International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 07 | July 2020 www.irjet.net p-ISSN: 2395-0072

Data Clustering: K-Means against Hierarchical Clustering, an Alternate Approach

Nimisha Bhide

[1]*Nimisha Bhide B.E. Computer Engineering Mumbai, Maharashtra, India.*

Abstract - *In today's world, analyzing data is the best way to extract information from any kind of data. There are many ways to analyze data, like regression, classification, deep learning, clustering and many more. In clustering, the elements are grouped together in different groups, which contain objects which are identical in some nature. Such groups are called clusters, the elements that are part of such clusters exhibit some kind of affinity (strong or weak) towards the other elements contained in such a group and disparity with the elements which are not part of the group. This paper aims to see which clustering algorithm is the best to derive insights and patterns from the dataset at hand. There are two techniques which form the basis of the clustering algorithm namely: partition clustering and hierarchical clustering. The two algorithm that are going to be used on the dataset here are k-means which is a part of partition clustering and hierarchical clustering. These algorithms are going to provide insights on factors as dataset size, dataset type, how many clusters are being created, the quality of the clusters being created, the accuracy of the model and the performance of the model.*

Key Words: **Clustering, Hierarchical Clustering, K-means clustering algorithm.**

1. INTRODUCTION

For deriving insights and solving a problem, a huge amount of data needs to be collected from different sources and databases because nowadays we use advanced methods of collecting data. In such situations, grouping of such data becomes extremely essential to derive insights from such data. Hence we use clustering for performing this operation as it does the job of grouping similar elements which exhibit affinity with the other elements in the same cluster and disparity with the elements which are not part of the same cluster.[3] If the data in the dataset is divided into lesser number of clusters, this will result into a loss of some details but still the data still holds some information. This is used to showcase data elements using lesser number of clusters and this leads to modelling of data by the data elements using its cluster which it owns. Hence cluster analysis can be defined as set of patterns being arranged into clusters by using the principle of similarity.[1] The patterns that are part of the same cluster are extremely similar, than the patterns that are observed with the cluster which is besides it. This allows us to not make an effort to know whether there exists a difference between a supervised and an unsupervised classification that usually exists in discriminate analysis and clustering.[13] We are provided with a set consisting of elements which are already classified in a approach which is supervised, the real task here to label the elements which were not encountered earlier and that are not labelled yet. The items which are already labeled are provided to know the description of the classes which will help us in labeling a new item. We are provided with a dataset containing elements which are not labelled and the job is to group them into appropriate clusters, this is done in unsupervised learning approach.[2]

This approach of clustering is highly exploited in the fields of Artificial Intelligence, image processing, data mining, pattern recognition, marketing, statistics, medicine and many more fields have also started considering the insights derived from such models.

Some of the researchers improved the existing data clustering algorithms, few of the others designed the new methods to cluster the data and few of the scholars examined and analyzed various data clustering methods [6]. The goal of clustering is to determine the intrinsic grouping in a set of unlabeled data. The clustering algorithms used on datasets are continuously being improved by researchers to improve their accuracy and performance, some of them found of the new ways to cluster data by making changes in the algorithm, while others analyzed and tested different types of clustering mechanisms.[6] The end goal of all the algorithm is to make out the efficient grouping in a set of data which has not been labelled.

2. Software for Implementation

What are the algorithms that are used on the data?

The aim of this paper is to find the suitable algorithm which perfectly fits our data and find good insights by examining, studying, analyzing and evaluating the two algorithms which are k-means and hierarchical clustering. The two algorithms are selected as:

1. They offer flexibility for different types of datasets.
2. They are capable of dealing with high dimensionality.
3. They are accepted in almost all fields.
4. They are very popular.

Now let us talk about each of these algorithms and the other reasons behind considering these two.

2.1 K-means algorithm

One of the most famous unsupervised clustering algorithm, which is used to group data is the K-means algorithm. This partitioning algorithm, creates clusters which are independent and bound.[11] There are two major steps that the algorithm is implemented by: first being, dividing the data into k clusters, where k is the number of clusters which needs to be assumed and decided before implementing the algorithm. From the data, now we have to pick k random elements and call them the centroids of the k number of clusters. The second step is to find the distance of the point from the centroid and then it is assigned to the cluster. This distance needs to be the least distance as this should bring the point the closest to the centroid.[9] This technique is used to lessen the number of rounds or iterations and change the position of elements in the clusters.

In such a situation, we provide 'O' as the number of elements and k to be the total number of clusters and O1,O2,O3,……,Ok as the output.[3]

The goal of the algorithm is to minimize the total cost. Fig1 describes the workflow of the k-means clustering algorithm using a flowchart. It helps us to understand the algorithm in a pictorial representation. The step wise procedure of how k-means algorithm works are given below [10].

The reason we are doing this is to reduce significantly the total cost. The figure below shows the flowchart of the k-means algorithm for clustering. The procedure of how to step-wise implement the k-means algorithm is given below.

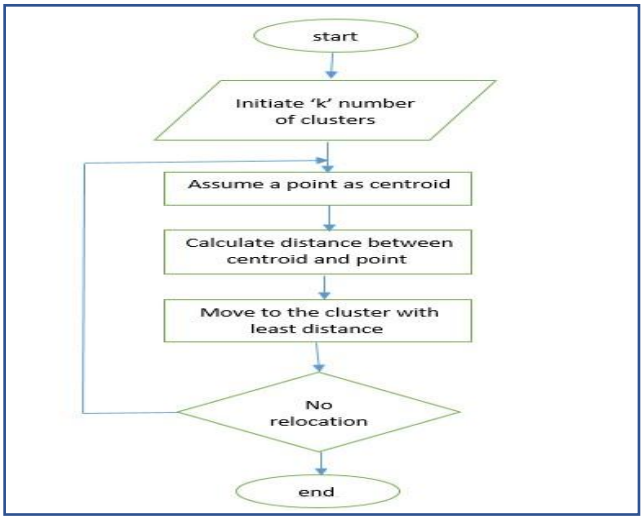

Chart-1. K-means clustering algorithm

1. Let 'k' be the total number of clusters.
2. Initialize the 'k' clusters
3. For each new vector:
   i. Measure the distance between every element and every centroid.
   ii. Find the closest centroid and add the new element to that specific cluster.

The Euclidean distance can be written as

$$= \sqrt{d_{\sum}(^{x\,x} - i)\ 2\ i=1}^{\ n}$$

Where x is the centroid, d is the distance between elements, x_i is the other element in the cluster,

We select k-means algorithm because

iii. time complexity=O(nki)

iv. space complexity=O (k + n)

v. it is not dependent on the order

Her o is the algorithm, k is the number of clusters, i is the number of iterations and n is the total number of elements.

2.2Hierarchical clustering algorithm

In organizing algorithms like the k-means algorithm, the number of clusters are decided at the beginning before the start of the clustering. The opposite happens in the hierarchical clustering in which we either combine or divide the groups that already exist which in return gives in which order the dividing or combining or the clustering takes place. A dendogram or a tree is used for representing a hierarchical clustering.

The hierarchical clustering can be performed in two ways, they are top-down and bottom-up. In this, we wither divide large clusters into smaller size clusters, or take smaller clusters and combine them into one big cluster.

Here, the distance between each point and every other point in the data set is calculated and the two points which have minimum distance is taken and combined to form a single cluster [3]. These two are now together taken as a single point or vector and then repeat the process of calculating the distance. This process will be continued till all points are combined to form a single cluster.

In this algorithm the following steps need to be performed, that is, to find the distance between each element with another element is computed and the elements with the least distance are combined to form a cluster.[3] This cluster is now treated as a single point or element and the step of finding the distances is continued. This is done until we have just one big cluster left. This is called the bottom up approach of the hierarchical clustering.

In top down approach, we have a single cluster to which we have to apply the algorithm that would give us small clusters having similar elements.

The flowchart of hierarchical clustering shows a step-wise procedure to implement it:

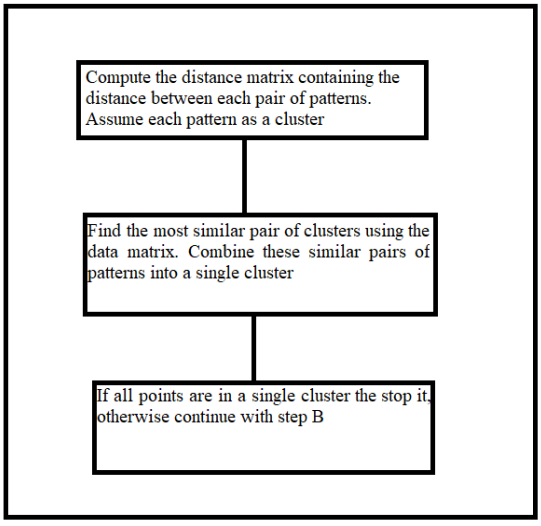

Chart-2. Flowchart of hierarchical clustering

The hierarchical algorithm is chosen because:

i. It exhibits embedded resilience at the granularity level.

ii. It can work with any kind of distances (Manhattan, Euclidean, etc.).

iii. It can be applied to any kind of attributes.

iv. It is very much adaptable.

3. How are the Algorithms Compared?

The two clustering algorithms are compared based on the following characteristics:

a) dataset size

b) dataset type

c) total number of clusters

d) model performance

e) model quality

f) results

g) accuracy of the model

4. Analysis and Observation

For each algorithm we need to have a predefined k that is the number of clusters except for hierarchical clustering before implementing the algorithm.

. K-means clustering is better than hierarchical clustering.

Based on dataset size, the quality of k-means is better when it is a large dataset. Hierarchical gives better performance with dataset which is smaller in size which is generally a subset of large dataset.

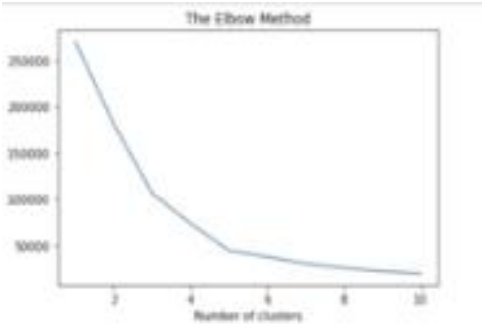

Chart- 3:Determining of how many clusters are required

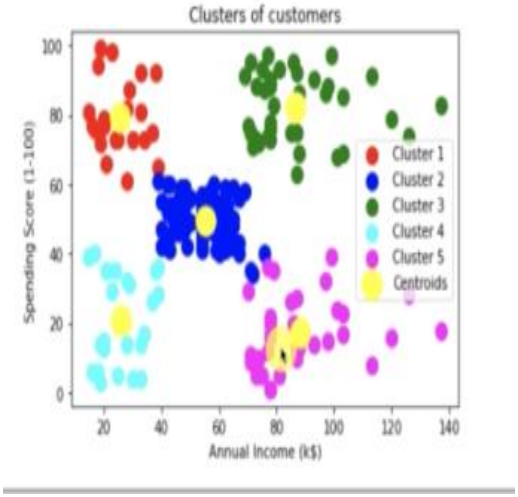

Chart-4: K-means clustering algorithm output

Hierarchical algorithm is better when cluster count increases in terms of accuracy. K-means algorithm give low accurate results.

Every algorithm exhibits some or the other kind of ambiguity and hence they need to be organized, hence an alternative algorithm designed to solve this problem is needed.

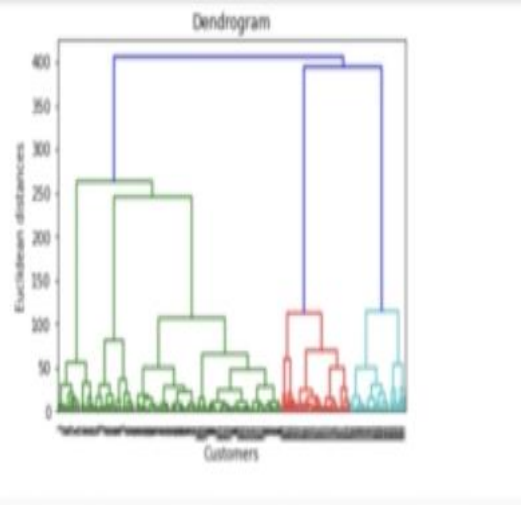

Chart-5.: Determining of how many clusters are required

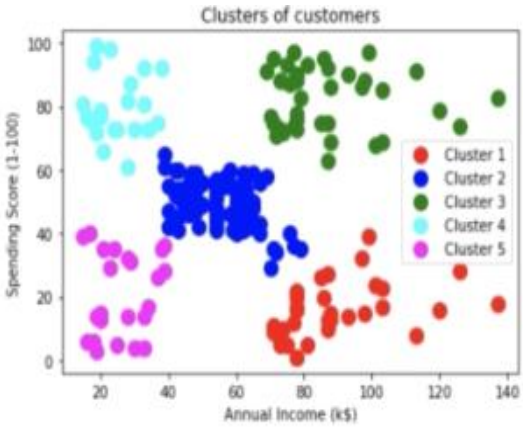

Chart- 6: Output of hierarchical clustering algorithm

An arbitrary dataset or a linear dataset can be used as per need. K-means performs better when linear data is under consideration because they are less resistant to noise. Such an arbitrary dataset can be easily found on the internet or some software. When such arbitrary data is used hierarchical algorithms give better results.

The process of assigning a cluster to an object is hindered by the noise in the dataset.

5. Conclusions

After comparing the results based on all the factors mentioned, here are some of the conclusions which are observed:

The results that were obtained were compared on the basis of the factors and the following conclusions were derived.

1. The performance is inversely proportional to the number of clusters considered.

2. The performance of k-means is better than Hierarchical algorithm.

3. When using large dataset, Quality of k-means becomes better and when using small dataset, hierarchical shows better results

4. Hierarchical gives better results when random dataset is used.

5. K-means is less resistant to noise.

We can also try to compare and analyze the algorithms using many more factors to come up with the most efficient method for clustering for any type of dataset which always provides the best results. Efforts must be made to increase the application area of algorithms of clustering and new methods must be developed to reduce the ambiguity in data

REFERENCES

[1] Kaur, Maninderjit, and Sushil Kumar Garg. "Survey on Clustering Techniques in Data Mining for Software Engineering." International Journal of Advanced and Innovative Research 3 (2014): 238-243.

[2] Sathya, R., and Annamma Abraham. "Comparison of supervised and unsupervised learning algorithms for pattern classification." Int J Adv Res Artificial Intell 2.2 (2013): 34-38.

[3] Singh, Nidhi, and Divakar Singh. "Performance Evaluation of KMeans and Heirarichal Clustering in Terms of Accuracy and Running Time." IJCSIT) International Journal of Computer Science and Information Technologies 3.3 (2012): 4119-4121.

[4] JIN, H., 2008. Multilevel spectral clustering with ascertainable clustering number. Journal of Computer Applications, 28(5), pp.1229-1231.

[5] International Journal of Science and Research (IJSR), 2016. Comparing EM Clustering Algorithm with Density Based Clustering Algorithm Using WEKA Tool. 5(7), pp.1199-1201.

[6] International Journal of Science and Research (IJSR), 2017. K Prototype Clustering with Efficient Summarization for Topic Evolutionary Tweet Stream Clustering. 6(1), pp.769-774

[7] Kriegel, H., Kröger, P. and Zimek, A., 2008. Detecting clusters in moderate-to-high dimensional data. Proceedings of the VLDB Endowment, 1(2), pp.1528-1529.

[8] Bandyopadhyay, S., 2011. Genetic algorithms for clustering and fuzzy clustering. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 1(6), pp.524-531.

[9] PAN, Z., 2010. Semi-supervised automatic clustering. Journal of Computer Applications, 30(10), pp.2614-2617.

[10] Kanungo, Tapas, et al. "An efficient k-means clustering algorithm: Analysis and implementation." Pattern Analysis and Machine Intelligence, IEEE Transactions on 24.7 (2002): 881-892.

[11] Ordonez, Carlos, and Paul Cereghini. "SQLEM: Fast clustering in SQL using the EM algorithm." ACM SIGMOD Record. Vol. 29. No. 2. ACM, 2000.

[12] Lailiyah, S. and Hafiyusholeh, M., 2016. PERBANDINGAN ANTARA METODE K-MEANS CLUSTERING DENGAN GATH-GEVA CLUSTERING. Jurnal Matematika "MANTIK", 1(2), p.26.

[13] Jain, Anil K., M. Narasimha Murty, and Patrick J. Flynn. "Data clustering: a review." ACM computing surveys (CSUR) 31.3 (1999): 264-323.