters.

Hence, the main goal of clustering technique is that the data or objects in the particular class/set to which they belongs to will be of similar in nature to one another and it shows different qualities when compared to other groups. The data in the clustering is associated and divided among the different clusters, on the basis of the logical relationships.

TYPES OF CLUSTERING

Clustering methods are broadly divided in the following types:-

- Clustering with Partitioning Method
- Clustering with Hierarchical Method
- Clustering with Density-based Method
- Clustering with Grid-based Method
- Clustering with Model-based Method
- Clustering with Constraint based Method

Partitioning Method:

In the provided class of the given data, total quantity of objects stated as 'n' , and 'k' number of partitions of the given data are constructed by the partitioning methods. Here:

- **n'** is defined as a number of objects.
- **k'** is for number of partitions.

And every single cluster is represented by the single divided partition that will be represented as 'k=<n'(k is less than equal to n). It means that it clusters the data into the k similar classes and satisfies the given requirements stated below:

- Every single divided class of cluster contains at least a single object.
- Each and every object must belongs to only a single class of cluster.

Partitioning method have several other techniques namely k-means, k- medoid and CLARANS, in which **k-means** is one of the most popular technique, found by MacQueen in 1967. Each cluster represented by the center or the means of the data point belonging to the cluster. It is the sensitive method to the out.

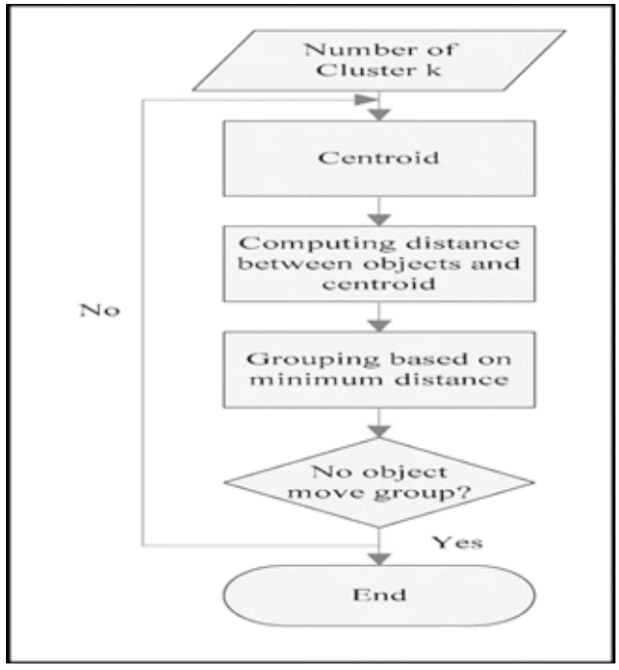

Figure 1: Working of K-means Algorithm

In the above figure(1), the flowchart elaborates the working of k-means algorithm.

Algorithm of k-means clustering:

Required Input:

Wanted number of resultant set of clusters, k, and a database that contains n number of data objects D={d1,

d2,…dn}.

Expected Output:

A set of clusters(k) Steps to be performed:

1) Firstly, there is a need to select k number of data objects from the given dataset D randomly as initially centered clusters.

2) Repeat the given procedure;

3) Now, the distance between the data objects would be calculated as dist(p <= q<=n) and all centered clusters k as c(p<=r<=k) and the data object 'dist' would be assigned to the nearest cluster.

4) Recalculate the center for each cluster.

5) until and unless there is no change in the center of clusters.

Hierarchical Clustering Method:

It is an alternative approach to the above defined method (Partitioning method) that is used to identify the groups in the dataset. In this technique, it is not required to predefine the quantity of clusters to be generated. The result of this technique is represented in a tree-based structure which is also known

ISSN - 0974-7869 (Print)
ISSN - 2395-6771 (Online)

as Dendrogram, in figure(2).

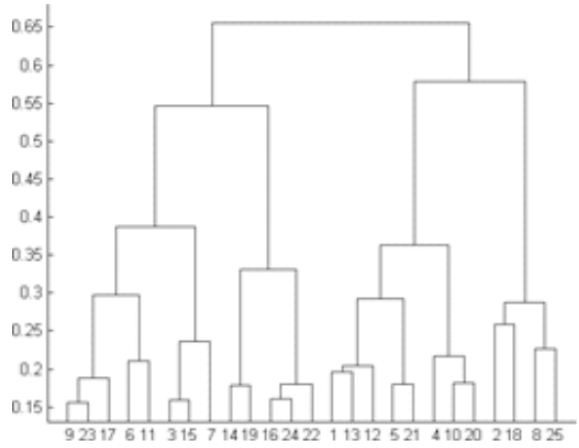

Figure 2: Dendrogram

Hierarchical Clustering further includes two ways to partition the data:

Agglomerative and **Divisive.**

Divisive method is also called as top-down clustering method. In this all observations are assigned into a one big cluster and then in each iterative step, separation of the points from the cluster is done, separation is done on the basis that, those clusters are excluded from the group which are not similar. Every single data point which is separated while the process, is considered as an individual cluster. And there will have n number of clusters which contains similar type of data. Here in divisive clustering method, the one big cluster is divided into 'n' small clusters.

Agglomerative method is also called as bottom-up clustering method. In this assign each observation into its own single cluster, then finds or processes the similarities among the clusters and combines the most similar ones, and repeat the process until there is a one cluster left.

The visualization of hierarchical clustering can be done by using a Dendrogram. A Dendrogram is a tree like structured figure, that is used to record the sequence of merges and splits of the data clusters.

It is very important to calculate the similarity between the clusters because, it helps in merging or splitting the clusters. Various approaches are provided to do that:

- MIN
- MAX
- Group Average
- Distance between Centroids
- Ward's Method

The basic algorithm is very simple:

1. In the beginning, consider each single point as a cluster in its own.

2. Repeat the following process, till there is only one cluster remains:

(a) Find out the most relatable and the closest pair of clusters.

(b) Then merge them, if they are the most closest to each other.

3. Return the merged tree of the formed clusters.

Density-based Method:

This method is a notion of density based method. The main point in this method is that, the cluster is expanding continuously as long as the neighbourhood's density will exceed to some threshold. In other words, this means that in the given cluster each data point and the radius of the cluster will contains minimum number of points.

Grid-based Model:

In this technique, a grid is formed by the objects together. The space of the objects has to be quantized into cells with finite numbers that will form a grid-based structure. It has fast processing speed and this is the main advantage of this model. It only rely on the cells' number in each proportion with the estimated space.

Model-based Method:

This is the technique in which, a model is conjectured to discover out the best fit out of the data for each cluster. This approach uses cluster density function to locate the clusters. In this technique, the data distribution reflects the spatial configuration.

Here the number of clusters are automatically determined, totally based on the statistical standards. It is one of the yield of robust clustering methods.

Constraint-based Method:

It is the technique in which, the congregation is carried out on the basis of user's incorporation and application oriented based restraints. A restraint can be defined as the expectation of the user the other properties of the desired results of the clustering. A constraint act as an interactive way that provides a measure to communicate with the process of clustering. Constraints might be user-specified or based on the requirements of the application.

Comparison on Hierarchical and K- Means Clustering

This paper is focusing on the comparison between the two clustering algorithms: one is Hierarchical clustering and the other one is K-means clustering.

- K-means clustering can handle the big data in a well manner as compared to the hierarchical clustering.
- The data is handled well because of the time complexity which is linear, as stated: O(n), in case of hierarchical clustering it is quadratic as stated: $O(n^2)$.
- The results generated using k-means clustering might varies as because of random clusters are selected at different point of time. While the results of hierarchical clustering are always same and reproducible.
- In case of the shape of clusters are hypo-spherical, k-means clustering is more suitable in such cases as compared to hierarchical clustering.

- In k-means clustering, the prior knowledge of k(no. of similar sets) in which we want to divide our data is required. But in case of hierarchical clustering one can generate a dendrogram and stop the process at whatever number of clusters are found.

Table 1: Comparison between K-means Clustering and Hierarchical Clustering

Basis	K-means Clustering	HierarchicalClustering
Definition	K-means clusteringbreaks down the datasets on the basis of predefined clusters which isgenerally named as 'k' over 'n' number of data objects. It breaks down the data into individuals group of clusters.	The results of hierarchical clustering issuch that the clusters areformed in the hierarchy manner, not like in k- means where number of clusters are predefined.
Clustering Criteria	It is well suited for generating global clusters.	It uses as the clustering criteria to break down the data into different clusters 'distance matrix.'
Performance	The performance is better than hierarchical clustering..	It is more suitable when the datasets are small in size.
Results	The results of k-means clustering varies at different point of time, because of random selection of clusters k.	The results are always same and more accurate as the clusters are predefined, not varies time to time.
Execution Time	Time is more required to execute the results.	Less time is required to execute the results as compared to k-means clustering.
Quality	K-means algorithm generates less quality results.	It generates high quality results.
Time Complexity	Linear time complexity: O(n).	Here time complexity is quadratic: O(n2).

About Dataset

This paper performs analysis on the dataset named "**Road Accidents** " which holds the records of road accidents happened in the country(India) on day to day basis in the year 2021-22.(The dataset holds the records of the accidents state- wise/union territories-wise and also holds the categories of the accidents, for example we can say that: major accidents, fatal accidents minor- injured accidents etc.

The main purpose of using this dataset is that the dataset holds the record of the accidents happens in the country. This dataset is used in this paper because, the number of accidents happens daily in the country is very big in count and the accidents happens are of different types like two-wheeler, four wheeler, bus, etc also keeps the record of the categorical data on the basis of type of accidents whether it is minor or major.

This paper compares the results of the above mentioned two algorithms of clustering, i.e. k-means clustering and hierarchical clustering algorithm. These two algorithms are applied on the same dataset using the statistical tool R and compare the outcome on the basis that which one's result is better and more accurate.

The attributes(types of accidents) that dataset holds are:

1. States/UT
2. Fatal Accidents

3. Grievous Injury Accidents

4. Minor Injury Accidents

5. Non-Injury Accidents

6. Total Accidents

Experiment

For comparing the results two algorithms are applied on the dataset which holds the records about accidents happened in our country using data mining tool **R version 3.6.3**, operating system: Windows 10. The algorithm is applied on R console and plot the respective plots to see the results of clusters formed.

For k-means Clustering:

We used k=4 here as per our dataset.

The resulted graph plot of clusters:

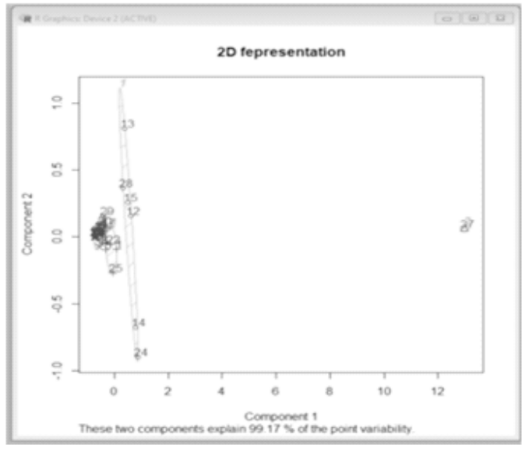

Figure 3: Cluster Plot formed using K-means Algorithm

It is observed in the above figure (3), here k-means clustering algorithm is applied on the respective dataset. Firstly, import the dataset, then scale the values of the dataset by -1 to make it compatible and ready to perform clustering. After this, **kmeans()** function is used, and set value of k equal to 4. Finally the resultant clusters will be plotted on a graph using **clustplot()** function .

OORJA
ISSN - 0974-7869 (Print)
ISSN - 2395-6771 (Online)

For Hierarchical Clustering:

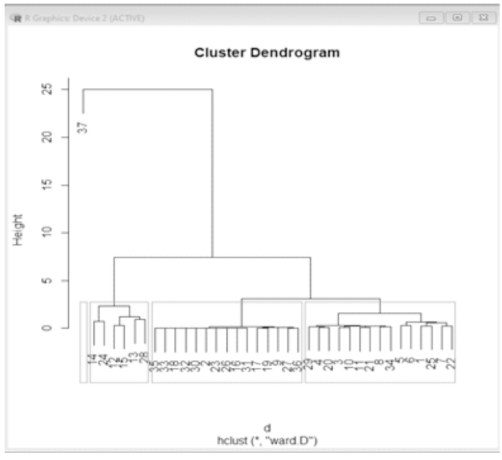

Figure 4: Cluster Plot formed using Hierarchical Clustering

Figure(4) states the hierarchical clustering procedure, here first we load the dataset then scale the dataset by -1 to make it compatible for clustering. Then used distance matrix and Euclidean method, then perform clustering using **hclust()** function using ward method then finally plot the resulting clusters on the graph by **using plot()** function.

Analysis After Experiment

After performing the experiment the results states that in terms of accuracy Hierarchical clustering is more efficient. In other words the clusters formed by using hierarchical clustering are more accurate as compared to k-means clustering.

And the time taken by both the algorithms in generating results, Hierarchical clustering takes less time as compared to k-means clustering. After the experiment, it is proved that Hierarchical clustering algorithm is more efficient as compared to K-means clustering algorithm.

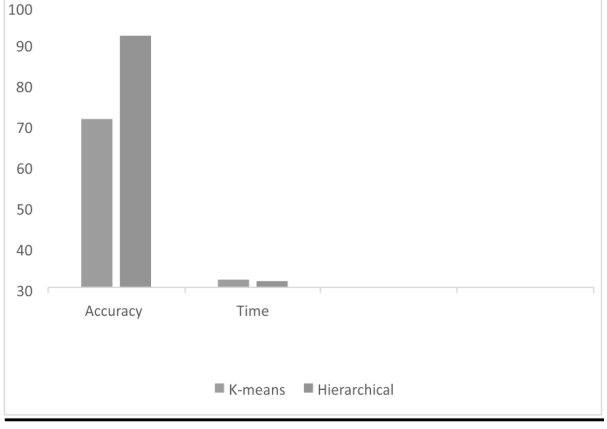

Figure 5: Results after Experiment

Conclusion

After applying both the algorithms on the respected dataset, plot the results on the graph to measure the compatibility of both the algorithm. After applying k-means algorithm, it is noticed that the algorithm is a bit complex to understand , and the algorithm is most suitable for the larger datasets. Here, after plotting on the graph, it is observed that the resultant clusters that are formed are not clearly visible on the graph (figure 5), because of the similarity among the grouped clusters, as because the dataset is small here.

When hierarchical clustering algorithm applied, it is observed that the algorithm is more simple and easy to understand as compared to k-means algorithm. After plotting the graph, it is observed that the resultant graph tree is very easy and simple to understand as compared to the k-means one. Hierarchical algorithm is more suitable for small datasets.

So for the dataset which is used in this paper, hierarchical algorithm is the most suitable to break-down the data as compared to k-means clustering algorithm in terms of simplicity, time complexity, understandability and accuracy.

References:

- M. Ambigavathi and D. Sridharan, “Analysis of Clustering Algorithms in Machine Learning for Healthcare Data”, Advances in Computing and Data Sciences (pp.117-128), Springer, July 2020
- R. C. Sonawane and H. D. Patil, "Clustering Techniques and Research Challenages in Machine Learning," 2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2020, pp. 290-293, doi: 10.1109/ICCMC48092.2020.ICCMC-00054.
- J. Oyelade et al., "Data Clustering: Algorithms and Its Applications," 2019 19th International Conference on Computational Science and Its Applications (ICCSA), St. Petersburg, Russia, 2019, pp. 71-81, doi: 10.1109/ICCSA.2019.000-1.
- H. Xu, "Research on clustering algorithms in data mining," 2022 3rd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE), Xi'an, China, 2022, pp. 652-655, doi: 10.1109/ICBAIE56435.2022.9985831.
- S. Kumar R, A. Arulanandham, S. Arumugam, G. Dinesh, R. Thirukkumaran and R. Subashmoorthy, "Analysis of classification and clustering techniques for ambient AQI using machine learning algorithms," 2022 4th International Conference on Smart Systems and Inventive Technology (ICSSIT), Tirunelveli, India, 2022, pp. 902-908, doi: 10.1109/ICSSIT53264.2022.9716359.
- S.R.Pande, Ms S.S.Sambare and V.M.Thakre, “Data Clustering Using Data Mining Techniques”, International Journal of Advance Research in Computer and Communication Engineering, Vol 1, Issue 8, October 2012.
- Joshi, A, Kaur, R, “A Review: Comparative study of various clustering techniques in data mining”, International Journal of Advance research in Computer Science and Software Engineering, March 2013.
- Manish Verma, Mauly Srivastava, Neha Chack, Atul Kumar Diswar and Nidhi Gupta, “A

Comparative Study of Various Clustering Algorithms in Data Mining", International Journal of Engineering Research and Applications(IJERA), Vol.2, Issue 3, 2012.

- Siddiqui, F., Isa A.M ,"Enhanced Moving K-Means (EMKM) Algorithm for Image Segmentation ", IEEE Trans. Consumer Electronics, 12 (4), 2014.
- Singh, N., & Singh, D. (2012). Performance evaluation of k- means and heirarichal clustering in terms of accuracy and running time. IJCSIT) International Journal of Computer Science and Information Technologies, 3(3), 4119-4121.
- Pamulaparty, L., Rao, C. G., & Rao, M. S. (2016). Cluster analysis of medical research data using R. Global Journal of Computer Science and Technology.
- Narendra, S., Aman, B., & Ratnesh, L. (2012). Comparison the various clustering algorithms of weka. International Journal of Emerging Technology and Advanced Engineering, 2(5), 80.
- Astha Joshi and Rajneet Kaur, "A Review: Comparative Study of Various Clustering Techniques in Data Mining", June 2013